# Users' Guides to the Medical Literature

## I. How to Get Started

Andrew D. Oxman, MD, MSc; David L. Sackett, MD, MSc; Gordon H. Guyatt, MD, MSc;
for the Evidence-Based Medicine Working Group

### CLINICAL SCENARIO

You are a primary care physician inspired by a recent editorial in *JAMA* about lifelong learning.[1] You decide to use some of the time you normally take for continuing medical education conferences for "practice-based education" tailored to your own practice. You begin by setting aside 2 hours every week to read about relevant clinical problems.

It is now Friday morning and you have 2 hours to spend in the hospital library. You review a one-page list of questions you have generated from the patients you've seen in the prior week. Your questions include these: What should you tell a 33-year-old woman with migraine headaches who has asked for a prescription for sumatriptan after reading a magazine article about it? Should you be screening older men in your practice for prostate cancer? What should you tell the mother of a 6-month-old boy who had a febrile seizure about his risk of developing epilepsy? Should you try to reduce a 25-year-old asthmatic man's reliance on inhaled β-agonists? What should you tell a 50-year-old menopausal woman asking about hormone replacement?

### INTRODUCTION

This series of articles will help you translate the results of medical research into clinical practice. We've written them from the perspective of the busy clinician who wants to provide effective medical care but is sharply restricted in time for reading. We do not attempt a course in research methods; the series is about using, not doing, research. It is designed to help provide our patients with care that is based on the best evidence currently available—"evidence-based medicine."[2] Evidence-based medicine emphasizes the need to move beyond clinical experience and physiological principles to rigorous evaluations of the consequences of clinical actions. Knowing how to use the clinical literature is imperative for ensuring we are providing optimal patient care.

In this article we will present a general approach to using one's clinical reading time effectively and some specific suggestions for deciding which clinical articles to read. In subsequent articles we will go into more detail on how this approach can contribute to solving clinical problems in the treatment, prevention, diagnosis, and prognosis of disease.

### NEED FOR THE USERS' GUIDES SERIES

Clinical information comes from two principal sources, the individual patient and research. To provide effective care, both types of information are needed. Information about the individual patient is elicited through a careful history, physical examination, and other investigations. The ways in which clinicians obtain information from scientific research is less clear, but of no less importance to the quality of care that patients receive.

To the extent that clinicians rely on community standards or opinion leaders to guide their practice, there is an implicit assumption that their needs for scientific information are being met through these means; ie, that community standards and the recommendations of clinical experts (opinion leaders) reflect the best available scientific information. However, the ways in which experts' opinions and "standard practice" evolve are complex.[3] Variation in clinical practice, comparisons of practice with evidence-based standards, and evaluations of the recommendations of clinical experts suggest that expert opinion and "standard practice" do not provide adequate mechanisms for the transfer of scientific information into clinical decision making.[4,5] Expert opinion often lags far behind the evidence and is not infrequently inconsistent with evidence.[6] This is not to say that expert opinion may not be important and useful, but it is clearly not sufficient.

The Editorial accompanying this article, the first of a series, reviews the reasons why clinicians need tools to evaluate and use the medical literature in their day-to-day clinical practice.[7] This series is designed to fill that need.

**For editorial comment see p 2096.**

For reasons of both logic and efficiency, we have sought uniformity in presentation of the Users' Guides by organizing each set into three basic questions:

1. Are the results of the study valid?
2. What are the results?
3. Will the results help me in caring for my patients?

Yes and no are often not adequate answers to these questions. This may contrast with readers' intuitive approach. After all, the Users' Guides are designed to help clinicians make decisions, and most clinical decisions are black and white; for example, we either start a treatment or we do not. It is understandable, therefore, that we seek black or white answers from the clinical literature. The article is right or wrong; the treatment works or it does not; the results apply to my patient or they do not. Unfortunately, evidence comes in shades of gray. Often, results may be valid, perhaps demonstrate an important effect, and might improve patient care.

The goal of the Users' Guides presented in this series of articles is to help clinicians sift through these shades of gray and make appropriate decisions, recognizing the "level" of certainty (or

From the Departments of Clinical Epidemiology and Biostatistics (Drs Oxman, Sackett, and Guyatt), Family Medicine (Dr Oxman), and Medicine (Drs Sackett and Guyatt), McMaster University, Hamilton, Ontario. A complete list of members of the Evidence-Based Medicine Working Group appears at the end of this article.

Reprint requests to McMaster University Health Sciences Centre, 1200 Main St W, Room 2C12, Hamilton, Ontario, Canada L8N 3Z5 (Dr Guyatt).

Guides for Selecting Articles That Are Most Likely to Provide Valid Results

| | |
|---|---|
| **Primary Studies** | |
| Therapy | • Was the assignment of patients to treatments randomized?<br>• Were all of the patients who entered the trial properly accounted for and attributed at its conclusion? |
| Diagnosis | • Was there an independent, blind comparison with a reference standard?<br>• Did the patient sample include an appropriate spectrum of the sort of patients to whom the diagnostic test will be applied in clinical practice? |
| Harm | • Were there clearly identified comparison groups that were similar with respect to important determinants of outcome (other than the one of interest)?<br>• Were outcomes and exposures measured in the same way in the groups being compared? |
| Prognosis | • Was there a representative patient sample at a well-defined point in the course of disease?<br>• Was follow-up sufficiently long and complete? |
| **Integrative Studies** | |
| Overview | • Did the review address a clearly focused question?<br>• Were the criteria used to select articles for inclusion appropriate?* |
| Practice guidelines | • Were the options and outcomes clearly specified?<br>• Did the guideline use an explicit process to identify, select, and combine evidence?* |
| Decision analysis | • Did the analysis faithfully model a clinically important decision?<br>• Was valid evidence used to develop the baseline probabilities and utilities?* |
| Economic analysis | • Were two or more clearly described alternatives compared?<br>• Were the expected consequences of each alternative based on valid evidence?* |

*Each of these guides makes an implicit or explicit reference to investigators' need to evaluate the validity of the studies that they are reviewing to produce their integrative article. The validity criteria one would use in making this evaluation would depend on the area being addressed (therapy, diagnosis, prognosis, or harm), and are those that are presented in the part of the Table dealing with primary articles.

strength of inference) underlying those decisions. The first key question—"Are the results of the study valid?"—and the last—"Will the results help me in caring for my patients?"—reflect the need to make a decision, despite the fact that the strength of the inferences that can be made based on a study spans a spectrum from strong to weak. Since this is a series on how to use research in taking care of patients, not how to do research, we will focus on flaws in study design or implementation that are most likely to weaken the strength of inference in ways that seriously distort clinical decisions based on them.

In the remainder of this article, we will introduce strategies for (1) framing clinical questions that are pertinent and answerable, (2) tracking down articles, and (3) deciding which articles to read, and which to believe.

## ASKING QUESTIONS THAT ARE PERTINENT AND ANSWERABLE

Clinical questions arise continuously in the course of providing routine medical care, but must be clearly formulated to ensure clear answers. Most clinical questions can be formulated in terms of a simple relationship between the patient, some "exposure" (to a treatment, a diagnostic test, or a potentially harmful agent), and one or more specific outcomes of interest, as shown in the following modifications of the questions from the scenario at the beginning of this article:

- Would sumatriptan (exposure) reduce the severity of headache pain (outcome) in this woman with frequent migraine attacks (patient)?—a question of therapy.
- Would a prostate-specific antigen test (exposure), if performed in this symptomless elderly man (patient), decrease his risk of dying from prostate cancer (outcome)?—a question of secondary prevention through early diagnosis.
- Does the febrile seizure (exposure) that this 6-month-old infant (patient) just had increase the likelihood that he will develop epilepsy (outcome)?—a question of prognosis.
- Do β-agonists (exposure) increase the risk of death (outcome) in this asthmatic man (patient)?—a question of harm.

The importance of such focused questions can be quickly assessed, and priority given to problems that are seen routinely and have practically important consequences. In general, those questions that are clearly related to a clinical decision about whether to use a therapeutic, preventive, or diagnostic intervention are the ones that warrant the most time. Focusing the question clarifies the target of the literature search and permits use of the appropriate guides for assessing validity in screening the titles and abstracts of the articles that are located.

For example, the question posed in the scenario at the beginning of this article about hormone replacement, while likely to be important in most primary care practices, is not well focused. It is worthwhile to clarify the type of patient and the outcomes of interest before beginning to look for an answer. Is the woman seeking treatment for hot flashes or is she asymptomatic? If the woman is asymptomatic and is wondering if she should take estrogen to prevent osteoporosis, clinically important outcomes that might be considered include hip fracture, cardiovascular disease, breast and endometrial cancer, and vaginal bleeding. In this case, a good approach might be to start by looking for published clinical practice guidelines instead of tracking down the evidence for each outcome. Later in this series we will present guides for how to critically appraise practice guidelines.

## TRACKING DOWN ARTICLES

Having posed a pertinent, answerable clinical question, you can proceed to track down the best available evidence. There are four routes for doing this: asking someone, checking reference lists in textbooks, finding a relevant article in your own reprint file, and using a bibliographic database such as MEDLINE. Asking a colleague or consultant is highly efficient, and makes most sense when the question concerns an exposure or treatment or patient you are unlikely to encounter again. If a recent textbook is at hand (published or updated within the previous year), you can follow your reading of the appropriate passage by checking the references cited by the author. Because a textbook is only as up-to-date as its most recent reference, all are at least partly out-of-date even before they are published. A new type of "subscription" textbook addresses this problem by providing periodic updates and often cites the evidence used in making its changes.[8,9] While frequent updates help protect against being out-of-date, they do not ensure that the conclusions of the clinical experts writing textbook chapters are valid. Prototypes of textbooks that are based on systematic reviews of validated evidence are available for obstetrical[10] and neonatal problems,[11] but most textbooks and review articles do not qualify as scientific overviews.[12]

A third starting point may be an article in your personal reprint file. Since the amount of time required to maintain an up-to-date file of clinical articles is formidable, you are unlikely to have the key article at hand. New methods for retrieving the current medical literature are rendering personal filing systems nonessential, if not obsolete.

The final route, conducting electronic searches of the medical literature, is fast becoming a basic skill for practicing modern, evidence-based medicine. Electronic access to MEDLINE is readily available in North America in a variety of on-line and CD-ROM formats. Clinicians can easily acquire the basic skills[13] and learn to retrieve the same number of relevant citations as librarians, even if their searches remain a bit messier.[14] The addition of structured abstracts to MEDLINE and the development of da-

tabases that have screened articles for their validity and clinical relevance, such as the *Oxford Database of Perinatal Trials*[15] and an electronic version of the ACP Journal Club, promise to make the task of retrieving information from the medical literature even easier. You can seek a review article (often the best place to start) by adding, to whatever Medical Subject Heading (MeSH) terms are used to identify the disorder and "exposure," in your MEDLINE search, the search term REVIEW (PT) (PT stands for publication type). You are more likely to find a methodologically sound review article by using the term META-ANALYSIS (PT) instead of REVIEW. Another potential place to start is with practice guidelines, which now have their own search term PRACTICE GUIDELINE (PT). Recruiting a librarian to help you with your first few searches may help you learn to avoid searches that are too broad and unfocused, or too narrow and thus risk missing key articles. Increasing numbers of physicians are finding that MEDLINE searches can help them solve clinical problems and improve patient care and clinical outcomes.[16]

## DECIDING IF AN ARTICLE IS LIKELY TO PROVIDE VALID RESULTS

The first question applied to any article tracked down in an effort to find an answer for a clinical problem concerns its closeness to the truth: are the results of this article valid? The Table presents two key guides to assess validity for primary studies (those that provide original data on a topic) and integrative studies (those that summarize data from primary studies). For each type of integrative study, the first criterion has to do with whether the question is appropriately framed, and the second with whether the evidence was appropriately collected and summarized. The clinician can use these most important criteria to rapidly screen an abstract to determine whether it warrants the additional time required to read it in detail. The busy clinician who has tracked down a number of articles on a question can use the guides to choose the one or two articles most likely to provide a valid answer. These criteria can also be used to reduce the clinical literature to a manageable size when trying to keep up with new advances that are pertinent to one's practice. If a more detailed review of an article's methods reveals that these "validity" guides are met, readers can turn their attention to the other guides designed to help them answer the next two key questions: what are the results and will they benefit my patient care?

## CONCLUSION

Subsequent articles in this series will describe strategies for efficiently selecting and using each of the types of articles in the Table. In doing so, they will describe the justification and application of guides for determining whether the results of an article are valid and applicable to the clinical decisions you must make.

Readers should be warned that the guides do not come with definitive answers. Learning to apply them can be challenging. However, it can also be extremely gratifying. More important, it is only by translating good evidence into good clinical decisions that we can be sure that we do more good than harm for our patients.

The Evidence-Based Medicine Working Group has been supported in part by Dr Sackett's Trillium Clinical Scientist Award.

The Evidence-Based Medicine Working Group includes the following: Gordon H. Guyatt (chair), MD, MSc, George Browman, MD, MSc, Deborah Cook, MD, MSc, Hertzel Gerstein, MD, MSc, Brian Haynes, MD, MSc, PhD, Robert Hayward, MD, MPH, Mitchell Levine, MD, MSc, Jim Nishikawa, MD, and David L. Sackett, MD, MSc, Departments of Medicine and Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario; Patrick Brill-Edwards, MD, Michael Farkouh, MD, Anne Holbrook, MD, PharmD, MSc, Roman Jaeschke, MD, MSc, Hui Lee, MD, MSc, Lori McDonald, MD, MSc, Ameen Patel, MD, Stephane Sauve, MD, MSc, Department of Medicine, McMaster University; Ted Haines, MD, MSc, Departments of Clinical Epidemiology and Biostatistics and Occupational Health Program, McMaster University; Elizabeth Juniper, MCSP, MSc, Bernie O'Brien, MD, MSc, K. S. Trout, FRCE, Stephen Walter, PhD, Department of Clinical Epidemiology and Biostatistics, McMaster University; Eric Bass, MD, MPH, Division of Internal Medicine, The Johns Hopkins University School of Medicine, Baltimore, Md; Allan Detsky, MD, PhD, Department of Clinical Epidemiology and Biostatistics, McMaster University, and the Departments of Health Administration and Medicine, University of Toronto (Ontario); Michael Drummond, BSc, MCom, DPhil, Centre for Health Economics, University of York, United Kingdom; Andreas Laupacis, MD, MSc, Departments of Medicine and Epidemiology and Community Medicine, University of Ottawa (Ontario) and Department of Clinical Epidemiology and Biostatistics, McMaster University; Virginia Moyer, MD, MPH, Department of Pediatrics, University of Texas, Houston; David Naylor, MD, DPhil, Clinical Epidemiology Research Programme, Sunnybrook Health Science Centre, Institute for Clinical Evaluative Sciences in Ontario, Departments of Health Administration, Medicine, and Behavioral Sciences, University of Toronto (Ontario); Andrew Oxman, MD, MSc, FACPM, Departments of Clinical Epidemiology and Biostatistics and Family Medicine, McMaster University; John Philbrick, MD, Department of Internal Medicine, University of Virginia, Charlottesville; W. Scott Richardson, MD, Department of Medicine, University of Rochester (NY) School of Medicine and Dentistry; Jack Sinclair, MD, Departments of Clinical Epidemiology and Biostatistics and Pediatrics, McMaster University; Brian L. Strom, MD, MPH, Center for Clinical Epidemiology and Biostatistics and Division of General Internal Medicine, University of Pennsylvania School of Medicine, Philadelphia; Peter Tugwell, MD, MSc, George Wells, MSc, PhD, Clinical Epidemiology Unit and Departments of Medicine and Epidemiology, University of Ottawa (Ontario); Sean Tunis, MD, MSc, Health Program, Office of Technology Assessment, US Congress, Washington, DC; John Williams, Jr, MD, MHS, Division of General Internal Medicine, The University of Texas Health Science Center at San Antonio; and Mark Wilson, MD, MPH, Department of Medicine, Bowman Gray School of Medicine, Winston-Salem, NC.

Drs Cook, Guyatt, Naylor, and Oxman are Career Scientists, and Dr Sackett is a Trillium Clinical Scientist, of the Ontario Ministry of Health. Dr Detsky holds a National Health Research Scholar award and Drs Haynes and Walter hold National Health Scientist awards from the National Health and Research Development Centre, Health and Welfare, Canada. Dr Cook is a Scholar of the St. Joseph's Hospital Foundation, Hamilton, Ontario. Dr Levine holds the Pharmaceutical Manufacturer's Association of Canada—Health Research Foundation/Medical Research Council of Canada Career Award in Medicine. Dr Williams is a Robert Wood Johnson Generalist Physician Faculty Scholar.

### References

1. Manning PR, DeBakey L. Lifelong learning tailored to individual clinical practice. *JAMA*. 1992; 268:1135-1136.
2. Evidence-Based Medicine Working Group. Evidence-based medicine: a new approach to teaching the practice of medicine. *JAMA*. 1992;268:2420-2425.
3. Eddy DM. Clinical policies and the quality of clinical practice. *N Engl J Med*. 1982;307:343-347.
4. Stross JK, Harlan WR. The dissemination of new medical information. *JAMA*. 1979;241:2622-2624.
5. Williamson JW, German PS, Weiss R, Skinner EA, Bowes F. Health science information management and continuing education of physicians: a survey of US primary care practitioners and their opinion leaders. *Ann Intern Med*. 1989;110: 151-160.
6. Antman EM, Lao J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts: treatments for myocardial infarction. *JAMA*. 1992;268:240-248.
7. Guyatt GH, Rennie D. Users' guides to the medical literature. *JAMA*. 1993;270:2096-2097.
8. Rubenstein E, Federman D, eds. *Medicine*. New York, NY: Scientific American Medicine; 1993.
9. Rubenstein E, Federman D, eds. *Care of the Surgical Patient*. New York, NY: Scientific American Medicine; 1993.
10. Chalmers I. Evaluating the effects of care during pregnancy and childbirth. In: Chalmers I, Enkin M, Keirse MJNC, eds. *Effective Care in Pregnancy and Childbirth*. Oxford, England: Oxford University Press; 1989:3-38.
11. Sinclair JC, Bracken ME, eds. *Effective Care of the Newborn Infant*. Oxford, England: Oxford University Press; 1992.
12. Mulrow CD. The medical review article: state of the science. *Ann Intern Med*. 1987;106: 485-488.
13. Haynes RB, McKibbon KA, Fitzgerald D, Guyatt GH, Walker CJ, Sackett DL. How to keep up with the medical literature, V: access by personal computer to the medical literature. *Ann Intern Med*. 1986;105:810-814.
14. Haynes RB, McKibbon KA, Walker CJ, Ryan C, Fitzgerald D, Ramsden ME. Online access to MEDLINE in clinical settings: a study of use and usefulness. *Ann Intern Med*. 1990;112:78-84.
15. Chalmers I, ed. *Oxford Database of Perinatal Trials*. Version 1.2, disk issue 7. Oxford, England: Oxford University Press; spring 1992.
16. Lindberg DAB, Siegel ER, Rapp BA, et al. Use of MEDLINE by physicians for clinical problem solving. *JAMA*. 1993;269:3124-3129.

# Users' Guides to the Medical Literature

## II. How to Use an Article About Therapy or Prevention

## A. Are the Results of the Study Valid?

Gordon H. Guyatt, MD, MSc; David L. Sackett, MD, MSc; Deborah J. Cook, MD, MSc;
for the Evidence-Based Medicine Working Group

### CLINICAL SCENARIO

You are working as an internal medicine resident in a rheumatology rotation and are seeing a 19-year-old woman who has had systemic lupus erythematosus diagnosed on the basis of a characteristic skin rash, arthritis, and renal disease. A renal biopsy has shown diffuse proliferative nephritis. A year ago her creatinine level was 140 μmol/L, 6 months ago it was 180 μmol/L, and in a blood sample taken a week before this clinic visit, 220 μmol/L. Over the last year she has been taking prednisone, and over the last 6 months, cyclophosphamide, both in appropriate doses.

From the Departments of Medicine and Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario.

A complete list of the members (with affiliations) of the Evidence-Based Medicine Working Group appears in the first article of this series (*JAMA*. 1993;270:2093-2095). The following members contributed to this article: Gordon Guyatt (Chair), MD, MSc; Eric Bass, MD, MPH; Patrick Brill-Edwards, MD; George Browman, MD, MSc; Deborah Cook, MD, MSc; Michael Farkouh, MD; Hertzel Gerstein, MD, MSc; Brian Haynes, MD, MSc, PhD; Robert Hayward, MD, MPH; Anne Holbrook, MD, PharmD, MSc; Roman Jaeschke, MD, MSc; Elizabeth Juniper, MCSP, MSc; Andreas Laupacis, MD, MSc; Hui Lee, MD, MSc; Mitchell Levine, MD, MSc; Virginia Moyer, MD, MPH; Jim Nishikawa, MD; Andrew Oxman, MD, MSc, FACPM; Ameen Patel, MD; John Philbrick, MD; W. Scott Richardson, MD; Stephane Sauve, MD, MSc; David Sackett, MD, MSc; Jack Sinclair, MD; K. S. Trout, FRCE; Peter Tugwell, MD, MSc; Sean Tunis, MD, MSc; Stephen Walter, PhD; John Williams, Jr, MD, MHS; and Mark Wilson, MD, MPH.

Reprint requests to McMaster University Health Sciences Centre, 1200 Main St W, Room 2C12, Hamilton, Ontario, Canada L8N 3Z5 (Dr Guyatt).

You are distressed by the rising creatinine level and the rheumatology fellow with whom you discuss the problem suggests that you contact the hematology service to consider a trial of plasmapheresis. The fellow states that plasmapheresis is effective in reducing the level of the antibodies responsible for the nephritis and cites a number of trials that have suggested therapy is beneficial. When you ask her if any of the studies were randomized clinical trials, she acknowledges that she is uncertain.

You present the dilemma to the attending physician who responds with a suggestion that, before you make a decision, you review the relevant literature. The attending recommends that you bring the patient back in 2 weeks, at which time you can offer her the appropriate therapy.

### THE SEARCH

You decide that the most helpful article would include patients with severe lupus that threatens renal function and who are already receiving immunosuppressive agents. Plasmapheresis must be compared with a control management strategy, and patients must be randomized to receive or not receive the plasmapheresis. Finally, the article must report clinically important outcomes, such as deterioration in renal function. You are familiar with the software program Grateful Med and use it for your search. The program provides a listing of Medical Subject Headings (MeSH), and you quickly find that "lupus nephritis" is one such heading and "plasmapheresis" another. You add a methodological term that will restrict your results to high-quality studies, "randomized controlled trial (PT)" (PT stands for publication type). The search, which you restrict to English-language articles, yields a total of three articles. One is a trial of prednisone and cyclophosphamide[1]; a second examines the effect of plasmapheresis on risk of infection.[2] The third citation, which describes "a controlled trial of plasmapheresis," appears most likely to address the issue at hand, the effectiveness of plasmapheresis in improving clinically important outcomes.

The relevant article is a randomized trial in which 46 patients received a standard therapeutic regimen of prednisone and cyclophosphamide, and 40 patients received standard therapy plus plasmapheresis.[3] Despite the fact that antibody levels decreased in those undergoing plasmapheresis, there was a trend toward a greater proportion of the plasmapheresis-treated patients dying (20% vs 13%) or developing renal failure (25% vs 17%). This seems to settle the issue of whether to offer your patient plasmapheresis. You wonder, however, whether the study could have led to an inaccurate or biased outcome. The remainder of this article will provide you with the tools to address this question.

Readers' Guides for an Article About Therapy

**Are the results of the study valid?**
- Primary guides:
  - Was the assignment of patients to treatments randomized?
  - Were all patients who entered the trial properly accounted for and attributed at its conclusion?
    - Was follow-up complete?
    - Were patients analyzed in the groups to which they were randomized?
- Secondary guides:
  - Were patients, health workers, and study personnel "blind" to treatment?
  - Were the groups similar at the start of the trial?
  - Aside from the experimental intervention, were the groups treated equally?

**What were the results?**
- How large was the treatment effect?
- How precise was the estimate of the treatment effect?

**Will the results help me in caring for my patients?**
- Can the results be applied to my patient care?
- Were all clinically important outcomes considered?
- Are the likely treatment benefits worth the potential harms and costs?

## INTRODUCTION

In the first article[4] in this series we introduced a framework for using the medical literature to solve patient problems and provide better clinical care. This second article begins the discussion of how to use a report dealing with therapy or prevention. In this article, we will use the term "therapy" in a broad sense. As we've described elsewhere,[5] the same guides can be applied to evaluation of therapeutic interventions (directed at reducing symptoms and curing disease) and preventive interventions (directed at reducing the risk of disease or disease complications).

## THE FRAMEWORK

As with articles on other clinical questions, one can usefully pose three questions about an article on therapy.

### Are the Results of the Study Valid?

This question has to do with the validity or accuracy of the results and considers whether the treatment effect reported in the article represents the true direction and magnitude of the treatment effect. Another way to state this question is this: Do these results represent an unbiased estimate of the treatment effect, or have they been influenced in some systematic fashion to lead to a false conclusion?

### What Were the Results?

If the results are valid and the study likely yields an unbiased assessment of treatment effect, then the results are worth examining further. This second question considers the size and precision of the treatment's effect. The best estimate of that effect will be the study findings themselves; the precision of the estimate will be superior in larger studies.

### Will the Results Help Me in Caring for My Patients?

This question has two parts. First, are the results applicable to your patient? You should hesitate to institute the treatment either if your patient is too dissimilar from those in the trial, or if the outcome that has been improved isn't important to your patient. Second, if the results are applicable, what is the net impact of the treatment? The impact depends on both benefits and risks (side effects and toxic effects) of treatment and the consequences of withholding treatment. Thus, even an effective therapy might be withheld when a patient's prognosis is already good without treatment, especially when the treatment is accompanied by important side effects and toxic effects.

We summarize our approach to evaluating and applying the results of articles addressing therapeutic effectiveness in the Table. House staff and practicing physicians alike need an approach that is both efficient and comprehensive. We have therefore labeled validity criteria as "primary"—those few that can quickly be applied by readers with limited time—and "secondary"—those that, though still important, can be reserved for articles that pass the initial guides and for readers who have both the need and the time for a deeper review.

## ARE THE RESULTS OF THIS ARTICLE VALID?

### Primary Guides

**Was the Assignment of Patients to Treatment Randomized?**—During the 1970s and early 1980s surgeons increasingly undertook extracranial-intracranial bypass (that is, anastomosis of a branch of the external carotid artery, the superficial temporal, to a branch of the internal carotid artery, the middle cerebral). They believed it prevented strokes in patients whose symptomatic cerebrovascular disease was otherwise surgically inaccessible. This conviction was based on the comparison of clinical outcomes among nonrandomized cohorts of patients who, for whatever reason, had and had not undergone this operation, for the former appeared to fare much better than the latter. To the surprise of many and the indignation of a few, a large multicenter randomized trial in which patients were allocated to receive or forego this operation using a process analogous to flipping a coin, demonstrated that the only effect of surgery was to make patients worse off in the immediate postsurgical period; long-term outcome was unaffected.[6]

Other surprises generated by randomized trials that contradicted the results of less rigorous trials include the demonstration that steroids may increase (rather than reduce) mortality in patients with sepsis,[7] that steroid injections do not ameliorate facet-joint back pain,[8] and that plasmapheresis does not benefit patients with polymyositis.[9] Such surprises may occur when treatments are assigned by random allocation, rather than by the conscious decisions of clinicians and patients. In short, clinical outcomes result from many causes, and treatment is just one of them: underlying severity of illness, the presence of comorbid conditions, and a host of other prognostic factors (unknown as well as known) often swamp any effect of therapy. Because these other features also influence the clinician's decision to offer the treatment at issue, nonrandomized studies of efficacy are inevitably limited in their ability to distinguish useful from useless or even harmful therapy. As confirmation of this fact, it turns out that studies in which treatment is allocated by any method other than randomization tend to show larger (and frequently false-positive) treatment effects than do randomized trials.[10-13] The beauty of randomization is that it assures, if sample size is sufficiently large, that both known and unknown determinants of outcome are evenly distributed between treatment and control groups.

What can the clinician do if no one has done a randomized trial of the therapeutic question she faces? She still has to make a treatment decision, and so must rely on weaker studies. In a later article in this series devoted to deciding whether a therapy or an exposure causes harm (a situation when randomization is usually not possible), we deal with how to assess weaker study designs. For now, you should bear in mind that nonrandomized studies provide much weaker evidence than do randomized trials.

**Were All Patients Who Entered the Trial Properly Accounted for and Attributed at Its Conclusion?**—This guide has two components: was follow-up complete and were patients analyzed in the groups to which they were randomized?

*Was Follow-up Complete?*—Every patient who entered the trial should be accounted for at its conclusion. If this is not done, or if substantial numbers of patients are reported as "lost to follow-up," the validity of the study is open to question. The greater the number of subjects who are lost, the more the trial may be subject to bias because patients who are lost often have different prognoses from those who are retained, and may disappear because they suffer adverse outcomes (even death) or because they are doing well (and so did not return to the clinic to be assessed).

Readers can decide for themselves when the loss to follow-up is excessive by assuming, in positive trials, that all patients lost from the treatment group did badly, and all lost from the control group did well, and then recalculating the outcomes under these assumptions. If the conclusions of the trial do not change, then the loss to follow-up was not excessive. If the conclusions would change, the strength of inference is weakened (that is, less confidence can be placed in the study results). The extent to which the inference is weakened will depend on how likely it is that treatment patients lost to follow-up all did badly, while control patients lost to follow-up all did well.

*Were Patients Analyzed in the Groups to Which They Were Randomized?*—As in routine practice, patients in randomized trials sometimes forget to take their medicine or even refuse their treatment altogether. Readers might, on first blush, agree that such patients who never actually received their assigned treatment should be excluded from analyses for efficacy. Not so.

The reasons people don't take their medication are often related to prognosis. In a number of randomized trials, noncompliant patients have fared worse than those who took their medication as instructed, even after taking into account all known prognostic factors, and even when their medications were placebos![14-19] Excluding noncompliant patients from the analysis leaves behind those who may be destined to have a better outcome and destroys the unbiased comparison provided by randomization.

The situation is similar with surgical therapies. Some patients randomized to surgery never have the operation because they are too sick or suffer the outcome of interest (such as stroke or myocardial infarction) before they get to the operating room. If investigators include such patients, who are destined to do badly, in the control arm but not in the surgical arm of a trial, even a useless surgical therapy will appear to be effective. However, the apparent effectiveness of surgery will come not from a benefit to those who have surgery, but the systematic exclusion of those with the poorest prognosis from the surgical group.

This principle of attributing all patients to the group to which they were randomized results in an intention-to-treat analysis. This strategy preserves the value of randomization: prognostic factors that we know about, and those we don't know about, will be, on average, equally distributed in the two groups, and the effect we see will be just that due to the treatment assigned.

## Secondary Guides

**Were Patients, Their Clinicians, and Study Personnel "Blind" to Treatment?**—Patients who know that they are on a new, experimental treatment are likely to have an opinion about its efficacy, as are their clinicians or the other study personnel who are measuring responses to therapy. These opinions, whether optimistic or pessimistic, can systematically distort both the other aspects of treatment and the reporting of treatment outcomes, thereby reducing our confidence in the study's results. In addition, unblinded study personnel who are measuring outcomes may provide different interpretations of marginal findings or differential encouragement during performance tests, either one of which can distort their results.[20]

The best way of avoiding all this bias is double-blinding (sometimes referred to as double-masking), which is achieved in drug trials by administering a placebo, indistinguishable from active treatment in appearance, taste, and texture, but lacking the putative active ingredient, to the control group. When you read reports on treatments (such as trials of surgical therapies) in which patients and treating clinicians cannot be kept blind, you should note whether investigators have minimized bias by blinding those who assess clinical outcomes.

**Were the Groups Similar at the Start of the Trial?**—For reassurance about a study's validity, readers would like to be informed that the treatment and control groups were similar for all the factors that determine the clinical outcomes of interest save one: whether they received the experimental therapy. Investigators provide this reassurance when they display the entry or baseline prognostic features of the treatment and control patients. Although we never will know whether similarity exists for the unknown prognostic factors, we are reassured when the known prognostic factors are nicely balanced.

Randomization doesn't always produce groups balanced for known prognostic factors. When the groups are small, chance may place those with apparently better prognoses in one group. As sample size increases, this is less and less likely (this is analogous to multiple coin flips: one wouldn't be too surprised to see seven heads out of 10 coin flips, but one would be very surprised to see 70 heads out of 100 coin flips).

The issue here is not whether there are statistically significant differences in known prognostic factors between treatment groups (in a randomized trial one knows in advance that any differences that did occur happened by chance), but rather the magnitude of these differences. If they are large, the validity of the study may be compromised. The stronger the relationship between the prognostic factors and outcome, and the smaller the trial, the more the differences between groups will weaken the strength of any inference about efficacy.

All is not lost if the treatment groups are not similar at baseline. Statistical techniques permit adjustment of the study result for baseline differences. Accordingly, readers should look for documentation of similarity for relevant baseline characteristics and, if substantial differences exist, should note whether the investigators conducted an analysis that adjusted for those differences. When both unadjusted and adjusted analyses reach the same conclusion, readers justifiably gain confidence in the validity of the study result.

**Aside From the Experimental Intervention, Were the Groups Treated Equally?**—Care for experimental and control groups can differ in a number of ways besides the test therapy, and differences in care other than that under study can weaken or distort the results. If one group received closer follow-up, events might be more likely to be reported, and patients may be treated more intensively with nonstudy therapies. For example, in trials of new forms of therapy for resistant rheumatoid arthritis, ancillary treatment with systemic steroids (extremely effective for relieving symptoms), if administered more frequently to the control group than to the treatment group, could obscure an experimental drug's true treatment effect (unless exacerbation requiring steroids were itself counted as an outcome).

Interventions other than the treatment under study, when differentially applied to the treatment and control groups, often are called "cointerventions." Cointervention is a more serious problem when double-blinding is absent, or when the use of very effective nonstudy treatments is permitted at the physicians' discretion. Clinicians gain greatest confidence in the results when permissible cointerventions are described in the "Methods" section and documented to be infrequent occurrences in the results.

The foregoing five guides (two primary and three secondary), applied in sequence, will help the reader determine whether the results of an article on therapy are likely to be valid. If the results are valid, then the reader can proceed to consider the magnitude of the effect and the applicability to her patients.

## ARE THE RESULTS OF THE STUDY VALID? THE PLASMAPHERESIS TRIAL

Readers may be interested in how well the trial of plasmapheresis in patients with lupus nephritis met the tests of validity. With respect to primary criteria, randomization was rigorously conducted, as treatment was assigned through a phone call to the study's Methods Center. One patient assigned to standard therapy was lost to follow-up, and all the other patients were analyzed in the group to which they had been assigned. With respect to secondary criteria, the study was not blinded, the two groups were comparable at the start of the trial, and the authors provide little information about comparability of other treatments.

In the introductory article in this series, we described the concept of strength of inference. The final assessment of validity is never a "yes" or "no" decision and must, to some extent, be subjective. We judge that the methods in this trial were, overall, strong and provide a valid start for deciding whether or not to administer plasmapheresis to our patient with severe lupus nephritis.

So, in part A of this two-part essay, we have described how to answer the question: Are the results of the study valid? Part B will describe how to answer the second and third questions: What are the results of the trial? and Will the results help me in caring for my patient?

### References


1. Levey AS, Lan SP, Corwin HL, et al. Progression and remission of renal disease in the Lupus Nephritis Collaborative Study: results of treatment with prednisone and short-term oral cyclophosphamide. *Ann Intern Med.* 1992;116:114-123.
2. Pohl MA, Lan SP, Berl T. Plasmapheresis does not increase the risk for infection in immunosuppressed patients with severe lupus nephritis: the Lupus Nephritis Collaborative Study Group. *Ann Intern Med.* 1991;114:924-929.
3. Lewis EJ, Hunsicker LG, Lan SP, Rohde RD, Lachin JM. A controlled trial of plasmapheresis therapy in severe lupus nephritis. *N Engl J Med.* 1992;326:1373-1379.
4. Oxman AD, Sackett DL, Guyatt GH, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, I: how to get started. *JAMA.* 1993;270:2093-2095.
5. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical Epidemiology: A Basic Science for Clinical Medicine.* 2nd ed. Boston, Mass: Little Brown & Co; 1991.
6. Haynes RB, Mukherjee J, Sackett DL, Taylor DW, Barnett HJM, Peerless SJ. Functional status changes following medical or surgical treatment for cerebral ischemia: results in the EC/IC Bypass Study. *JAMA.* 1987;257:2043-2046.
7. Bone RC, Fisher CJ, Clemmer TP, et al. Controlled trial of high-dose methylprednisolone in the treatment of severe sepsis and septic shock. *N Engl J Med.* 1987;317:653-658.
8. Carette S, Marcoux S, Truchon R, et al. A controlled trial of corticosteroid injections into facet joints for chronic low back pain. *N Engl J Med.* 1991;325:1002-1007.
9. Miller FW, Leitman SF, Cronin ME, et al. Controlled trial of plasma exchange and leukapheresis in polymyositis and dermatomyositis. *N Engl J Med.* 1992;326:1380-1384.
10. Sacks HS, Chalmers TC, Smith H Jr. Randomized versus historical assignment in controlled clinical trials. *Am J Med.* 1983;309:1353-1361.
11. Chalmers TC, Celano P, Sacks HS, Smith H Jr. Bias in treatment assignment in controlled clinical trials. *N Engl J Med.* 1983;309:1358-1361.
12. Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy, I: medical. *Stat Med.* 1989;8:441-454.
13. Emerson JD, Burdick E, Hoaglin DC, et al. An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. *Controlled Clin Trials.* 1990; 11:339-352.
14. Coronary Drug Project Research Group. Influence of adherence treatment and response of cholesterol on mortality in the Coronary Drug Project. *N Engl J Med.* 1980;303:1038-1041.
15. Asher WL, Harper HW. Effect of human chorionic gonadotropin on weight loss, hunger, and feeling of well-being. *Am J Clin Nutr.* 1973; 26:211-218.
16. Hogarty GE, Goldberg SC. Drug and sociotherapy in the aftercare of schizophrenic patients: one-year relapse rates. *Arch Gen Psychiatry.* 1973;28:54-64.
17. Fuller R, Roth H, Long S. Compliance with disulfiram treatment of alcoholism. *J Chronic Dis.* 1983;36:161-170.
18. Pizzo PA, Robichaud KJ, Edwards BK, Schumaker C, Kramer BS, Johnson A. Oral antibiotic prophylaxis in patients with cancer: a double-blind randomized placebo-controlled trial. *J Pediatr.* 1983; 102:125-133.
19. Horwitz RI, Viscoli CM, Berkman L, et al. Treatment adherence and risk of death after myocardial infarction. *Lancet.* 1990;336:542-545.
20. Guyatt GH, Pugsley SO, Sullivan MJ, et al. Effect of encouragement on walking test performance. *Thorax.* 1984;39:818-822.

# Users' Guides to the Medical Literature

## II. How to Use an Article About Therapy or Prevention

## B. What Were the Results and Will They Help Me in Caring for My Patients?

Gordon H. Guyatt, MD, MSc; David L. Sackett, MD, MSc; Deborah J. Cook, MD, MSc; for the Evidence-Based Medicine Working Group

### CLINICAL SCENARIO

You are a general internist who is asked to see a 65-year-old man with controlled hypertension and a 6-month history of atrial fibrillation resistant to cardioversion. Although he has no evidence for valvular or coronary heart disease, the family physician who referred him to you wants your advice on whether the benefits of long-term anticoagulants (to reduce the risk of embolic stroke) outweigh their risks (of hemorrhage from anticoagulant therapy). The patient shares these concerns and doesn't want to receive a treatment that would do more harm than good. You know that there have been randomized trials of warfarin for nonvalvular atrial fibrillation and decide that you'd better review one of them.

From the Departments of Medicine and Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario.

A complete list of the members (with affiliations) of the Evidence-Based Medicine Working Group appears in the first article of this series (*JAMA*. 1993;270:2093-2095). The following members contributed to this article: Gordon Guyatt (Chair), MD, MSc; Eric Bass, MD, MPH; Patrick Brill-Edwards, MD; George Browman, MD, MSc; Deborah Cook, MD, MSc; Michael Farkouh, MD; Hertzel Gerstein, MD, MSc; Brian Haynes, MD, MSc, PhD; Robert Hayward, MD, MPH; Anne Holbrook, MD, PharmD, MSc; Roman Jaeschke, MD, MSc; Elizabeth Juniper, MCSP, MSc; Andreas Laupacis, MD, MSc; Hui Lee, MD, MSc; Mitchell Levine, MD, MSc; Virginia Moyer, MD, MPH; Jim Nishikawa, MD; Andrew Oxman, MD, MSc, FACPM; Ameen Patel, MD; John Philbrick, MD; W. Scott Richardson, MD; Stephane Sauve, MD, MSc; David Sackett, MD, MSc; Jack Sinclair, MD; K.S. Trout, FRCE; Peter Tugwell, MD, MSc; Sean Tunis, MD, MSc; Stephen Walter, PhD; John Williams Jr, MD, MHS; and Mark Wilson, MD, MPH.

Reprint requests to McMaster University Health Sciences Centre, 1200 Main St W, Room 2C12, Hamilton, Ontario, Canada L8N 3Z5 (Dr Guyatt).

### THE SEARCH

The ideal article addressing this clinical problem would include patients with nonvalvular atrial fibrillation and would compare the effect of warfarin and a control treatment, ideally a placebo, on the risk of emboli (including embolic stroke) and also on the risk of the complications of anticoagulation. Randomized, double-blind studies would provide the strongest evidence.

In the software program GRATEFUL MED you select a Medical Subject Heading (MeSH) that identifies your population, "atrial fibrillation," another that specifies the intervention, "warfarin," and a third that specifies the outcome of interest, "stroke" (which the software automatically converts to "explode cerebrovascular disorders" meaning that all articles indexed under cerebrovascular disorders or its subheadings are potential targets of the search), while restricting the search to English-language studies. To ensure that, at least on your first pass, you identify only the highest quality studies, you include the methodological term "randomized controlled trial (PT)" (PT stands for publication type). The search yields nine articles. Three are editorials or commentaries, one addresses prognosis, and one focuses on quality of life for patients receiving anticoagulants. You decide to read the most recent of the four randomized trials.[1]

Reading the study, you find it meets the validity criteria you learned about in a prior article in this series.[2] To answer your patient's and the referring physician's concerns, however, you need to delve further into the relation between benefits and risks.

### INTRODUCTION

The previous article in this series dealt with whether a study of effectiveness of therapy was valid (Table 1). In this installment, we will show you how to proceed further to understand and use the results of valid studies of therapeutic interventions. We have summarized calculations in the Tables for easy reference.

#### What Were the Results?

**How Large Was the Treatment Effect?**—Most frequently, randomized clinical trials carefully monitor how often patients experience some adverse event or outcome. Examples of these dichotomous outcomes (yes or no outcomes that either happen or don't happen) include cancer recurrence, myocardial infarction, and death. Patients either do or do not suffer an event, and the article reports the proportion of patients who develop such events. Consider, for example, a study in which 20% (0.20) of a control group died, but only 15% (0.15) of those receiving a new treatment died. How might these results be expressed? Table 2 provides a summary of ways of presenting the effects of therapy.

One way would be as the absolute difference (known as the *absolute risk reduction* or risk difference), between the proportion who died in the control group (X) and the proportion who died in the treatment group (Y), or X − Y =

0.20−0.15=0.05. Another way to express the impact of treatment would be as a *relative risk* (RR): the risk of events among patients receiving the new treatment, relative to that among controls, or Y/X =0.15/0.20=0.75.

The most commonly reported measure of dichotomous treatment effects is the complement of this RR, and is called the *relative risk reduction* (RRR). It is expressed as a percent: $[1-(Y/X)] \times 100\% = [1-0.75] \times 100\% = 25\%$. An RRR of 25% means that the new treatment reduced the risk of death by 25% relative to that occurring among control patients; the greater the RRR, the more effective the therapy.

**How Precise Was the Estimate of Treatment Effect?**—The true risk reduction can never be known; all we have is the estimate provided by rigorous controlled trials, and the best estimate of the true treatment effect is that observed in the trial. This estimate is called a "point estimate" in order to remind us that although the true value lies somewhere in its neighborhood, it is unlikely to be precisely correct. Investigators tell us the neighborhood within which the true effect likely lies by the statistical strategy of calculating confidence intervals (CIs).[3]

We usually (though arbitrarily) use the 95% CI, which can be simply interpreted as defining the range that includes the true RRR 95% of the time. You'll seldom find the true RRR toward the extremes of this interval, and you'll find the true RRR beyond these extremes only 5% of the time, a property of the CI that relates closely to the conventional level of "statistical significance" of $P<.05$. We illustrate the use of CIs in the following examples.

If a trial randomized 100 patients each to treatment and control groups, and there were 20 deaths in the control group and 15 deaths in the treatment group, the authors would calculate a point estimate for the RRR of 25%: X=20/100 or 0.20, Y=15/100 or 0.15, and $[1-(Y/X)] \times 100\% = [1-0.75] \times 100 = 25\%$. You might guess, however, that the true RRR might be much smaller or much greater than this 25%, based on a difference of just five deaths. In fact, you surmise that the treatment might provide no benefit (an RRR of 0%) or even harm (a negative RRR). And you would be right—in fact, these results are consistent with both an RRR of −38% (that is, patients given the new treatment might be 38% more likely to die than control patients), and an RRR of nearly 59% (that is, patients subsequently receiving the new treatment might have a risk of dying almost 60% less than that of the risk in those who are not treated). In other words, the 95% CI on this RRR is −38% to 59%, and the trial really hasn't helped us decide whether to offer the new treatment. What sort of study would be more helpful?

What if the trial enrolled not 100 patients per group, but 1000 patients per group, and observed the same event rates as before, so that there were 200 deaths in the control group (X=200/1000=0.20) and 150 deaths in the treatment group (Y=150/1000=0.15). Again, the point estimate of the RRR is 25%: $[1-(Y/X)] \times 100\% = [1-(0.15/0.20)] \times 100\% = 25\%$. In this larger trial, you might think that the true reduction in risk is much closer to 25% and, again, you would be right; the 95% CI on the RRR for this set of results is all on the positive side of 0 and runs from 9% to 41%.

What these examples show is that the larger the sample size of a trial, the larger the number of outcome events and the greater our confidence that the true RRR (or any other measure of efficacy) is close to what we have observed. In the second example above, the lowest plausible value for the RRR was 9% and the highest value 41%. The point estimate—in this case 25%—is the one value most likely to represent the true RRR. As one considers values farther and farther from the point estimate, they become less and less consistent with the observed RRR. By the time one crosses the upper or lower boundaries of the 95% CI, the values are extremely unlikely to represent the true RRR, given the point estimate (that is, the observed RRR).

The Figure represents the CIs around the point estimate of an RRR of 25% in these two examples, with a risk reduction of 0 representing no treatment effect. In both scenarios the point estimate of the RRR is 25%, but the CI is far narrower in the second scenario.

It is evident that the larger the sample size, the narrower the CI. When is the sample size big enough?[4] In a "positive" study—a study in which the authors conclude that the treatment is effective—one can look at the lower boundary of the CI. In the second example, this lower boundary was +9%. If this risk reduction (the lowest that is consistent with the study results) is still important, or "clinically significant," (that is, it is large enough for you to want to offer it to your patient), then the investigators have enrolled sufficient patients. If, on the other hand, you do not consider an RRR of 9% clinically significant, then the study cannot be considered definitive, even if its results are statistically significant (that is, they exclude a risk reduction of 0). Keep in mind that the probability of the true value being less than the lower boundary of the CI is only 2.5%, and that a different criterion for the CI (a 90% CI, for instance) might be as or more appropriate.

The CI also helps us interpret "negative" studies in which the authors have concluded that the experimental treatment is no better than control therapy. All we need do is look at the upper boundary of the CI. If the RRR at this upper boundary would, if true, be clinically important, the study has failed to exclude an important treatment effect. In the first example we presented in this section, the upper boundary of the CI was an RRR of 59%. Clearly, if this represented the truth, the benefit of the treatment would be substantial, and we would conclude that although the investigators had failed to prove that experimental treatment was better than placebo, they also had failed to prove that it was not; they could not exclude a large, positive treatment effect. Once again the clinician must bear in mind the proviso about the arbitrariness of the choice of 95% boundaries for the CI. A reason-

Table 1.—Readers' Guides for an Article About Therapy

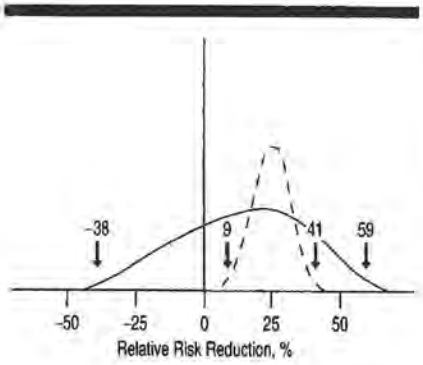
**Are the results of the study valid?**
- Primary guides:
  - Was the assignment of patients to treatments randomized?
  - Were all patients who entered the trial properly accounted for and attributed at its conclusion?
    - Was follow-up complete?
    - Were patients analyzed in the groups to which they were randomized?
- Secondary guides:
  - Were patients, health workers, and study personnel "blind" to treatment?
  - Were the groups similar at the start of the trial?
  - Aside from the experimental intervention, were the groups treated equally?

**What were the results?**
- How large was the treatment effect?
- How precise was the estimate of the treatment effect?

**Will the results help me in caring for my patients?**
- Can the results be applied to my patient care?
- Were all clinically important outcomes considered?
- Are the likely treatment benefits worth the potential harms and costs?

Table 2.—Introducing Some Measures of the Effects of Therapy

| Risk without therapy (baseline risk): X | Risk with therapy: Y | Absolute risk reduction (risk difference): X−Y | Relative risk: Y/X | Relative risk reduction (RRR): [1−(Y/X)]×100% or [(X−Y)/X]×100% | 95% confidence interval for the RRR: |
|---|---|---|---|---|---|
| 20/100=0.20 or 20% | 15/100=0.15 or 15% | 0.20−0.15=0.05 | 0.15/0.20=0.75 | [1−0.75]×100%=25%<br>[0.05/0.20]×100%=25% | −38% to +59% |




The solid line represents the confidence interval around the first example in which there were 100 patients per group and the number of events in the active and control groups were two and four, respectively. The broken line represents the confidence interval around the second example in which there were 1000 patients per group and the number of events in the active and control groups were 20 and 40, respectively.

able alternative, a 90% CI, would be somewhat narrower.

What can the clinician do if the CI around the RRR is not reported in the article? There are three approaches, and we present them in order of increasing complexity. The easiest approach is to examine the *P* value. If the *P* value is exactly .05, then the lower bound of the 95% confidence limit for the RRR has to lie exactly at 0 (an RR of 1), and you cannot exclude the possibility that the treatment has no effect. As the *P* value decreases below .05, the lower bound of the 95% confidence limit for the RRR rises above 0.

A second approach, involving some quick mental arithmetic or a pencil and paper, can be used when the article includes the value for the standard error (SE) of the RRR (or of the RR). This is because the upper and lower boundaries of the 95% CI for an RRR are the point estimate plus and minus twice this SE.

The third approach involves calculating the CIs yourself[5] or asking the help of someone else (a statistician, for instance) to do so. Once you obtain the CIs, you know how high and low the RRR might be (that is, you know the precision of the estimate of the treatment effect) and can interpret the results as described above.

Not all randomized trials have dichotomous outcomes, nor should they. For example, a new treatment for patients with chronic lung disease may focus on increasing their exercise capacity. Thus, in a study of respiratory muscle training for patients with chronic airflow limitation, one primary outcome measured how far patients could walk in 6 minutes in an enclosed corridor.[6] This 6-minute walk improved from an average of 406 to 416 meters (up 10 meters) in the experimental group receiving respiratory muscle training, and from 409 to 429 (up 20 meters) in the control group. The point estimate for improvement in the 6-minute walk due to respiratory muscle training therefore was negative, at −10 meters (or a 10-meter difference in favor of the control group).

Here too you should look for the 95% CIs around this difference in changes in exercise capacity and consider their implications. The investigators tell us that the lower boundary of the 95% CI was −26 meters (that is, the results are consistent with a difference of 26 meters in favor of the control treatment) and the upper boundary was +5 meters. Even in the best of circumstances, adding 5 meters to the 400 recorded at the start of the trial would not be important to the patient, and this result effectively excludes a clinically significant benefit of respiratory muscle training as applied in this study.

Having determined the magnitude and precision of the treatment effect, readers now can turn to the final question of how to apply the article's results to their patients and clinical practice.

## Will the Results Help Me in Caring for My Patients?

**Can the Results Be Applied to My Patient Care?**—The first issue to address is how confident you are that you can apply the results to a particular patient or patients in your practice. If the patient would have been enrolled in the study had she been there—that is, she meets all the inclusion criteria, and doesn't violate any of the exclusion criteria—there is little question that the results are applicable. If this is not the case, and she would not have been eligible for the study, judgment is required. The study result probably applies even if, for example, she was 2 years too old for the study, had more severe disease, had previously been treated with a competing therapy, or had a comorbid condition. A better approach than rigidly applying the study's inclusion and exclusion criteria is to ask whether there is some compelling reason why the results should *not* be applied to the patient. A compelling reason usually won't be found, and most often you can generalize the results to your patient with confidence.

A final issue arises when our patient fits the features of a subgroup of patients in the trial report. In articles reporting the results of a trial (especially when the treatment doesn't appear to be efficacious for the average patient), the authors may have examined a large number of subgroups of patients at different stages of their illness, with different comorbid conditions, with different ages at entry, and the like. Quite often these subgroup analyses were not planned ahead of time, and the data are simply "dredged" to see what might turn up. Investigators may sometimes overinterpret these "data-dependent" analyses as demonstrating that the treatment really has a different effect in a subgroup of patients—those who are older or sicker, for instance, may be held up as benefitting substantially more or less than other subgroups of patients in the trial. You can find guides for deciding whether to believe these subgroup analyses,[7] summarized as follows: the treatment is really likely to benefit the subgroup more or less than the other patients if the difference in the effects of treatment in the subgroups (1) is large; (2) is very unlikely to occur by chance; (3) results from a analysis specified as a hypothesis before the study began; (4) was one of only a very few subgroup analyses that were carried out; and (5) is replicated in other studies. To the extent that the subgroup analysis fails these criteria, clinicians should be increasingly skeptical about applying them to their patients.

**Were All Clinically Important Outcomes Considered?**—Treatments are indicated when they provide important benefits. Demonstrating that a bronchodilator produces small increments in forced expired volume in patients with chronic airflow limitation, that a vasodilator improves cardiac output in heart failure patients, or that a lipid-lowering agent improves lipid profiles does not necessarily provide a sufficient reason for administering these drugs. What is required is evidence that the treatments improve outcomes that are important to patients, such as reducing shortness of breath during the activities required for daily living, avoiding hospitalization for heart failure, or decreasing the risk of myocardial infarction. We can consider forced expired volume in 1 second, cardiac output, and the lipid profile "substitute end points." That is, the authors have substituted these physiologic measures for the important outcomes (shortness of breath, hospitalization, or myocardial infarction), usually because to confirm benefit on the latter they would have had to enroll many more patients and followed them for far longer periods of time.

A dramatic recent example of the danger of substitute end points was found in the evaluation of the usefulness of antiarrhythmic drugs following myocardial infarction. Because such drugs had been shown to reduce abnormal ventricular depolarizations (the substitute end points) in the short run, it made

Table 3.—Two Men With Contrasting Prognoses Following Myocardial Infarction

| If the risk of death at 1 year without therapy (baseline risk) is: X | And the relative risk of death with therapy (a β blocker) is: Y/X | And the relative risk reduction is: [1−(Y/X)]×100% or [(X−Y)/X]×100% | Then the risk of death with treatment is: Y | And the absolute risk reduction is: X−Y | And the number needed to be treated to prevent one event is: 1/(X−Y) |
|---|---|---|---|---|---|
| 1% or 0.01 | 75% or 0.75 | 25% | 0.01×0.75=0.0075 | 0.01−0.0075=0.0025 | 1/0.0025=400 |
| 10% or 0.10 | 75% or 0.75 | 25% | 0.10×0.75=0.075 | 0.10−0.075=0.025 | 1/0.025=40 |

Table 4.—Incorporating Side Effects Into the Number Needed to Be Treated

| If the risk of death at 1 year without therapy (baseline risk) is: X | And the risk of death with propranolol is: Y | Then the absolute risk reduction is: X−Y | And the number needed to be treated to prevent one event is: 1/(X−Y) | And if the incidence of clinically important fatigue on propranolol is: | Then the number of fatigued patients per life saved is: |
|---|---|---|---|---|---|
| 1% or 0.01 | 0.01×0.75=0.0075 | 0.01−0.0075=0.0025 | 1/0.0025=400 | 10% or 0.10 | 400×0.1=40 |
| 10% or 0.10 | 0.10×0.75=0.075 | 0.10−0.075=0.025 | 1/0.025=40 | | 40×0.1=4 |

sense that they should reduce the occurrence of life-threatening arrhythmias in the long run. A group of investigators performed randomized trials on three agents (encainide, flecainide, and moricizine) previously shown to be effective in suppressing the substitute end point of abnormal ventricular depolarizations in order to determine whether they reduced mortality in patients with asymptomatic or mildly symptomatic arrhythmias following myocardial infarction. The investigators had to stop the trials when they discovered that mortality was substantially higher in patients receiving antiarrhythmic treatment than in those receiving a placebo.[8,9] Clinicians relying on the substitute end point of arrhythmia suppression would have continued to administer the three drugs, to the considerable detriment of their patients.

Even when investigators report favorable effects of treatment on one clinically important outcome, clinicians must take care that there are no deleterious effects on other outcomes. For instance, as this series was in preparation, the controversy continued over whether reducing lipids unexpectedly increases noncardiovascular causes of death.[10] Cancer chemotherapy may lengthen life but may also decrease its quality. Finally, surgical trials often document prolonged life for those who survive the operation (yielding higher 3-year survival in those receiving surgery), but an immediate risk of dying during or shortly after surgery. Accordingly, users of the reports of surgical trials should look for information on immediate and early mortality (typically higher in the surgical group) in addition to longer-term results.

**Are the Likely Treatment Benefits Worth the Potential Harm and Costs?**—If the article's results are generalizable to your patient and its outcomes are important, the next question concerns whether the probable treatment benefits are worth the effort that you and your patient must put into the enterprise. A 25% reduction in the risk of death may sound quite impressive, but its impact on your patient and practice may nevertheless be minimal. This notion is illustrated using a concept called "number needed to treat" (NNT).[11]

The impact of a treatment is related not only to its RRR, but also to the risk of the adverse outcome it is designed to prevent. β-Blockers reduce the risk of death following myocardial infarction by approximately 25%, and this RRR is consistent across subgroups, including those at higher and lower "baseline" risk of recurrence and death when they are untreated. Table 3 considers two patients with recent myocardial infarctions.

First, consider a 40-year-old man with a small infarct, normal exercise capacity, and no sign of ventricular arrhythmia who is willing to stop smoking, begin exercising, lose weight, and take aspirin daily. This individual's risk of death in the first year after infarction may be as low as 1%. β-Blockers would reduce this risk by a quarter, to 0.75%, for an absolute risk reduction of 0.25% or 0.0025. The inverse of this absolute risk reduction (that is, 1 divided by the absolute risk reduction) equals the number of such patients we'd have to treat in order to prevent one event (in this case, to prevent one death following a mild heart attack in a low-risk patient). In this case, we would have to treat 400 such patients for 1 year to save a single life (1/0.0025=400).

An older man with limited exercise capacity and frequent ventricular extrasystoles who continues to smoke following his infarction may have a risk of dying in that next year as high as 10%. A 25% risk reduction for death in such a high-risk patient generates an absolute risk reduction of 2.5% or 0.025, and we would have to treat only 40 such individuals for 1 year to save a life (1/0.025=40).

These examples underscore a key element of the decision to start therapy: before deciding on treatment, we must consider our patient's risk of the adverse event if left untreated. For any given RRR, the higher the probability that a patient will experience an adverse outcome if we don't treat, the more likely the patient will benefit from treatment, and the fewer such patients we need to treat to prevent one event. Thus, both patients and our own clinical efficiency benefit when the NNT to prevent an event is low.

We might not hesitate to treat even as many as 400 patients to save one life if the treatment were cheap, easy to apply and comply with, and safe. In reality, however, treatments usually are expensive and they carry risks. When these risks or adverse outcomes are documented in trial reports, users can apply the NNT to judge both the relative benefits and costs of therapy. If, for instance, β-blockers cause clinically important fatigue in 10% of the patients who use them, the NNT to cause fatigue is 1/0.10 or 10. This is shown in Table 4, where it is seen that a policy of treating low-risk patients after myocardial infarction (NNT=400 to prevent one death) will result in 40 being fatigued for every life saved. On the other hand, a policy of treating just high-risk patients will result in four being fatigued for every life saved.

Clinicians don't, however, treat groups of patients uniformly. Rather, we consider individual responses and tailor our therapy accordingly. One response to the problem of common, relatively minor side effects (such as fatigue) is to discontinue therapy in patients suffering from that problem. If we think of fatigued low-risk patients as a group, we would make 400 patients fatigued to save a life, a trade-off that probably wouldn't be worth it. By discontinuing treatment in these people, we can treat the remainder without making anyone fatigued.

We cannot apply this approach, however, to severe, episodic events. Examples include the risk of bleeding in patients given anticoagulants, throm-

Table 5.—Summary of the Effect of Warfarin Therapy on Patients With Nonvalvular Atrial Fibrillation*

| If the risk of stroke at 1 year without therapy (baseline risk) is: X | And the risk of stroke with treatment is: Y | Then the absolute risk reduction is: X−Y | And the number needed to be treated to prevent one stroke is: 1/(X−Y) (95% confidence interval) | And if the incidence of clinically important bleeding on warfarin is: | Then the number of bleeds per stroke prevented is: |
|---|---|---|---|---|---|
| 0.043 | 0.009 | 0.043−0.009=0.034 | 1/0.034=30 (26 to 45) | 0.01 | 29×0.01=0.29 |

*Data from Ezekowitz et al.[1]

bolytic agents, or aspirin, or the risk of rare but devastating drug reactions. In each of these examples the number of adverse events per life saved (or, if the events are rare enough, the number of lives saved per adverse event) can provide a compelling picture of the trade-offs associated with the intervention.

## RESOLUTION OF THE SCENARIO

In the randomized trial of warfarin in nonvalvular atrial fibrillation that you selected for reading (Ezekowitz et al[1]), 260 patients received warfarin and 265 received placebo. The results are summarized in Table 5.

Over the next 1½ years, just four of the former (0.9% per year), but 19 of the latter (4.3% per year) suffered cerebral infarction. Thus, the RRR is (0.043−0.009)/0.043=79%, the absolute risk reduction is 0.043−0.009=0.034, and the NNT to prevent one stroke is 1/0.034 =29 (or approximately 30). Applying CIs to this NNT, the NNT could be (using the lower boundary of the CI around the RRR, which was 0.52) as great as 45, or (using the upper boundary of the CI around the RRR, which was 0.90) as few as 26. Now, you know that warfarin is a potentially dangerous drug, and that about 1% of patients on this treatment will suffer clinically important bleeding as a result of treatment each year.[12] Therefore, there will be one episode of bleeding in every 100 treated patients, and if the NNT to prevent a stroke is 30, then for every three strokes prevented, one major episode of bleeding would occur. If the lower boundary of the CI for the benefit of oral anticoagulants represents the truth, the NNT is 45 and for every two strokes prevented, one would cause a major episode of bleeding; if, on the other hand, the upper boundary represents the truth, the NNT is 26 and approximately four strokes would be prevented for every major bleeding episode. The true risk-benefit ratio probably lies somewhere between these extremes, closer to that asssociated with the point estimate.

And what about the woman with lupus nephritis, whose plight, described in part A of this two-part essay, prompted us to find a trial of adding plasmapheresis to a regimen of prednisone and cyclophosphamide? Unfortunately, although plasmapheresis did produce sharp declines in the substituted end points of anti-dsDNA antibodies and cryoprecipitable immune complexes, the trial did not find any benefit from plasmapheresis in the clinically important measures of renal failure or mortality. When a careful statistical analysis of the emerging data suggested little hope of ever showing clinical benefit, the trial was stopped.

## CONCLUSION

Having read the introduction to this series and the two articles on using articles about therapy, we hope that you are developing a sense of how to use the medical literature to resolve a treatment decision. First, define the problem clearly, and use one of a number of search strategies to obtain the best available evidence. Having found an article relevant to the therapeutic issue, assess the quality of the evidence. To the extent that the quality of the evidence is poor, any subsequent inference (and the clinical decision it generates) will be weakened. If the quality of the evidence is adequate, determine the range within which the true treatment effect likely falls. Then, consider the extent to which the results are generalizable to the patient at hand, and whether the outcomes that have been measured are important. If the generalizability is in doubt, or the importance of the outcomes questionable, support for a treatment recommendation will be weakened. Finally, by taking into account the patient's risk of adverse events, assess the likely results of the intervention. This involves a balance sheet looking at the probability of benefit and the associated costs (including monetary costs, and issues such as inconvenience) and risks. The bottom line of the balance sheet will guide your treatment decision.

While this may sound like a challenging route to deciding on treatment, it is what clinicians implicitly do each time they administer therapy.[13] Making the process explicit and being able to apply guidelines to help assess the strength of evidence will, we think, result in better patient care.

### References


1. Ezekowitz MD, Bridgers SL, James KE, et al, for the Veterans Affairs Stroke Prevention in Nonrheumatic Atrial Fibrillation Investigators. Warfarin in the prevention of stroke associated with nonrheumatic atrial fibrillation. *N Engl J Med.* 1992; 327:1406-1412.
2. Guyatt GH, Sackett DL, Cook DJ, for the Evidence-Based Working Group. Users' guides to the medical literature, II: how to use an article about therapy or prevention, A: are the results of the study valid? *JAMA.* 1993;270:2598-2601.
3. Altman DG, Gore SM, Gardner MJ, Pocock SJ. Statistical guidelines for contributors to medical journals. In: Gardner MJ, Altman DG, eds. *Statistics With Confidence: Confidence Intervals and Statistical Guidelines.* London, England: British Medical Journal; 1989:83-100.
4. Detsky AS, Sackett DL. When was a 'negative' trial big enough? how many patients you needed depends on what you found. *Arch Intern Med.* 1985; 145:709-715.
5. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. 2nd ed. *Clinical Epidemiology: A Basic Science for Clinical Medicine.* Boston, Mass: Little Brown & Co Inc; 1991:218.
6. Guyatt GH, Keller J, Singer J, Halcrow S, Newhouse M. A controlled trial of respiratory muscle training in patients with chronic airflow limitation. *Thorax.* 1992;47:598-602.
7. Oxman AD, Guyatt GH. A consumer's guide to subgroup analysis. *Ann Intern Med.* 1992;116:78-84.
8. Echt DS, Liebson PR, Mitchell LB, et al. Mortality and morbidity in patients receiving encainide, flecainide, or placebo: the Cardiac Arrhythmia Suppression Trial. *N Engl J Med.* 1991;324:781-788.
9. The Cardiac Arrhythmia Suppression Trial II Investigators. Effect of the antiarrhythmic agent moricizine on survival after myocardial infarction. *N Engl J Med.* 1992;327:227-233.
10. Muldoon ME, Manuck SB, Matthews KA. Lowering cholesterol concentrations and mortality: a quantitative review of primary prevention trials. *BMJ.* 1990;301:309-314.
11. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *N Engl J Med.* 1988;318: 1728-1733.
12. Levine M, Hirsch J, Ladefeld S, Raskob G. Hemorrhagic complications of anticoagulant treatment. *Chest.* 1992;102 (suppl):352S-363S.
13. Eddy DM. Clinical policies and the quality of clinical practice. *N Engl J Med.* 1982;307:343-347.

# Users' Guides to the Medical Literature

## III. How to Use an Article About a Diagnostic Test

## A. Are the Results of the Study Valid?

Roman Jaeschke, MD, MSc; Gordon Guyatt, MD, MSc; David L. Sackett, MD, MSc;
for the Evidence-Based Medicine Working Group

### CLINICAL SCENARIO

You are a medical consultant asked by a surgical colleague to see a 78-year-old woman, now 10 days after abdominal surgery, who has become increasingly short of breath over the last 24 hours. She has also been experiencing what she describes as chest discomfort, which is sometimes made worse by taking a deep breath (but sometimes not). Abnormal findings on physical examination are restricted to residual tenderness in the abdomen and scattered crackles at both lung bases. Chest roentgenogram reveals a small right pleural effusion, but this is the first roentgenogram since the operation. Arterial blood gases show a $Po_2$ of 70 mm Hg, with a saturation of 92%. The electrocardiogram shows only nonspecific changes.

You suspect that the patient, despite receiving 5000 U of heparin twice a day, may have had a pulmonary embolus (PE). You request a ventilation-perfusion scan (V/Q scan), and the result reported to the nurse over the telephone is "intermediate probability" for PE. Though still somewhat uncertain about the diagnosis, you order full anticoagulation. Although you have used the V/Q scan frequently in the past and think you have a fairly good notion of how to use the results, you realize that your understanding is based on intuition and local practice rather than on the properties of V/Q scanning from the original literature. Consequently, on your way to the nuclear medicine department to review the scan, you stop off in the library.

### THE SEARCH

Your plan is to find a study that will tell you about the properties of V/Q scanning as it applies to your clinical practice in general and this patient in particular. You are familiar with the software program GRATEFUL MED and use this for your search. The program provides a listing of Medical Subject Headings (MeSH), and your first choice is "pulmonary embolism." Since there are 1749 articles with that MeSH heading published between 1989 and 1992 (the range of your search), you are going to have to pare down your search. You choose two strategies: you will pick only articles that have "radionuclide imaging" as a subheading and also have the associated MeSH heading "comparative study" (since you will need a study comparing V/Q scanning with some reference standard). This search yields 31 articles, of which you exclude 11 that evaluate new diagnostic techniques, nine that relate to the diagnosis and treatment of deep venous thrombosis, and one that examines the natural history of PE. The remaining 11 address V/Q scanning in PE. One, however, is an editorial; four are limited in their scope (dealing with perfusion scans only, with situations in which the diagnostic workup should begin with pulmonary angiography, or with a single perfusion defect). Of the remainder, the Prospective Investigation of Pulmonary Embolism Diagnosis (PIOPED) study[1] catches your eye, both because it is in a widely read journal with which you are familiar and because it is referred to in the titles of several of the other articles. You print the abstract of this article and find it includes the following piece of information: among people with an intermediate result of the V/Q scan, 33% had PE. You conclude you have made a good choice and retrieve the article from the library shelves.

This article in the "Users' Guides to the Medical Literature" series and the one that follows will demonstrate an approach to making optimal use of the article.

### INTRODUCTION

Clinicians regularly confront dilemmas when ordering and interpreting diagnostic tests. The continuing proliferation of medical technology renders the clinician's ability to assess articles about diagnostic tests ever more important. Accordingly, this article will present the principles of efficiently assessing articles about diagnostic tests and optimally using the information they provide. Once you decide, as was illustrated in the clini-

From the Departments of Medicine and Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario.

A complete list of the members (with affiliations) of the Evidence-Based Medicine Working Group appears in the first article of this series (*JAMA*. 1993;270:2093-2095). The following members contributed to this article: Gordon Guyatt (Chair), MD, MSc; Eric Bass, MD, MPH; Patrick Brill-Edwards, MD; George Browman, MD, MSc; Deborah Cook, MD, MSc; Michael Farkouh, MD; Hertzel Gerstein, MD, MSc; Brian Haynes, MD, MSc, PhD; Robert Hayward, MD, MPH; Anne Holbrook, MD, PharmD, MSc; Roman Jaeschke, MD, MSc; Elizabeth Juniper, MCSP, MSc; Hui Lee, MD, MSc; Mitchell Levine, MD, MSc; Virginia Moyer, MD, MPH; Jim Nishikawa, MD; Andrew Oxman, MD, MSc, FACPM; Ameen Patel, MD; John Philbrick, MD; W. Scott Richardson, MD; Stephane Sauve, MD, MSc; David Sackett, MD, MSc; Jack Sinclair, MD; K. S. Trout, FRCE; Peter Tugwell, MD, MSc; Sean Tunis, MD, MSc; Stephen Walter, PhD; and Mark Wilson, MD, MPH.

Reprint requests to McMaster University Health Sciences Centre, 1200 Main St W, Room 2C12, Hamilton, Ontario, Canada L8N 3Z5 (Dr Guyatt).

cal scenario with the PIOPED article, that an article is potentially relevant (that is, the title and abstract suggest the information is directly relevant to the patient problem you are addressing), you can invoke the same three questions that we suggested in the "Introduction" and the articles on therapy[2-4] (Table 1).

### Are the Results of the Study Valid?

Whether one can believe the results of a study is determined by the methods used to carry it out. To say that the results are valid implies that the accuracy of the diagnostic test, as reported, is close enough to the truth to render the further examination of the study worthwhile. First, you must determine if you can believe the results of the study by considering how the authors assembled their patients and how they applied the test and an appropriate reference (or "gold" or "criterion") standard to the patients.

### What Are the Results of the Study?

If you decide that the study results are valid, the next step is to determine the diagnostic test's accuracy. This is done by examining (or calculating for yourself) the test's likelihood ratios (often referred to as the test's "properties").

### Will the Results Help Me in Caring for My Patients?

The third step is to decide how to use the test, both for the individual patient and for your practice in general. Are the results of the study generalizable—ie, can you apply them to this particular patient and to the kind of patients you see most often? How often are the test results likely to yield valuable information? Does the test provide additional information above and beyond the history and physical examination? Is it less expensive or more easily available than other diagnostic tests for the same target disorder? Ultimately, are patients better off if the test is used?

In this article we deal with the first question in detail, while in the next article in the series we address the second and third questions. We use the PIOPED article to illustrate the process.

In the PIOPED study, 731 consenting patients suspected of having PE underwent both V/Q scanning and pulmonary angiography. The pulmonary angiogram was considered to be the best way to prove whether a patient really had a PE and therefore was the reference standard. Each angiogram was interpreted as showing one of three results: PE present, PE uncertain, or PE absent. The accuracy of the V/Q scan was compared with the angiogram, and the V/Q scan results were reported in one of four categories: high probability (for PE), intermediate probability, low probability, or near normal or normal. The comparisons of the V/Q scans and angiograms are shown in Tables 2 and 3. We'll get to the differences between these tables later; for now, let's apply the first of the three questions to this article.

Table 1.—Evaluating and Applying the Results of Studies of Diagnostic Tests

| |
|---|
| **Are the results of the study valid?** |
| Primary guides: |
| Was there an independent, blind comparison with a reference standard? |
| Did the patient sample include an appropriate spectrum of patients to whom the diagnostic test will be applied in clinical practice? |
| Secondary guides: |
| Did the results of the test being evaluated influence the decision to perform the reference standard? |
| Were the methods for performing the test described in sufficient detail to permit replication? |
| **What were the results?** |
| Are likelihood ratios for the test results presented or data necessary for their calculation provided? |
| **Will the results help me in caring for my patients?** |
| Will the reproducibility of the test result and its interpretation be satisfactory in my setting? |
| Are the results applicable to my patient? |
| Will the results change my management? |
| Will patients be better off as a result of the test? |

Table 2.—The Relationship Between the Results of Pulmonary Angiograms and Ventilation-Perfusion Scan Results in Patients With Successful Angiograms

| | Angiogram | |
|---|---|---|
| Scan Category | Pulmonary Embolus Present | Pulmonary Embolus Absent |
| High probability | 102 | 14 |
| Intermediate probability | 105 | 217 |
| Low probability | 39 | 199 |
| Near normal/normal | 5 | 50 |
| **Total** | **251** | **480** |

Table 3.—The Relationship Between the Results of Pulmonary Angiograms and Ventilation-Perfusion Scan Results*

| | Angiogram | |
|---|---|---|
| Scan Category | Pulmonary Embolus Present | Pulmonary Embolus Absent |
| High probability | 102 | 14 |
| Intermediate probability | 105 | 217 |
| Low probability | 39 | 273 |
| Near normal/normal | 5 | 126 |
| **Total** | **251** | **630** |

*Includes 150 patients with low probability and near normal/normal ventilation-perfusion scans, no (136) or uninterpretable (14) angiograms, and no clinically important thromboembolism on follow-up.

## ARE THE RESULTS OF THE STUDY VALID?

### Primary Guides

**Was There an Independent, Blind Comparison With a Reference Standard?**—The accuracy of a diagnostic test is best determined by comparing it with the "truth." Accordingly, readers must assure themselves that an appropriate reference standard (such as biopsy, surgery, autopsy, or long-term follow-up) has been applied to every patient, along with the test under investigation.[5] In the PIOPED study, the pulmonary angiogram was used as the reference standard and this was as "gold" as could be achieved without sacrificing the patients. In reading articles about diagnostic tests, if you can't accept the reference standard (within reason, that is—nothing is perfect!), then the article is unlikely to provide valid results for your purposes.

If you do accept the reference standard, the next question is whether the test results and the reference standard were assessed independently of each other (that is, by interpreters who were unaware of the results of the other investigation). Our own clinical experience shows us why this is important. Once we have been shown a pulmonary nodule on a computed tomographic scan, we see the previously undetected lesion on the chest roentgenogram; once we learn the results of the echocardiogram, we hear the previously inaudible cardiac murmur. The more likely it is that the interpretation of a new test could be influenced by knowledge of the reference standard result (or vice versa), the greater the importance of the independent interpretation of both. The PIOPED investigators did not state explicitly that the tests were interpreted blindly in the article. However, one could deduce from the effort they put into ensuring reproducible, independent readings that the interpreters were in fact blinded, and we have confirmed through correspondence with one of the authors that this was so. When such matters are in doubt, most authors are happy to clarify if directly contacted.

**Did the Patient Sample Include an Appropriate Spectrum of Patients to Whom the Diagnostic Test Will Be Applied in Clinical Practice?**—A diagnostic test is really useful only to the extent it distinguishes between target disorders or states that might otherwise be confused. Almost any test can distinguish the healthy from the severely affected; this ability tells us nothing about the clinical utility of a test. The true, pragmatic value of a test is therefore established only in a study that closely resembles clinical practice.

A vivid example of how the hopes raised with the introduction of a diagnostic test can be dashed by subsequent investigations comes from the story of carcinoembryonic antigen (CEA) in colorectal cancer. Carcinoembryonic antigen levels, when measured in 36 people with known advanced cancer of the co-

lon or rectum, were elevated in 35 of them. At the same time, much lower levels were found in normal people and in a variety of other conditions.[6] The results suggested that measurement of CEA levels might be useful in diagnosing colorectal cancer or even in screening for the disease. In subsequent studies of patients with less advanced stages of colorectal cancer (and, therefore, lower disease severity) and patients with other cancers or other gastrointestinal disorders (and, therefore, different but potentially confused disorders), the accuracy of CEA measurements plummeted, and the use of CEA levels for cancer diagnosis and screening was abandoned. Carcinoembryonic antigen is now recommended only as one element in the follow-up of patients with known colorectal cancer.[7]

In the PIOPED study, the whole spectrum of patients suspected of having PE were eligible and recruited, including those who entered the study with high, medium, and low clinical suspicion of PE. We thus may conclude that the appropriate patient sample was chosen.

### Secondary Guides

Once you are convinced that the article is describing an appropriate spectrum of patients who underwent the independent, blind comparison of a diagnostic test and a reference standard, most likely its results represent an unbiased estimate of the real accuracy of the test—that is, an estimate that doesn't systematically distort the truth. However, you can further reduce your chances of being misled by considering a number of other issues.

**Did the Results of the Test Being Evaluated Influence the Decision to Perform the Reference Standard?**—The properties of a diagnostic test will be distorted if its result influences whether patients undergo confirmation by the reference standard. This situation, sometimes called "verification bias"[8,9] or "work-up bias",[10,11] would apply, for example, when patients with suspected coronary artery disease and positive exercise tests were more likely to undergo coronary angiography (the reference standard) than those with negative exercise tests.

Verification bias was a problem for the PIOPED study; patients whose V/Q scans were interpreted as normal or near normal and low probability were less likely to undergo pulmonary angiography (69%) than those with more positive V/Q scans (92%). This is not surprising, since clinicians might be reluctant to subject patients with a low probability of PE to the risks of angiography. The results of the PIOPED study restricted to those patients with successful angiography are presented in Table 2.

Most articles would stop here, and readers would have to conclude that the magnitude of the bias resulting from different proportions of patients with high and low probability V/Q scans undergoing adequate angiography is uncertain but perhaps large. However, the PIOPED investigators applied a second reference standard to the 150 patients with low probability or normal/near normal scans who failed to undergo angiography (136 patients) or in whom angiographic interpretation was uncertain (14 patients): they would be judged to be free of PE if they did well without treatment. Accordingly, they followed every one of them for 1 year without treating them with anticoagulants. Not one of these patients developed clinically evident PE during this time, from which we can conclude that clinically important PE (if we define clinically important PE as requiring anticoagulation to prevent subsequent adverse events) was not present at the time they underwent V/Q scanning. When these 150 patients, judged free of PE by this second reference standard of a good prognosis without anticoagulant therapy, are added to the 480 patients with negative angiograms in Table 2, the result is Table 3. We hope you agree with us that the better estimate of the accuracy of V/Q scanning comes from Table 3, which includes the 150 patients who, from follow-up, did not have clinically important PE. Accordingly, we will use these data in subsequent calculations.

There were still another 50 patients with either high or intermediate probability scans who either did not undergo angiography or whose angiograms were uninterpretable. It is possible that these individuals could bias the results. However, they are a relatively small proportion of the population, and if their clinical characteristics are not clearly different from those who underwent angiography, it is unlikely that the test properties would differ systematically in this subpopulation. Therefore, we can proceed with relative confidence in the PIOPED results.

**Were the Methods for Performing the Test Described in Sufficient Detail to Permit Replication?**—If the authors have concluded that you should use a diagnostic test, they must tell you how to use it. This description should cover all issues that are important in the preparation of the patient (diet, drugs to be avoided, precautions after the test), the performance of the test (technique, possibility of pain), and the analysis and interpretation of its results.

Once the reader is confident that the article's results constitute an unbiased estimate of the test properties, she can determine exactly what (and how helpful) those test properties are. While not pristine (studies almost never are), we can strongly infer that the results are a valid estimate of the properties of the V/Q scan. We will describe how to interpret and apply the results in the next article of this series.

#### References


1. The PIOPED Investigators. Value of ventilation/perfusion scan in acute pulmonary embolism: results of the Prospective Investigation of Pulmonary Embolism Diagnosis (PIOPED). *JAMA*. 1990; 263:2753-2759.
2. Oxman AD, Sackett DL, Guyatt GH, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, I: how to get started. *JAMA*. 1993;270:2093-2095.
3. Guyatt GH, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, II: how to use an article about therapy or prevention, A: are the results of the study valid? *JAMA*. 1993;270:2598-2601.
4. Guyatt GH, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, II: how to use an article about therapy or prevention, B: what were the results and will they help me in caring for my patients? *JAMA*. 1994;271:59-63.
5. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical Epidemiology: A Basic Science for Clinical Medicine*. 2nd ed. Boston, Mass: Little Brown and Co; 1991:53-57.
6. Thomson DMP, Krupey J, Freedman SO, Gold P. The radioimmunoassay of circulating carcinoembryonic antigen of the human digestive system. *Proc Natl Acad Sci U S A*. 1969;64:161-167.
7. Bates SE. Clinical applications of serum tumor markers. *Ann Intern Med*. 1991;115:623-638.
8. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics*. 1983;39:207-215.
9. Gray R, Begg CB, Greenes RA. Construction of receiver operating characteristic curves when disease verification is subject to selection bias. *Med Decis Making*. 1984;4:151-164.
10. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med*. 1978;299:926-930.
11. Choi BCK. Sensitivity and specificity of a single diagnostic test in the presence of work-up bias. *J Clin Epidemiol*. 1992;45:581-586.

# Users' Guides to the Medical Literature

## III. How to Use an Article About a Diagnostic Test

## B. What Are the Results and Will They Help Me in Caring for My Patients?

Roman Jaeschke, MD, MSc; Gordon H. Guyatt, MD, MSc; David L. Sackett, MD, MSc; for the Evidence-Based Medicine Working Group

### CLINICAL SCENARIO

You are back where we put you in the previous article[1] on diagnostic tests in this series on how to use the medical literature: in the library studying an article that will guide you in interpreting ventilation-perfusion (V/Q) lung scans. Using the criteria in Table 1, you have decided that the Prospective Investigation of Pulmonary Diagnosis (PIOPED) study[2] will provide you with valid information. Just then, another physician comes looking for an article to help with the interpretation of V/Q scanning. Her patient is a 28-year-old man whose acute onset of shortness of breath and vague chest pain began shortly after completing a 10-hour auto trip. He experienced several episodes of similar discomfort in the past, but none this severe, and is very apprehensive about his symptoms. After a normal physical examination, electrocardiogram and chest radiograph, and blood gas measurements that show a $P_{CO_2}$ of 32 mm Hg and a $P_{O_2}$ of 82 mm Hg, your colleague has ordered a V/Q scan. The results are reported as an "intermediate-probability" scan.

You tell your colleague how you used GRATEFUL MED to find an excellent article addressing the accuracy of V/Q scanning. She is pleased that you found the article valid, and you agree to combine forces in applying it to both your patients.

In the previous article on diagnostic tests, we presented an approach to deciding whether a study was valid, and the results therefore worth considering. In this installment, we explore the next steps, which involve understanding and using the results of valid studies of diagnostic tests.

### WHAT ARE THE RESULTS?

#### Are Likelihood Ratios for the Test Results Presented or Data Necessary for Their Calculation Included?

**Pretest Probability.**—The starting point of any diagnostic process is the patient, presenting with a constellation of symptoms and signs. Consider the two patients who opened this exercise—the 78-year-old woman 10 days after surgery and the 28-year-old anxious man, both with shortness of breath and nonspecific chest pain. Our clinical hunches about the probability of pulmonary embolus (PE) as the explanation for these two patients' complaints, that is, their pretest probabilities, are very different: the probability in the elderly woman is high, and in the young man the probability is low. As a result, even if both have intermediate-probability V/Q scans, subsequent management is likely to differ. One might well treat the elderly woman but order additional investigations in the young man.

Two conclusions emerge from this line of reasoning. First, whatever the results of the V/Q scan, they do not tell us whether PE is present. What they do accomplish is to modify the pretest probability of PE, yielding a new posttest probability. The direction and magnitude of this change from pretest to posttest probability are determined by the test's properties, and the property that we shall focus on in this series is the likelihood ratio (LR).

The second conclusion we can draw from our two contrasting patients is that the pretest probability exerts a major influence on the diagnostic process. Each item of the history and physical examination is a diagnostic test that either increases or decreases the probability of a target disorder. Consider the young man who presented to your colleague. The fact that he presents with shortness of breath raises the possibility of PE. The fact that he has been immobile for 10 hours increases this probability, but his age, lack of antecedent disease, and normal physical examination, chest radiograph, and arterial blood gas measurements all decrease this probability. If we knew the properties of each of these pieces of information (and for some of them, we do[3,4]), we could move sequentially through them, incorporating each piece of information as we go and continuously recalculating the probability of the target disorder. Clinicians do proceed in this fashion, but because the properties of the individual items of history and physical examination usually are not available, they often must rely on clinical experience and intuition to arrive at the pretest probability that precedes ordering a diagnostic test. For some clinical problems, including the diagnosis of PE, their intuition has proved surprisingly accurate.[2]

Nevertheless, the limited information about the properties of items of history and physical examination often results in clinicians' varying widely in their estimates of pretest probabilities. There are a number of solutions to this problem. First, clinical investigators should study the history and physical exami-

From the Departments of Medicine and Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario.

A complete list of the members (with affiliations) of the Evidence-Based Medicine Working Group appears in the first article of this series (*JAMA*. 1993;270:2093-2095). The following members contributed to this article: Gordon Guyatt (chair), MD, MSc; Eric Bass, MD, MPH; Patrick Brill-Edwards, MD; George Browman, MD, MSc; Deborah Cook, MD, MSc; Michael Farkouh, MD; Hertzel Gerstein, MD, MSc; Brian Haynes, MD, MSc, PhD; Robert Hayward, MD, MPH; Anne Holbrook, MD, PharmD, MSc; Roman Jaeschke, MD, MSc; Elizabeth Juniper, MCSP, MSc; Hui Lee, MD, MSc; Mitchell Levine, MD, MSc; Virginia Moyer, MD, MPH; Jim Nishikawa, MD; Andrew Oxman, MD, MSc, FACPM; Ameen Patel, MD; John Philbrick, MD; W. Scott Richardson, MD; Stephane Sauve, MD, MSc; David Sackett, MD, MSc; Jack Sinclair, MD; K.S. Trout, FRCE; Peter Tugwell, MD, MSc; Sean Tunis, MD, MSc; Stephen Walter, PhD; and Mark Wilson, MD, MPH.

Reprint requests to McMaster University Health Sciences Centre, 1200 Main St W, Room 2C12, Hamilton, Ontario, Canada L8N 3Z5 (Dr Guyatt).

Table 1.—Evaluating and Applying the Results of Studies of Diagnostic Tests

**Are the results of the study valid?**
Primary guides:
- Was there an independent, blind comparison with a reference standard?
- Did the patient sample include an appropriate spectrum of patients to whom the diagnostic test will be applied in clinical practice?

Secondary guides:
- Did the results of the test being evaluated influence the decision to perform the reference standard?
- Were the methods for performing the test described in sufficient detail to permit replication?

**What are the results?**
- Are likelihood ratios for the test results presented or data necessary for their calculation provided?

**Will the results help me in caring for my patients?**
- Will the reproducibility of the test result and its interpretation be satisfactory in my setting?
- Are the results applicable to my patient?
- Will the results change my management?
- Will patients be better off as a result of the test?

Table 2.—Test Properties of Ventilation-Perfusion (V/Q) Scanning

| | Pulmonary Embolism | | | | |
|---|---|---|---|---|---|
| | Present | | Absent | | |
| V/Q Scan Result | No. | Proportion | No. | Proportion | Likelihood Ratio |
| High probability | 102 | 102/251 = 0.406 | 14 | 14/630 = 0.022 | 18.3 |
| Intermediate probability | 105 | 105/251 = 0.418 | 217 | 217/630 = 0.344 | 1.2 |
| Low probability | 39 | 39/251 = 0.155 | 273 | 273/630 = 0.433 | 0.36 |
| Normal/near normal | 5 | 5/251 = 0.020 | 126 | 126/630 = 0.200 | 0.10 |
| **Total** | **251** | ... | **630** | ... | ... |

nation to learn more about the properties of these diagnostic tests. Fortunately, such investigations are becoming common. Panzer and colleagues[5] have summarized much of the available information in the form of a medical text, and overviews on the accuracy and precision of the history and physical examination are being published concurrently with the Users' Guides in the *JAMA* series on The Rational Clinical Examination.[6] In addition, for some target disorders such as myocardial ischemia, multivariable analyses can provide physicians with ways of combining information to generate very precise pretest probabilities.[7] Second, when we don't know the properties of history and physical examination we can consult colleagues about their probability estimates; the consensus view is likely to be more accurate than our individual intuition. Finally, when we remain uncertain about the pretest probability, we can assume the highest plausible pretest probability, and the lowest possible pretest probability, and see if this changes our clinical course of action. We will illustrate how one might do this later in this discussion.

**Likelihood Ratios.**—The clinical usefulness of a diagnostic test is largely determined by the accuracy with which it identifies its target disorder, and the accuracy measure we shall focus on is the LR. Let's now look at Table 2, constructed from the results of the PIOPED study. There were 251 people with angiographically proven PE and 630 people whose angiograms or follow-up excluded PE. For all patients, V/Q scans were classified into four levels, from high probability to normal or near normal. How likely is a high-probability scan among people who do have PE? Table 2 shows that 102 of 251 people (or 0.406) with PE had high-probability scans. How often is the same test result, a high-probability scan, found among people who, although suspected of it, do not have PE? The answer is 14 of 630 or 0.022. The ratio of these two likelihoods is called the LR and for a high-probability scan equals 0.406 divided by 0.022 or 18.3. In other words, a high-probability lung scan is 18.3 times as likely to occur in a patient with, as opposed to a patient without, a PE. In a similar fashion, the LR can be calculated for each level of the diagnostic test result. Each calculation involves answering two questions: first, how likely it is to get a given test result (eg, a low-probability V/Q scan) among people with the target disorder (PE), and second, how likely it is to get the same test result (again, a low-probability scan) among people without the target disorder (no PE). For a low-probability V/Q scan these likelihoods are 39/251 (0.155) and 273/630 (0.433), and their ratio (the LR for a low-probability scan) is 0.36. As shown in Table 2, we can repeat these calculations for the other scan results.

What do all these numbers mean? The LRs indicate by how much a given diagnostic test result will raise or lower the pretest probability of the target disorder. An LR of 1 means that the posttest probability is exactly the same as the pretest probability. Likelihood ratios greater than 1 increase the probability that the target disorder is present, and the higher the LR the greater this increase. Conversely, LRs less than 1 decrease the probability of the target disorder, and the smaller the LR, the greater the decrease in probability and the smaller its final value.

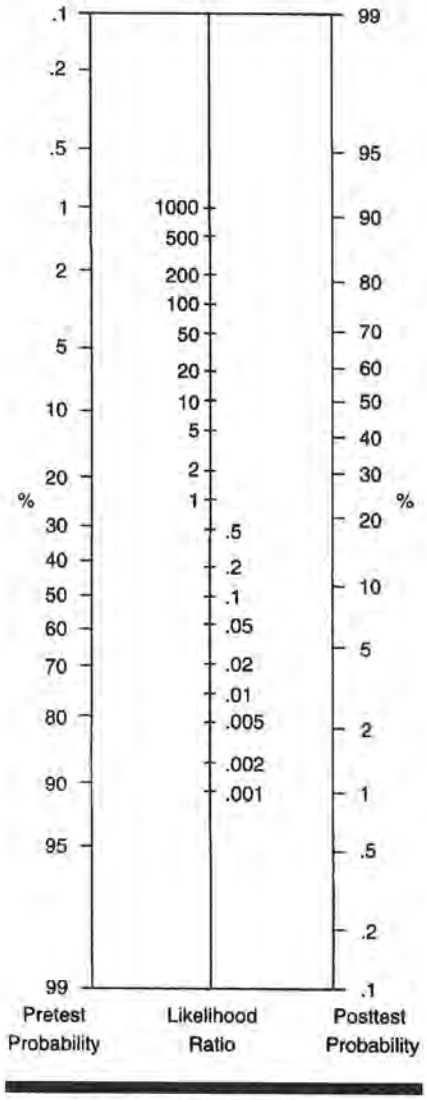
How big is a big LR, and how small is a small one? Using LRs in your day-to-day practice will lead to your own sense of their interpretation, but as a rough guide:

- Likelihood ratios greater than 10 or less than 0.1 generate large and often conclusive changes from pretest to posttest probability.
- Likelihood ratios of 5 to 10 and 0.1 to 0.2 generate moderate shifts in pretest to posttest probability.
- Likelihood ratios of 2 to 5 and 0.5 to 0.2 generate small (but sometimes important) changes in probability.
- Likelihood ratios of 1 to 2 and 0.5 to 1 alter probability to a small (and rarely important) degree.

Having determined the magnitude and significance of the LRs, how do we use them to go from pretest to posttest probability? We can't combine likelihoods directly, the way we can combine probabilities or percentages; their formal use requires converting pretest probability to odds, multiplying the result by the LR, and converting the consequent posttest odds into a posttest probability. While not too difficult,[2] this calculation can be tedious and involves the following: the equation to convert probabilities into odds is (probability/[1 − probability]), which is equivalent to probability of having the target disorder/probability of not having the target disorder. A probability of 0.5 represents odds of 0.50/0.50, or 1 to 1; a probability of 0.80 represents odds of 0.80/0.20, or 4 to 1; a probability of 0.25 represents odds of 0.25/0.75, or 1 to 3, or 0.33. Once you have the pretest odds, the posttest odds can be calculated by multiplying the pretest odds by the LR. The posttest odds can be converted back into probabilities using a formula of probability = odds/(odds + 1).

Fortunately, there is an easier way. A nomogram proposed by Fagan[8] (Figure) does all the conversions for us and allows us to go very simply from pretest to posttest probability. The first column of this nomogram represents the pretest probability, the second column represents the LR, and the third shows the posttest probability. You obtain the posttest probability by anchoring a ruler at the pretest probability and rotating it until it lines up with the LR for the observed test result.

Recall our elderly woman with suspected PE after abdominal surgery. Most clinicians would agree that the probability of this patient's having PE is quite high, at about 70%. This value then represents the pretest probability. Suppose that her V/Q scan was reported as high probability. Anchoring a ruler at her pre-




Nonogram for interpreting diagnostic test results. Adapted from Fagan.[8]

test probability of 70% and aligning it with the LR of 18.3 associated with a high-probability scan, her posttest probability is over 97%. What if her V/Q scan yielded a different result? If her V/Q scan result is reported as intermediate (LR, 1.2), the probability of PE hardly changes (to 74%), while a near-normal result yields a posttest probability of 19%.

We have pointed out that the pretest probability is an estimate, and that one way of dealing with the uncertainty is to examine the implications of a plausible range of pretest probabilities. Let us assume the pretest probability in this case is as low as 60%, or as high as 80%. The posttest probabilities that would follow from these different pretest probabilities appear in Table 3.

The same exercise may be repeated for our second patient, the young man with nonspecific chest pain and hyperventilation. If you consider that his presentation is compatible with a 20% probability of PE, using our nomogram the posttest probability with a high-probability scan result is 82%, an intermediate probability is 23%, and a near-normal probability is 2%. The pretest probability (with a range of possible pretest probabilites from 10% to 30%), LRs, and posttest probabilities associated with each of the four possible scan results also appear in Table 3.

Table 3.—Pretest Probabilities, Likelihood Ratios (LRs) of Ventilation-Perfusion Scan Results, and Posttest Probabilities in Two Patients With Pulmonary Embolus

| Pretest Probability, % (Range)* | Scan Result (LR) | Posttest Probability, % (Range)* |
|---|---|---|
| **78-Year-Old Woman With Sudden Onset of Dyspnea Following Abdominal Surgery** | | |
| 70 (60-80) | High probability (18.3) | 97 (96-99) |
| 70 (60-80) | Intermediate probability (1.2) | 74 (64-83) |
| 70 (60-80) | Low probability (0.36) | 46 (35-59) |
| 70 (60-80) | Normal/near normal (0.1) | 19 (13-29) |
| **28-Year-Old Man With Dyspnea and Atypical Chest Pain** | | |
| 20 (10-30) | High probability (18.3) | 82 (67-89) |
| 20 (10-30) | Intermediate probability (1.2) | 23 (12-34) |
| 20 (10-30) | Low probability (0.36) | 8 (4-6) |
| 20 (10-30) | Normal/near normal (0.1) | 2 (1-4) |

*The values in parentheses represent a plausible range of pretest probabilities. That is, while the best guess as to the pretest probability is 70%, values of 60% to 80% would also be reasonable estimates.

Readers who have followed the discussion to this point will understand the essentials of interpretation of diagnostic tests and can stop here. They should consider the next section, which deals with sensitivity and specificity, optional. We include it largely because clinicians will still encounter studies that present their results in terms of sensitivity and specificity and may wish to understand this alternative framework for summarizing the properties of diagnostic tests.

**Sensitivity and Specificity.**—You may have noted that our discussion of LRs ignored any talk of normal and abnormal tests. Instead, we presented four different V/Q scan interpretations, each with their own LR. This is not, however, the way the PIOPED investigators presented their results. They fell back on the older (but less useful) concepts of sensitivity and specificity.

Sensitivity is the proportion of people with the target disorder in whom the test result is positive, and specificity is the proportion of people without the target disorder in whom the test result is negative. To use these concepts we have to divide test results into normal and abnormal; in other words, create a 2×2 table. The general form of a 2×2 table, which we use to understand sensitivity and specificity, is presented in Table 4. Look again at Table 2 and observe that we could transform our 4×2 table into any of three such 2×2 tables, depending on what we call normal or abnormal (or what we call negative and positive test results). Let's assume that we call only high-probability scans abnormal (or positive). The resulting 2×2 table is presented in Table 5.

To calculate sensitivity from the data in Table 2 we look at the number of people with proven PE (251) who were diagnosed as having the target disorder on V/Q scan: 102 – sensitivity of 102/251, or approximately 41% ($a/[a+c]$). To calculate specificity we look at the number of people without the target disorder (630) who were was classified on V/Q scan as normals: 616 – specificity of 616/630, or 98% ($d/[b+d]$). We can also calculate LRs for the positive and negative test results using this cut point, 18.3 and 0.6, respectively.

Let's see how the test performs if we decide to put the threshold of positive vs negative in a different place in Table 2. For example, let's call only the normal/near-normal V/Q scan result negative. This 2×2 table (Table 6) shows the sensitivity is now 246/251, or 98% (among 251 people with PE, 246 are diagnosed on V/Q scan), but what has happened to specificity? Among 630 people without PE, only 126 have a negative test result (specificity of 20%). The corresponding LRs are 1.23 and 0.1. Note that with this cut we not only lose the diagnostic information associated with the high-probability scan result, but also interpret intermediate- and low-probability results as if they increase the likelihood of PE, when in fact they decrease the likelihood. You can generate the third 2×2 table by setting the cut point in the middle—if your sensitivity and specificity are 82% and 63%, respectively, and associated LRs of a positive and negative test 2.25 and 0.28, you have it right. (If you were to create a graph where the vertical axis will denote sensitivity [or true-positive rate] for different cutoffs and the horizontal axis will display 1–specificity [or false-posi-

tive rate] for the same cutoffs, and you connect the points generated by using cut points, you would have what is called a receiver operating characteristic [ROC curve]; an ROC curve can be used to formally compare the value of different tests by examining the area under each curve, but once one has abandoned the need for a single cut point, it has no other direct clinical application.)

You can see that in using sensitivity and specificity you have to either throw away important information or recalculate sensitivity and specificity for every cut point. We recommend the LR approach because it is simpler and more efficient.

Thus far, we have established that the results are likely true for the people who were included in the PIOPED study, and ascertained the LRs associated with different results of the test. How useful is the test likely to be in our clinical practice?

## WILL THE RESULTS HELP ME IN CARING FOR MY PATIENT?

### Will the Reproducibility of the Test Result and Its Interpretation Be Satisfactory in My Setting?

The value of any test depends on its ability to yield the same result when reapplied to stable patients. Poor reproducibility can result from problems with the test itself (eg, variations in reagents in radioimmunoassay kits for determining hormone levels). A second cause for different test results in stable patients arises whenever a test requires interpretation (eg, the extent of ST-segment elevation on an electrocardiogram). Ideally, an article about a diagnostic test will tell readers how reproducible the test results can be expected to be. This is especially important when expertise is required in performing or interpreting the test (and you can confirm this by recalling the clinical disagreements that arise when you and one or more colleagues examine the same electrocardiogram, ultrasound, or computed tomographic scan, even when all of you are experts).

If the reproducibility of a test in the study setting is mediocre and disagreement between observers is common, and yet the test still discriminates well between those with and without the target condition, it is very useful. Under these circumstances, it is likely that the test can be readily applied to your clinical setting. If reproducibility of a diagnostic test is very high and observer variation very low, either the test is simple and unambigious or those interpreting it are highly skilled. If the latter applies, less skilled interpreters in your own clinical setting may not do as well.

Table 4.—Comparison of the Results of Diagnostic Test With the Result of Reference Standard*

| | Reference Standard | |
|---|---|---|
| Test Result | Disease Present | Disease Absent |
| Disease present | True positive (*a*) | False positive (*b*) |
| Disease absent | False negative (*c*) | True negative (*d*) |

*Sensitivity = $a/(a+c)$.
Specificity = $d/(b+d)$.
Likelihood ratio for positive test result = $[a/(a+c)]/[b/(b+d)]$.
Likelihood ratio for negative test result = $[c/(a+c)]/[d/(b+d)]$.

In the PIOPED study, the authors not only provided a detailed description of their diagnostic criteria for V/Q scan interpretation, they also reported on the agreement between their two independent readers. Clinical disagreements over intermediate- and low-probability scans were common (25% and 30%, respectively), and they resorted to adjudication by a panel of experts.

### Are the Results Applicable to My Patient?

The issue here is whether the test will have the same accuracy among your patients as was reported in the article. Test properties may change with a different mix of disease severity or a different distribution of competing conditions. When patients with the target disorder all have severe disease, LRs will move away from a value of 1 (sensitivity increases). If patients are all mildly affected, LRs move toward a value of 1 (sensitivity decreases). If patients without the target disorder have competing conditions that mimic the test results seen in patients who do have the target disorder, the LRs will move closer to 1 and the test will appear less useful. In a different clinical setting in which fewer of the nondiseased have these competing conditions, the LRs will move away from 1 and the test will appear more useful.

The phenomenon of differing test properties in different subpopulations has been most strikingly demonstrated for exercise electrocardiography in the diagnosis of coronary artery disease. For instance, the more extensive the severity of coronary artery disease, the larger are the LRs of abnormal exercise electrocardiography for angiographic narrowing of the coronary arteries.[9] Another example comes from the diagnosis of venous thromboembolism, where compression ultrasound for proximal-vein thrombosis has proved more accurate in symptomatic outpatients than in asymptomatic postoperative patients.[10]

Sometimes, a test fails in just the patients one hopes it will best serve. The LR of a negative dipstick test for the rapid diagnosis of urinary tract infection is approximately 0.2 in patients with clear symptoms and thus a high probability of urinary tract infection, but is over 0.5 in those with low probability,[11] rendering it of little help in ruling out infection in the latter, low-probability patients.

Table 5.—Comparison of the Results of Diagnostic Test (Ventilation-Perfusion Scan) With the Result of Reference Standard (Pulmonary Angiogram) Assuming Only High-Probability Scans Are Positive (Truly Abnormal)*

| | Angiogram | |
|---|---|---|
| Scan Category | Pulmonary Embolus Present | Pulmonary Embolus Absent |
| High probability | 102 | 14 |
| Others | 149 | 616 |
| **Total** | **251** | **630** |

*Sensitivity, 41%; specificity, 98%; likelihood ratio of a high-probability test result, 18.3; likelihood ratio of other results, 0.61.

Table 6.—Comparison of the Results of Diagnostic Test (Ventilation-Perfusion Scan) With the Result of Reference Standard (Pulmonary Angiogram) Assuming Only Normal/Near-Normal Scans Are Negative (Truly Normal)*

| | Angiogram | |
|---|---|---|
| Scan Category | Pulmonary Embolus Present | Pulmonary Embolus Absent |
| High, intermediate, and low probability | 246 | 504 |
| Near normal/normal | 5 | 126 |
| **Total** | **251** | **630** |

*Sensitivity, 98%; specificity, 20%; likelihood ratio of high, intermediate, and low probability, 1.23; likelihood ratio of near normal/normal, 0.1.

If you practice in a setting similar to that of the investigation and your patient meets all the study inclusion criteria and does not violate any of the exclusion criteria, you can be confident that the results are applicable. If not, a judgment is required. As with therapeutic interventions, you should ask whether there are compelling reasons why the results should not be applied to your patients, either because the severity of disease in your patients, or the mix of competing conditions, is so different that generalization is unwarranted. The issue of generalizability may be resolved if you can find an overview that pools the results of a number of studies.

The patients in the PIOPED study were a representative sample of patients with suspected PE from a number of large general hospitals. The results are therefore readily applicable to most clinical practices in North America. There are groups to whom we might be reluctant to generalize the results, such as critically ill patients (who were excluded from the study, and who are likely to have a different spectrum of competing conditions than other patients).

### Will the Results Change My Management?

It is useful in making, learning, teaching, and communicating management decisions to link them explicitly to the probability of the target disorder. Thus, for any target disorder there are probabilities below which a clinician would dismiss a diagnosis and order no further tests (a "test" threshold). Similarly, there are probabilities above which a clinician would consider the diagnosis confirmed, and would stop testing and initiate treatment (a "treatment" threshold). When the probability of the target disorder lies between the test and treatment thresholds, further testing is mandated.[12]

Once we decide what our test and treatment thresholds are, posttest probabilities have direct treatment implications. Let us suppose that we are willing to treat those with a probability of PE of 80% or more (knowing that we will be treating 20% of our patients unnecessarily). Furthermore, let's suppose we are willing to dismiss the diagnosis of PE in those with a posttest probability of 10% or less. You may wish to apply different numbers here; the treatment and test thresholds are a matter of judgment, and differ for different conditions depending on the risks of therapy (if risky, you want to be more certain of your diagnosis) and the danger of the disease if left untreated (if the danger of missing the disease is high—such as in PE—you want your posttest probability very low before abandoning the diagnostic search). In our young man, a high-probability scan results in a posttest probability of 82% and may dictate treatment (or, at least, further investigation), an intermediate-probability scan (23% posttest probability) will dictate further testing (perhaps bilateral leg venography, serial impedance plethysmography or ultrasound, or pulmonary angiography), while a low-probability or normal scan (probabilities of <10%) will allow exclusion of the diagnosis of PE. In the elderly woman, a high-probability scan dictates treatment (97% posttest probability of PE), an intermediate result (74% posttest probability) may be compatible with either treatment or further testing (likely a pulmonary angiogram), while any other result mandates further testing.

If most patients have test results with LRs near 1, the test will not be very useful. Thus, the usefulness of a diagnostic test is strongly influenced by the proportion of patients suspected of having the target disorder whose test results have very high or very low LRs so that the test result will move their probability of disease across a test or treatment threshold. In the patients suspected of having PE in our V/Q scan example, review of Table 2 allows us to determine the proportion of patients with extreme results (either high probability with an LR of over 10, or near-normal/normal scans with an LR of 0.1). The proportion can be calculated as (102+14+5+126)/881 or 247/881=28%. Clinicians who have repeatedly been frustrated by frequent intermediate- or low-probability results in their patients with suspected PE will already know that this proportion (28%) is far from optimal. Thus, despite the high LR associated with a high-probability scan, and the low LR associated with a normal/near-normal result, V/Q scanning is of limited usefulness in patients with suspected PE.

A final comment has to do with the use of sequential tests. We have demonstrated how each item of history, or each finding on physical examination, represents a diagnostic test. We generate pretest probabilities that we modify with each new finding. In general, we can also use laboratory tests or imaging procedures in the same way. However, if two tests are very closely related, application of the second test may provide little or no information, and the sequential application of LRs will yield misleading results. For instance, once one has the results of the most powerful laboratory test for iron deficiency, serum ferritin, additional tests such as serum iron or transferrin saturation add no further information.[13]

### Will Patients Be Better Off as a Result of the Test?

The ultimate criterion for the usefulness of a diagnostic test is whether it adds information beyond that otherwise available, and whether this information leads to a change in management that is ultimately beneficial to the patient.[14] The value of an accurate test will be undisputed when the target disorder, if left undiagnosed, is dangerous, the test has acceptable risks, and effective treatment exists. High probability or near-normal/normal results of a V/Q scan may well eliminate the need for further investigation and result in anticoagulants' being appropriately given or appropriately withheld (either course of action having a substantial influence on patient outcome).

In other clinical situations, tests may be accurate, and management may even change as a result of their application, but their impact on patient outcome may be far less certain. Examples include right heart catheterization for many critically ill patients, or the incremental value of magnetic resonance imaging scanning over computed tomography for a wide variety of problems.

## HOW YOU CAN USE THESE GUIDES FOR CLINICAL PRACTICE AND FOR READING

By applying the principles described in this and the preceding article you will be able to assess and use information from articles about diagnostic tests. You are now equipped to decide whether an article concerning a diagnostic test represents a believable estimate of the true value of a test, what the test properties are, and the circumstances under which the test should be applied to your patients. Because LRs are now being published for an increasing number of tests,[5] the approach we have outlined has become directly applicable to the day-to-day practice of medicine.

#### References


1. Jaeschke R, Guyatt G, Sackett DL, for the Evidence-Based Working Group. Users' guides to the medical literatuere, III: how to use an article about a diagnostic test, A: are the results of the study valid? *JAMA*. 1994;271:389-391.
2. The PIOPED Investigators. Value of ventilation/perfusion scan in acute pulmonary embolism: results of the Prospective Investigation of Pulmonary Embolism Diagnosis (PIOPED). *JAMA*. 1990; 263:2753-2759.
3. Mayeski RJ. Pulmonary embolism. In: Panzer RJ, Black ER, Griner PF, eds. *Diagnostic Strategies for Common Medical Problems*. Philadelphia, Pa: American College of Physicians; 1991.
4. Stein PD, Terrin NL, Hales CA, et al. Clinical, laboratory, roentgenographic, and electrocardiographic findings in patients with acute pulmonary embolism and no pre-existing cardiac or pulmonary disease. *Chest*. 1991;100:598-603.
5. Panzer RJ, Black ER, Griner PF. *Diagnostic Strategies for Common Medical Problems*. Philadelphia, Pa: American College of Physicians; 1991.
6. Sackett DL, Rennie D. The science and art of the clinical examination. *JAMA*. 1992;267:2650-2652.
7. Pozen MW, D'Agostino RB, Selker HP, et al. A predictive instrument to improve coronary care-unit admission practices in acute ischemic heart disease. *N Engl J Med*. 1984;310:1273-1278.
8. Fagan TJ. Nomogram for Bayes's theorem (C). *N Engl J Med*. 1975;293:257.
9. Hlatky MA, Pryor DB, Harrell FE. Factors affecting sensitivity and specificity of exercise electrocardiography. *Am J Med*. 1984;77:64-71.
10. Ginsberg JS, Caco CC, Brill-Edwards PA, et al. Venous thrombosis in patients who have undergone major hip or new surgery: detection with compress US and impedance plethysmography. *Radiology*. 1991;181;651-654.
11. Lachs MS, Nachamkin I, Edelstein PH, et al. Spectrum bias in the evaluation of diagnostic tests: lessons from the repid dipstick test for urinary tract infection. *Ann Intern Med*. 1992;117:135-140.
12. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical Epidemiology: A Basic Science for Clinical Medicine*. 2nd ed. Boston, Mass: Little Brown & Co Inc; 1991:145-148.
13. Guyatt GH, Oxman A, Ali M. Diagnosis of iron deficiency. *J Gen Intern Med*. 1992;7:145-153.
14. Guyatt GH, Tugwell P, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. *Can Med Assoc J*. 1986; 134:587-594.

# Users' Guides to the Medical Literature

## IV. How to Use an Article About Harm

Mitchell Levine, MD, MSc; Stephen Walter, PhD; Hui Lee, MD, MSc; Ted Haines, MD, MSc;
Anne Holbrook, MD, PharmD, MSc; Virginia Moyer, MD, MPH; for the Evidence-Based Medicine Working Group

### CLINICAL SCENARIO

You are having lunch in the hospital cafeteria when one of your colleagues raises the issue of the safety of β-adrenergic agonists in the treatment of asthma. Your colleague feels uncertain about how to respond to patients asking him about media reports of an increased risk of death associated with these medications. Another colleague mentions a key article on this topic that generated much of the publicity, but she cannot recall the details. You all agree that this is an issue that arises frequently enough in your practices that you should become familiar with the evidence contained in the article that your patients have heard about. You volunteer to search the literature for the key article and report back to your colleagues in the next few days.

From the Departments of Clinical Epidemiology and Biostatistics (Drs Levine, Walter, and Haines), Medicine (Drs Lee and Holbrook), the Occupational Health Program (Dr Haines), McMaster University, Hamilton, Ontario; and the Department of Pediatrics, University of Texas, Houston (Dr Moyer).

A complete list of members (with affiliations) of the Evidence-Based Medicine Working Group appears in the first article of this series (*JAMA*. 1993;270:2093-2095). The following members contributed to this article: Gordon Guyatt (chair), MD, MSc; Eric Bass, MD, MPH; George Browman, MD, MSc; Deborah Cook, MD, MSc; Michael Farkouh, MD; Hertzel Gerstein, MD, MSc; Ted Haines, MD, MSc; Brian Haynes, MD, MSc, PhD, Robert Hayward, MD, MPH; Anne Holbrook, MD, PharmD, MSc; Roman Jaeschke, MD, MSc; Elizabeth Juniper, MCSP, MSc; Andreas Laupacis, MD, MSc; Hui Lee, MD, MSc; Mitchell Levine, MD, MSc; Virginia Moyer, MD, MPH; David Naylor, MD, DPhil; Jim Nishikawa, MD; Andrew Oxman, MD, MSc, FACPM; Ameen Patel, MD; John Philbrick, MD; Scott Richardson, MD; Stephane Sauve, MD, MSc; David Sackett, MD, MSc; Jack Sinclair, MD; Brian Strom, MD, MPH; K.S. Trout, FRCE; Sean Tunis, MD, MSc; Stephen Walter, PhD; John Williams Jr, MD, MHS; and Mark Wilson, MD, MPH.

Reprint requests to Room 2C12, McMaster University Health Sciences Centre, 1200 Main St W, Hamilton, Ontario, Canada, L8N 3Z5 (Dr Guyatt).

### THE SEARCH

The next day you do a MEDLINE search using the following terms: asthma (MH) (MH stands for MeSH heading, indexing terms used by National Library of Medicine personnel); adrenergic beta receptor agonists (MH); adverse effects (SH) (SH stands for Subheading). You limit the search to *Abridged Index Medicus* journals knowing that you will likely find the article your colleague recalled seeing within this list of major medical journals. Your MEDLINE search (1990 through 1993) identifies 38 citations. There were nine original studies, seven review articles, and 22 letters, editorials, and commentaries. Of the nine original articles, only one is an epidemiologic study assessing the risk of death associated with inhaled β-adrenergic agonists, and you think this is the article to which your colleague referred. The study describes a 2.6-fold increased risk of death from asthma associated with the use of β-adrenergic agonist metered-dose inhalers.[1]

### INTRODUCTION

Clinicians often encounter patients who may be facing harmful exposures, either to medical interventions or environmental agents. Are pregnant women at increased risk of miscarriage if they work in front of video display terminals? Do vasectomies increase the risk of prostate cancer? Do hypertension management programs at work lead to increased absenteeism? When examining these questions, physicians must evaluate the validity of the data, the strength of the association between the putative cause and the adverse outcome, and the relevance to patients in their practice (Table 1).

This article in our series of "Users' Guides to the Medical Literature" will help you evaluate an individual article assessing an issue of harm. To fully assess the cause-and-effect relationship implied in any question of harm requires consideration of all the information available. Systematic overviews (eg, meta-analyses) can provide an objective summary of all the available evidence, and we will deal with how to use an overview in a subsequent article in this series. Using such an overview requires a prior understanding of the rules of evidence for individual studies, and this article covers the basic rules for observational (nonrandomized) studies.

### ARE THE RESULTS OF THE STUDY VALID?

#### Primary Guides

**Were There Clearly Identified Comparison Groups That Were Similar With Respect to Important Determinants of Outcome Other Than the One of Interest?**—In a study that identifies a harmful exposure, the choice of comparison groups has an enormous influence on the credibility of the results. Because the design of the study determines the comparison groups, we will review the basic study designs that clinicians encounter when assessing whether their patients have been or might be exposed to a potentially harmful factor (Table 2).

*Randomized Controlled Trials.*—A randomized controlled trial (RCT) is a true experiment in which patients are assigned, by a mechanism analogous to

Table 1.—User's Guides for an Article About Harm

**Are the results of the study valid?**
- Primary guides:
  - Were there clearly identified comparison groups that were similar with respect to important determinants of outcome, other than the one of interest?
  - Were the outcomes and exposures measured in the same way in the groups being compared?
  - Was follow-up sufficiently long and complete?
- Secondary guides:
  - Is the temporal relationship correct?
  - Is there a dose-response gradient?

**What are the results?**
- How strong is the association between exposure and outcome?
- How precise is the estimate of the risk?

**Will the results help me in caring for my patients?**
- Are the results applicable to my practice?
- What is the magnitude of the risk?
- Should I attempt to stop the exposure?

Table 2.—Directions of Inquiry and Key Methodologic Strengths and Weaknesses for Different Study Designs

| Design | Starting Point | Assessment | Strengths | Weaknesses |
|---|---|---|---|---|
| Randomized controlled trial | Exposure status | Adverse event status | Internal validity | Feasibility, generalizability |
| Cohort | Exposure status | Adverse event status | Feasible when randomization of exposure not possible | Susceptible to threats to internal validity |
| Case control | Adverse event status | Exposure status | Overcomes temporal delays, may only require small sample size | Susceptible to threats to internal validity |

a coin flip, to either the putative causal agent or some alternative experience (either another agent or no exposure at all). Investigators then follow the patients forward in time and assess whether they have experienced the outcome of interest. The great strength of the RCT is that we can be confident that the study groups were similar not only with respect to determinants of outcome that we know about, but also those we do not know about.

In prior articles in this series, we have shown how readers of articles about therapy can use the results of RCTs.[2,3] Randomized controlled trials are rarely done to study possible harmful exposures, but if a well-designed RCT demonstrates an important relationship between an agent and an adverse event, clinicians can be confident of the results. For instance, the Cardiac Arrhythmia Suppression Trial is an RCT that demonstrated an association between the antiarrhythmic agents encainide, flecainide, and moricizine, and excessive mortality.[4,5] As a result, clinicians have curtailed their use of these drugs and have become much more cautious in using other antiarrhythmic agents in the treatment of nonsustained ventricular arrhythmias.

*Cohort Studies.*—When it is either not feasible or not ethical to randomly assign patients to be exposed or not exposed to a putative causal agent, investigators must find an alternative to an RCT. In a cohort study, the investigator identifies exposed and nonexposed groups of patients and then follows them forward in time, monitoring the occurrence of the outcome. You can appreciate the practical need for cohort studies when subjects cannot be "assigned" to an exposure group, as occurs when one wants to evaluate the effects of an occupational exposure. For example, investigators assessed perinatal outcomes among children of men exposed to lead and organic solvents in the printing industry using a cohort of all males who had been members of printers' unions in Oslo, Norway, and on the basis of job classification, they categorized fathers as to their exposure to lead and solvents. In this study, exposure was associated with an eightfold increase in preterm births, but no significant impact on birth defects.[6]

Cohort studies may also be performed when harmful outcomes are infrequent. For example, clinically apparent upper gastrointestinal hemorrhage in nonsteroidal anti-inflammatory drug (NSAID) users occurs approximately 1.5 times per 1000 person years of exposure, in comparison with 1.0 per 1000 person years in those not taking NSAIDs (assuming a stable risk over time).[7] An RCT to study this effect would require approximately 6000 patient-years of exposure to achieve a 95% probability of observing at least one additional serious gastrointestinal hemorrhage among treated patients, and a substantially larger sample size (approximately 75 000 patient-years per group) for adequate power to test the hypothesis that NSAIDs cause the additional hemorrhages.[8] Such an RCT would not be feasible, but a cohort study, particularly one in which the information comes from a large administrative database, would be.

Because subjects in a cohort study select themselves (or are selected by a physician) for exposure to the putative harmful agent, there is no particular reason they should be similar to nonexposed persons with respect to other important determinants of outcome. It therefore becomes crucial for investigators to document the characteristics of the exposed and nonexposed subjects and either demonstrate their comparability or use statistical techniques to adjust for differences. In the association between NSAIDs and the increased risk of upper gastrointestinal bleeding, age is associated both with exposure to NSAIDs and with gastrointestinal bleeding, and is therefore called a possible "confounding variable." In other words, since patients taking NSAIDs will be older, it may be difficult to tell if their increased risk of bleeding is because of their age or because of their NSAID exposure. When such a confounding variable is unequally distributed in the exposed and nonexposed populations, investigators use statistical techniques that correct or adjust for the imbalances.

Even if investigators document the comparability of potentially confounding variables in exposed and nonexposed cohorts or use statistical techniques to adjust for differences, there may be an important imbalance in prognostic factors that the investigators don't know about or have not measured that may be responsible for differences in outcome. It may be, for instance, that illnesses that require NSAIDs, rather than the NSAIDs themselves, are responsible for the increased risk of bleeding. Thus, the strength of inference from a cohort study will always be less than that of a rigorously conducted RCT.

*Case-Control Studies.*—When the outcome of interest either is very rare or takes a long time to develop, cohort studies also may not be feasible. Investigators may use an alternative design in which they identify cases, patients who have already developed the outcome of interest (eg, a disease, hospitalization, death). The investigators then choose controls, persons who do not have the outcome of interest, but who are otherwise similar to the cases with respect to important determinants of outcome such as age, sex, and concurrent medical conditions. Investigators can then assess retrospectively the relative frequency of exposure to the putative harmful agent among the cases and controls. This observational design is called a case-control study.

Using a case-control design, investigators demonstrated the association between diethylstilbestrol ingestion by pregnant women and the development of vaginal adenocarcinoma in their daughters many years later.[9] A prospective cohort study designed to test this cause-and-effect relationship would have required at least 20 years from the time when the association was first suspected until the completion of the study. Further, given the infrequency of the disease, a cohort study would have required hundreds of thousands of subjects. Using the case-control strategy, the investigators defined two groups of young

women—those who had suffered the outcome of interest (vaginal adenocarcinoma) were designated as the cases (n=8), and those who did not have the outcome, as the controls (n=32). Then, working backward in time, the exposure rates to diethylstilbestrol were determined for the two groups. Analogous to the situation with a cohort study, investigators had to ensure balance, or adjust for imbalances, in important risk factors in cases and controls (eg, intrauterine x-ray exposure). The investigators found a strong association between in utero diethylstilbestrol exposure and vaginal adenocarcinoma that was extremely unlikely to be attributable to the play of chance ($P<.00001$), without a delay of 20 years, and requiring only 40 women.

As with cohort studies, case-control studies are susceptible to unmeasured confounders. Therefore, the strength of inference that can be drawn from the results may be limited.

*Case Series and Case Reports.*—Case series and case reports do not provide any comparison group and are therefore unable to satisfy the requirements of the first primary guide. Although descriptive studies occasionally demonstrate dramatic findings mandating an immediate change in physician behavior (eg, thalidomide and birth defects), there are potentially undesirable consequences when actions are taken in response to weak evidence. Bendectin (a combination of doxylamine, pyridoxine, and dicyclomine used as an antiemetic in pregnancy) was withdrawn as a result of case reports suggesting it was teratogenic.[10] Later, a number of comparative studies demonstrated the relative safety of the drug,[11] but they could not eradicate a litigious atmosphere that prompted the manufacturer to withdraw the drug from the market. Thus, many pregnant women who could have benefited were denied the symptom relief the drug could have offered. In general, clinicians should not draw conclusions about cause-and-effect relationships from case series, but recognize that the results may generate questions for regulatory agencies and clinical investigators to address.

*Design Issues—Summary.*— It is apparent that, just as for questions of therapeutic effectiveness, clinicians should look for RCTs to resolve issues of harm. It is also apparent that they will often be disappointed in this search, and must be satisfied with studies of weaker design. Whatever the design, however, they should look for an appropriate control population before making a strong inference about a putative harmful agent.

**Were the Exposures and Outcomes Measured in the Same Way in the Groups Being Compared?**—In case-control studies, ascertainment of the exposure is a key issue. Patients with leukemia, when asked about prior exposure to solvents, may be more likely to recall exposure than would a control group, either because of increased patient motivation (recall bias) or greater probing by an interviewer (interviewer bias). Clinicians should attend to whether investigators used strategies, such as blinding subjects and interviewers to the hypothesis of the study, to minimize bias. For example, in a case-control study describing the association between psychotropic drug use and hip fracture, investigators established drug exposure by examining computerized claims files of the Michigan Medicaid program, a strategy that avoided both recall and interviewer bias.[12] As a result, the clinician has more confidence in the study's findings of a twofold increase in the risk of hip fracture.

Exposure opportunity should also be similar among cases and controls. There is evidence suggesting a 2.7-fold increased risk of homicide among individuals keeping a gun in their home. It would be important to know that the control group had a similar opportunity for gun possession, otherwise the true risk could be different from the study results—increased if the controls had a greater opportunity, decreased if the controls had a lesser opportunity for gun possession.[13]

In RCTs and cohort studies, ascertainment of outcome is the key issue. Investigators have reported a threefold increase in risk of malignant melanoma in individuals working with radioactive materials. One possible explanation for some of the increased risk might be that physicians, aware of a possible risk, search more diligently and therefore detect disease that might otherwise go unnoticed (or detect disease at an earlier point in time). This could result in the exposed cohort having an apparent, but spurious, increase in risk—a situation we refer to as surveillance bias.[14]

**Was Follow-up Sufficiently Long and Complete?**—An additional point relating to the measurement of outcomes is the need for adequate follow-up in RCTs and cohort studies. As discussed in a previous article in this series,[2] patients unavailable for follow-up threaten the validity of the results because the patients who are unavailable may have very different outcomes from those available for assessment. The longer the follow-up period required, the greater the possibility that the follow-up will be incomplete.

In a well-executed study, investigators determined the vital status of 1235 of 1261 white males (98%) employed in chrysotile asbestos textile operation between 1940 and 1975. The relative risk (RR) for lung cancer death increased monotonically from 1.4 to 18.2 with cumulative exposure among asbestos workers with at least 15 years since first exposure.[15] Because the 2% missing data were unlikely to affect the results and the follow-up was sufficiently long, the study allows relatively strong inference about the increase in cancer risk with asbestos exposure.

### Secondary Guides

**Is the Temporal Relationship Correct?**—Does exposure to the harmful agent precede the adverse outcome? The reports of increased suicidal ideation associated with the use of the antidepressant fluoxetine illustrate the importance of this question.[16] Did the thoughts of suicide occur after the fluoxetine was administered, or were the patients given this drug because they were already showing signs of clinical deterioration? A meta-analysis of controlled trials of treatment for depression did not confirm the apparent association.[17]

**Is There a Dose-Response Gradient?**—We are more confident attributing an adverse outcome to a particular exposure if, as the quantity or the duration of exposure to the putative harmful agent increases, risk of the adverse outcome also increases. The risk of dying from lung cancer in male physician smokers is dose-dependent; the risk increases by 50%, 132%, and 220% for one to 14, 15 to 24, and 25 or more cigarettes smoked per day, respectively.[18]

## WHAT ARE THE RESULTS?

**How Strong Is the Association Between Exposure and Outcome?**—We have described the most common way of expressing an association between exposure and outcome, the RR, in detail in an earlier article in this series.[3] In brief, the RR is the risk (or incidence) of the adverse effect in the exposed group divided by the risk of the adverse effect in the unexposed group. Values greater than 1 represent an increase in risk associated with the exposure, while values less than 1 represent a reduction in risk. To illustrate, in a cohort study assessing inhospital mortality following noncardiac surgery in male veterans, 23 of 289 patients with a history of hypertension died, compared with three of 185 patients without hypertension. The RR of death for hypertensive men was 4.9.[19] The RR tells us that death occurs almost five times more often in the hypertensive patients than in normotensive patients.

Table 3.—Estimate of Relative Risks and Odds Ratios for Exposed and Unexposed Patients

| Patient | Adverse Event (Case) | No Adverse Event (Control) |
|---|---|---|
| Exposed | a | b |
| Not exposed | c | d |

*Relative risk = [a/(a+b)]/[c/(c+d)].
Odds ratio = (a/c)/(b/d).

The estimate of RR depends on having samples of exposed and unexposed patients, where the proportion of the patients with the outcome of interest can be determined. The RR is therefore not applicable to case-control studies in which the number of cases and controls, and therefore the proportion of individuals with the outcome, is chosen by the investigator. For case-control studies, instead of using a ratio of risks, we use a ratio of odds: the odds of a case patient being exposed divided by the odds of a control patient being exposed. Using a simple 2×2 table, RRs and odds ratios (ORs) can be represented as depicted in Table 3.

When the outcome of interest is rare in the population from which the sample of cases was drawn, which is often the reason for using a case-control design in the first place, the OR closely approximates the RR.

When considering both study design and strength of association, we may be ready to interpret a small increase in risk as representing a true harmful effect when the study design is strong (such as an RCT). A much higher increase in risk might be required of weaker designs (such as cohort or case-control studies) as subtle findings are more likely to be because of subtle flaws in design. Very large values for RRs or ORs represent strong associations that are less likely to be caused by confounding or bias.

**How Precise Is the Estimate of the Risk?**—In a previous article in this series we have shown how the clinician can evaluate the precision of the estimate of treatment effect by examining the confidence interval (CI) around that estimate.[3] The clinician can take the same approach with articles assessing risk. In a study in which the investigators have shown an association between an exposure and an adverse outcome, the lower limit of the estimate of RR associated with the adverse exposure provides a minimal estimate of the strength of the association. In a study where the investigators fail to demonstrate an association (a "negative" study), the upper boundary of the CI around the RR tells the clinician just how big an adverse effect may still be present, despite the failure to show a statistically significant association.

## WHAT ARE THE IMPLICATIONS FOR MY PRACTICE?

**Are the Results Applicable to My Practice?**—If you are convinced that the results of the study are valid for the population that was studied, you then have to decide whether you can extrapolate the results to patients in your own practice. Are your patients similar to those described in the study with respect to morbidity, age, race, or other potentially important factors? Are there clinically important differences in the treatments or exposures between your patients and the patients studied? For example, the risk of thrombophlebitis associated with oral contraceptives described in the 1970s may not be applicable to the patient of the 1990s because of the lower estrogen doses currently in use. Similarly, increases in uterine cancer secondary to postmenopausal estrogen probably don't apply to women who are also taking concomitant progestins tailored to produce monthly withdrawal bleeding.

**What Is the Magnitude of the Risk?**—The RR and the OR do not tell us how frequently the problem occurs, only that the observed effect occurs more or less often in the exposed group compared with the unexposed group. Thus, the reader needs a method for assessing clinical importance. In our discussion of therapy we described how the clinician can calculate the number of patients she must treat to prevent an adverse event.[3] When the issue is harm, the clinician can use data from an RCT or cohort study to make an analogous calculation to determine how many people must be exposed to the harmful agent to cause an adverse outcome. From the Cardiac Arrhythmia Suppression Trial, over an average of 10 months of follow-up, mortality was 3.0% and 7.7% for placebo and encainide/flecainide patients, respectively. The absolute risk increase was 4.7%, the reciprocal of which tells us that, on average, for every 21 patients we treat with encainide or flecainide for about a year, we will cause one excess death.[4] This contrasts with NSAIDs and upper gastrointestinal bleeding. Of 2000 unexposed patients, two will suffer a hemorrhage each year. Of 2000 patients taking NSAIDs, three will suffer a hemorrhage each year. Thus, if we treat 2000 patients with NSAIDs, we can expect a single additional bleeding event.[6]

**Should I Attempt to Stop the Exposure?**—After evaluating the evidence that an exposure is harmful, determining subsequent actions may not be simple. There are at least three aspects the physician must consider in making a clinical decision.[20]

One is the strength of inference; how strong was the study or studies that demonstrated harm in the first place? Second, what is the magnitude of the risk to patients if exposure to the harmful agent continues? Third, what are the adverse consequences of reducing or eliminating exposure to the harmful agent?

Clinical decision making is simple when both the likelihood of harm and its magnitude are great. Because the evidence of increased mortality from encainide and flecainide came from an RCT, we can be confident of the causal connection. Since treating only 21 people will result in an excess death, it is no wonder that clinicians quickly curtailed their use of these antiarrhythmic agents when the study results became available.

The clinical decision is also made easier when an acceptable alternative for avoiding the risk is available. For example, β-blockers prescribed for the treatment of hypertension can result in a symptomatic increase in airways resistance in patients with asthma or chronic airflow limitation, mandating the use of an alternative drug, such as a thiazide diuretic, in susceptible patients.[21] Even if the evidence is relatively weak, the availability of an alternative can result in a clear decision. The early case-control studies demonstrating the association between aspirin use and Reye's syndrome were relatively weak and left considerable doubt about the causal relationship. Although the strength of inference was not great, the availability of a safe, inexpensive, and well-tolerated alternative, acetaminophen, justified use of this alternative agent in children at risk of Reye's syndrome.[22]

In contrast to the early studies regarding aspirin and Reye's syndrome, multiple well-designed cohort and case-control studies have consistently demonstrated an association between NSAIDs and upper gastrointestinal bleeding, and our inference about harm has therefore been relatively strong. However, the risk of an upper gastrointestinal hemorrhage is quite low, and we don't have safer, equally efficacious anti-inflammatory alternatives available. We are therefore probably right in continuing to prescribe NSAIDs for the appropriate clinical conditions.

## RESOLUTION OF THE SCENARIO

The study you retrieved on the risks of inhaled β-adrenergic therapy used a case-control design relying on computer record linkages between health insurance data and a drug plan.[1] The database for the study included 95% of the population of the province of Saskatch-

ewan in western Canada. The investigators matched 129 cases of fatal or near-fatal asthma with 655 controls who were also asthmatics. The investigators attempted to control for potential confounders, such as disease severity. Their measures of disease severity included the number of hospitalizations in the previous 24 months and an index of the aggregate use of medications. They found an association between the routine use of large doses of β-adrenergic agonist metered-dose inhalers and death from asthma (OR, 2.6 per canister per month; 95% CI, 1.7 to 3.9).

The study satisfied the validity criteria in Table 1 quite well. The investigators chose an appropriate control population and corrected for measurable potential differences in important prognostic factors in the treatment and control groups; exposure and outcome were measured the same way in treatment and control groups; the temporal relationship is correct; and they found a dose-response gradient. However, the study used a case-control design rather than an RCT, and we remain uncertain whether differences in unmeasured prognostic variables between the treatment and control groups explain the results. In other words, it is still possible that the patients who used more β-agonists were sicker, and this (rather than their increased use of the drug) explains the increased risk of death.

The magnitude of the association is moderate, and although the baseline risk of death from asthma (44 deaths in 12 301 asthmatic patients receiving medication, 0.3%) is low enough that we would have to treat large numbers of patients before the drugs were responsible for a death, reducing preventable deaths is extremely important. The fact that the data came from a population-based study suggests the results are widely generalizable.

Thus, as an individual study on the subject, you find the results of an "association" between inhaled β-adrenergic agonist use and death both believable and relevant to your practice. Because it is not an RCT, you are less certain about a true causal relationship underlying the observed association. Full assessment of the likelihood of a causal relationship would require a systematic review of all the evidence in the literature. You tell your inquiring patients that there is an increased risk of death in heavy users of inhaled β-adrenergic agonists, but that you cannot be certain whether it is because of the drug or possibly the consequence of having severe disease. Intermittent use of inhaled β-agonist therapy in patients with reversible airflow obstruction provides an attractive alternative to more intensive administration, and many clinicians have responded to the results of this and other studies by choosing this alternative approach.

#### References


1. Spitzer WO, Suissa S, Ernst P, et al. The use of beta-agonists and the risk of death and near death from asthma. *N Engl J Med.* 1992;326:501-506.
2. Guyatt GH, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, II: how to use an article about therapy or prevention, A: are the results of the study valid? *JAMA.* 1993;270:2598-2601.
3. Guyatt GH, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, II: how to use an article about therapy or prevention, B: what were the results and will they help me in caring for my patients? *JAMA.* 1994;271:59-63.
4. Echt DS, Liebson PR, Mitchell LB, et al. Mortality and morbidity in patients receiving encainide, flecainide, or placebo: the Cardiac Arrhythmia Suppression Trial. *N Engl J Med.* 1991;324:781-788.
5. The Cardiac Arrhythmia Suppression Trial II Investigators. Effect of the antiarrhythmic agent moricizine on survival after myocardial infarction. *N Engl J Med.* 1992;327:227-233.
6. Kristenson P, Irgens LM, Daltveit AK, Andersen A. Perinatal outcomes among children of men exposed to lead and solvents in the printing industry. *Am J Epidemiol.* 1993;137:134-143.
7. Carson JL, Strom BL, Soper KA, West SL, Morse ML. The association of nonsteroidal anti-inflammatory drugs with upper gastrointestinal tract bleeding. *Arch Intern Med.* 1987;147:85-88.
8. Walter SD. Determination of significant relative risks and optimal sampling procedures in prospective and retrospective comparative studies of various sizes. *Am J Epidemiol.* 1977;105:387-397.
9. Herbst AL, Ulfelder H, Poskanzer DC. Adenocarcinoma of the vagina: association of maternal stilbestrol therapy with tumor appearance in young women. *N Engl J Med.* 1971;284:878-881.
10. Soverchia G, Perr PF. Two cases of malformation of a limb in infants of mothers treated with an antiemetic in a very early phase of pregnancy. *Pediatr Med Chir.* 1981;3:97-99.
11. Holmes LB. Teratogen update: bendectin. *Teratology.* 1983;27:277-281.
12. Ray WA, Griffin MR, Schaffner W, Baugh DK, Melton LJ III. Psychotropic drug use and the risk of hip fracture. *N Engl J Med.* 1987;316:363-369.
13. Kellermann AL, Rivara FP, Rushforth NB, et al. Gun ownership as a risk factor for homicide in the home. *N Engl J Med.* 1993;329:1084-1091.
14. Hiatt RA, Fireman B. The possible effect of increased surveillance on the incidence of malignant melanoma. *Prev Med.* 1986;15:652-660.
15. Dement JM, Harris RL, Symons MJ, Shy CM. Exposures and mortality among chrysotile asbestos workers: part II, mortality. *Am J Med.* 1983; 1:421-433.
16. Teicher MH, Glod C, Cole JO. Emergence of intense suicidal preoccupation during fluoxetine treatment. *Am J Psychiatry.* 1990;147:207-210.
17. Beasley CM, Dornseif BE, Bosomworth JC, et al. Fluoxetine and suicide: a meta-analysis of controlled trials of treatment for depression. *BMJ.* 1991;303:685-692.
18. Doll R, Hill AB. Mortality in relation to smoking: ten years' observation of British doctors. *BMJ.* 1964;II:1399-1410, 1460-1467.
19. Browner WS, Li J, Mangano DT, for the Study of Perioperative Ischemia Research Group. In-hospital and long-term mortality in male veterans following noncardiac surgery. *JAMA.* 1992;268:228-232.
20. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical Epidemiology: A Basic Science for Clinical Medicine.* 2nd ed. Boston, Mass: Little Brown & Co Inc; 1991.
21. Ogilvie RI, Burgess ED, Cusson JR, Feldman RD, Leiter LA, Myers MG. Report of the Canadian Hypertension Society Consensus Conference, 3: pharmacologic treatment of essential hypertension. *Can Med Assoc J.* 1993;149:575-584.
22. Soumerai SB, Ross-Degnan D, Kahn JS. Effects of professional and media warnings about the association between aspirin use in children and Reye's syndrome. *Milbank Q.* 1992;70:155-167.

# Users' Guides to the Medical Literature

## V. How to Use an Article About Prognosis

Andreas Laupacis, MD, MSc; George Wells, MSc, PhD; W. Scott Richardson, MD; Peter Tugwell, MD, MSc; for the Evidence-Based Medicine Working Group

### CLINICAL SCENARIO

You are about to see a 76-year-old retired schoolteacher for the second time. You first saw her in the clinic a month ago because of cognitive problems. Your evaluation at that time included a Standardized Mini-Mental State Examination,[1] on which she scored 18 out of a possible 30 points, and a physical examination that was normal including no focal neurological signs. You arranged investigations for the treatable causes of dementia that were negative, and you thus feel she has probable Alzheimer's disease.

The patient has lived with her son since her husband died 6 years ago. Her son thinks that she first developed significant problems with her memory about 3 years ago. However, she has become increasingly agitated and paranoid during the last year. She has refused to allow him to look after her financial affairs, despite the fact that she owns three pieces of property and isn't able to manage them herself. Her son asked you about her prognosis, and whether she is likely to die soon from the dementia. You indicated that you would discuss this with him at the second visit once the results of all the tests are available.

### SEARCH

Hoping to provide the son with the most specific information possible about his mother's prognosis, after the initial visit you searched the medical library's MEDLINE CD-ROM system via the hospital's network on the clinic computer. You entered "*Alzheimer's Disease," which yielded 3687 articles from 1990 onward. Next, you entered "prognosis," which yielded 23 004 articles; crossing the two sets yielded 27 articles. Scanning the abstracts on screen, you found several articles of potential interest, including one that seemed precisely on target: "Survival of Outpatients With Alzheimer-Type Dementia" by Walsh and colleagues.[2]

### INTRODUCTION

In this article we will suggest a framework that you can use to efficiently assess articles that deal with prognosis, using the article on patients with dementia as an example. We will follow the usual format of this series and discuss how to determine whether the results are valid, how to interpret the results, and whether the information will benefit your patients (Table).

"Prognosis" refers to the possible outcomes of a disease and the frequency with which they can be expected to occur (eg, death in a patient with dementia). Sometimes the characteristics of a particular patient can be used to more accurately predict that patient's eventual outcome (eg, a patient with dementia and behavioral problems may have a worse prognosis than someone without behavioral problems). These characteristics are called "prognostic factors." Prognostic factors can be any of several types, such as demographic (eg, age), disease-specific (eg, tumor stage), or comorbid (eg, other conditions accompanying the disease in question). They can predict any outcome, whether good (eg, cure or survival) or bad (eg, death or complication). Prognostic factors need not necessarily cause the outcomes, just be associated with them strongly enough to predict their development. In the literature, prognostic factors are usually distinguished from "risk factors," those patient characteristics associated with the development of the disease in the first place. For example, smoking is an important risk factor for the development of lung cancer, but is not as important a prognostic factor as tumor stage in someone who has lung cancer.

It is usually impossible or unethical to randomize patients to different prognostic factors. Therefore, the best study design to identify the presence of and determine the increased risk associated with a prognostic factor is a cohort study. As we described in a previous article in this series,[3] investigators conducting a cohort study follow one or more groups (cohorts) of individuals who have not yet suffered an adverse event and monitor the number of outcome events over time. An ideal cohort study consists of a well-defined sample of individuals representative of the population of interest and uses objective outcome criteria. One cohort study conducted in Framingham, Mass, in which investigators have followed a cohort of 5209 individuals since 1948, has provided clinicians with a great deal of useful information about the prognostic importance of many determinants of cardiovascular disease.[4] Since rigorous randomized trials include careful documentation of inclusion criteria and strict protocols for follow-up, patients in such trials form cohorts that can also generate information about the prognosis of a disease. However, the patients entered into the trial are often not representative of the population with the disorder.[5]

From the Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario (Dr Laupacis); the Departments of Medicine and Epidemiology and Community Medicine, University of Ottawa (Ontario) (Drs Laupacis, Wells, and Tugwell); and the Department of Medicine, University of Rochester (NY) School of Medicine and Dentistry (Dr Richardson).

A complete list of members (with affiliations) of the Evidence-Based Medicine Working Group appears in the first article of this series (*JAMA*. 1993;270:2093-2095). The following members contributed to this article: Gordon H. Guyatt (chair), MD, MSc; George Browman, MD, MSc; Deborah Cook, MD, MSc; Hertzel Gerstein, MD, MSc; Brian Haynes, MD, MSc, PhD; Robert Hayward, MD, MPH; Mitchell Levine, MD, MSc; Jim Nishikawa, MD; David L. Sackett, MD, MSc; Patrick Brill-Edwards, MD; Michael Farkouh, MD; Anne Holbrook, MD, PharmD, MSc; Roman Jaeschke, MD, MSc; Hui Lee, MD, MSc; Stephane Sauve, MD, MSc; Virginia Moyer, MD, MPH; David Naylor, MD, DPhil, Andrew Oxman, MD, MSc, FACPM; John Philbrick, MD; Jack Sinclair, MD; Brian L. Strom, MD, MPH; Sean Tunis, MD, MSc; John Williams, Jr, MD, MHS; and Mark Wilson, MD, MPH.

Reprint requests to Room 2C12, McMaster University Health Sciences Centre, 1200 Main St W, Hamilton, Ontario, Canada L8N 3Z5 (Dr Guyatt).

Users' Guides to the Medical Literature Section Editor: Drummond Rennie, MD, Deputy Editor (West), *JAMA*.

**Are the results of the study valid?**
- Primary guides:
  - Was there a representative and well-defined sample of patients at a similar point in the course of the disease?
  - Was follow-up sufficiently long and complete?
- Secondary guides:
  - Were objective and unbiased outcome criteria used?
  - Was there adjustment for important prognostic factors?

**What are the results?**
- How large is the likelihood of the outcome event(s) in a specified period of time?
- How precise are the estimates of likelihood?

**Will the results help me in caring for my patients?**
- Were the study patients similar to my own?
- Will the results lead directly to selecting or avoiding therapy?
- Are the results useful for reassuring or counseling patients?

To study prognostic factors, investigators can also collect "cases" of individuals who have already suffered the outcome event and compare them with "controls" who have not. In these "case-control" studies, the investigators count the number of individuals in each group with a particular prognostic factor (eg, were the patients with dementia who died more likely to have had behavioral problems than those who did not die?). The potential for bias when selecting cases and controls, as well as the retrospective nature of data collection about prognostic factors (which often depends on the memory of the patients or their relatives or the accuracy of medical charts), limits the strength of inference clinicians can draw from case-control studies.[3] Also, case-control studies cannot provide information about the absolute risk of an event, but only about the relative risk (RR). Nevertheless, case-control studies can provide useful information and are particularly appropriate when the outcome is rare or the required duration of follow-up is long.

## ARE THE RESULTS OF THE STUDY VALID?

### Primary Guides

**Was There a Representative and Well-Defined Sample of Patients at a Similar Point in the Course of the Disease?**—This guide addresses two related issues. The first concerns how well defined the individuals in the study are, and whether they are representative of the underlying population. The authors should describe and specify their criteria for establishing that the patient has the disorder of interest (in this case, Alzheimer-type dementia) and how they selected their patient sample.

Several biases related to the assembly of the patients can distort the results of a study. For example, the sequence of referrals that leads patients from primary to tertiary centers raises the proportion of more severe or unusual cases, thus increasing the likelihood of adverse or nonfavorable outcomes. In one example of this "referral filter bias," the likelihood of a subsequent nonfebrile seizure in children with their first febrile seizure was much lower in community-based populations than in those drawn from hospitals.[6]

The second issue concerns whether the study patients are all at a similar, well-defined point in the course of their disease. Authors should provide a clear description of the stage of disease at which patients entered the study. For instance, since the duration of illness is often associated with outcome, the investigators should report the duration of illness for the sample patients. Within reason, all or most of the study patients should be at a similar point, such as survivors of a first myocardial infarction or patients newly diagnosed with lung cancer. However, the similar point in the course of disease need not be early on.

Walsh and colleagues[2] studied 126 outpatients with Alzheimer's disease who were consecutively referred to a multidisciplinary clinic for evaluation between 1980 and 1982. The diagnosis was made by consensus by a group consisting of an internist, psychiatrist, psychologist, neurologist or neuropathologist, and research nurse using the conventional *Diagnostic and Statistical Manual of Mental Disorders, Fourth Revision* criteria for dementia.[7] The tests used to exclude other causes of dementia were not described. However, given the multidisciplinary nature and expertise of the group, it seems reasonable to assume that the appropriate tests were done to exclude disorders such as hypothyroidism, depression, and space-occupying lesions of the brain that can be confused with Alzheimer's disease.

Walsh and colleagues reported survival from two different points in time: (1) referral to the clinic and (2) the point at which symptoms of memory loss were first noticed. The former is a more certain point in time, but suffers from the disadvantage that patients come to medical attention at different stages in the progression of their disease. The latter provides a more uniform starting point, but is potentially imprecise because dementia develops insidiously and the time of onset is identified retrospectively. Survival after presentation to clinic is probably more relevant for your patient's son.

**Was Follow-up Sufficiently Long and Complete?**—Since the presence of a prognostic factor often precedes the development of an outcome event by a long period, investigators must follow patients for long enough to detect the outcomes of interest. For example, recurrence in some women with early breast cancer can occur many years after initial diagnosis and treatment.[8] Patients in the dementia study were enrolled between 1980 and 1982 and followed until 1988 or their death. Thus, the follow-up was quite long, and 61% of the cohort died during this time.

Ideally, investigators will succeed in following all patients (as they did in the dementia study) but this is often not the case. Patients are not usually unavailable for follow-up for inconsequential reasons. Patients may fail to return because they have suffered exactly those events in which the investigators are interested (eg, they died or have been institutionalized). Conversely, patients who feel entirely healthy may also be less likely to return for evaluation because they feel so well. Simply put, the greater the number of patients unavailable for follow-up, the less accurate the estimate regarding the risk of the adverse outcome.

Under what circumstances does unavailability for follow-up compromise the validity of a study? It is important that you consider the relation between the proportion of patients who are unavailable and the proportion of patients who have suffered the adverse outcome of interest. The larger the number of patients whose fate is unknown relative to the number who have suffered an event, the greater the threat to the study's validity. For instance, let us assume that 30% of a particularly high-risk group (such as elderly diabetics) have suffered an adverse outcome (such as cardiovascular death) during long-term follow-up. If 5% of the patients have been lost, the true rate of patients who had died may be as high as 35%. Even if this were so, the clinical implications would not change, and the unavailability for follow-up doesn't threaten the validity of the study. However, in a much lower-risk patient sample (otherwise healthy middle-aged men, for instance) the observed event rate may be 1%. In this case, if one assumed that all 5% of the patients unavailable for follow-up had died, the event rate of 6% would have very different implications. If the number of patients unavailable for follow-up potentially jeopardizes the study's validity, you should look for the reasons for patients being unavailable, and compare the important demographic and clinical characteristics of the patients who were unavailable with the patients in whom follow-up was complete. To the extent that the reasons for disappearance are unrelated to outcome and the

patients who are unavailable are similar to those for whom information is complete, you may feel reassured. If investigators omit information about reasons for unavailability for follow-up or the characteristics of the patients who are unavailable, the strength of inference from the study results will be weaker.

### Secondary Guides

**Were Objective and Unbiased Outcome Criteria Used?**—Investigators must provide a clear and sensible definition of adverse outcomes before the study starts. Outcome events can vary from those that are objective and easily measured (death), to those that require some judgment (myocardial infarction), to those that require considerable judgment and may often be difficult to measure (eg, disability, quality of life). To minimize bias, the individual determining the outcomes should not know whether the patient had a potential prognostic factor. This is not always possible and, for unequivocal events such as death, may not be necessary. However, blinding is essential for outcomes requiring a great deal of judgment, such as transient ischemic attacks or unstable angina. In the study by Walsh and colleagues, the method and intensity of follow-up were not described. However, all patients were accounted for at the end of the study and the date of death was known for those who died.

**Was There Adjustment for Important Prognostic Factors?**—When comparing the prognosis of two groups of patients, investigators should consider whether their clinical characteristics are similar and adjust the analysis for any differences they find. The Framingham Study investigators reported that the rate of stroke in patients with atrial fibrillation and rheumatic heart disease was 41 per 1000 person years, which was very similar to the rate for patients with atrial fibrillation but no rheumatic heart disease.[9] However, patients with rheumatic heart disease were younger than those who did not have rheumatic heart disease. Once adjustments were made for the age, sex, and hypertensive status of the patients, the investigators found that the rate of stroke was sixfold greater in patients with rheumatic heart disease and atrial fibrillation than in patients with atrial fibrillation who did not have rheumatic heart disease.

Many studies of prognosis break the study group into cohorts based on suspected prognostic factors. Comparison of the pattern and frequencies of outcomes between these groups can determine the RR associated with the prognostic factor in question. For example, Pincus and colleagues[10] followed a cohort of patients with rheumatoid arthritis for 15 years. They separated the patients into a number of cohorts depending on their demographic characteristics, disease variables, and functional status. They found that some demographic variables (eg, age and education level) and functional status (eg, modified walking time and activities of daily living) were strongly predictive of mortality.

Since treatments can also alter patient outcomes, they should be taken into account when analyzing prognostic factors. While such treatments are not, strictly speaking, prognostic factors, the investigators should adjust for differences in treatment in the analysis. For example, in a study from Framingham that examined the prognosis of Q-wave vs non–Q-wave first myocardial infarction, the investigators adjusted for age, sex, and the presence of hypertension, angina pectoris, congestive heart failure, and cardiovascular disease prior to the infarct.[11] However, they did not take into account treatment with aspirin or β-blockers, which clinicians may have administered at the time of the infarct, and which we know have an impact on mortality.

In the study by Walsh and colleagues, no attempt was made to compare the mortality rate of the demented patients with a group of patients without dementia. However, they did evaluate the importance of 20 potential prognostic factors in their cohort. Age at symptom onset, dementia severity, wandering and falling, behavioral problems, and hearing loss all had a statistically significant relation to mortality.

## WHAT ARE THE RESULTS?

The quantitative results from studies of prognosis or risk are the number of events that occur over time. We will describe three common expressions of this relationship that provide complementary information about prognosis.

**How Large Is the Likelihood of the Outcome Event(s) in a Specified Period of Time?**—Your patient's son asked, "What are the chances that my mother will still be alive in 5 years?" You can provide a simple and direct answer in absolute terms. Five years after presentation to the clinic about one-half the patients (50%) had died. Thus, there is about a 50:50 chance that his mother will be alive in 5 years.

Your patient's son might then indicate that the only person he knows with Alzheimer's disease is a 65-year-old uncle who was diagnosed 10 years ago and is still living. He is surprised that his mother's chance of dying in the next 5 years is so high. This gives you the chance to discuss some of the prognostic factors for death in patients with Alzheimer's disease. As mentioned previously, the statistically significant prognostic factors for death were increasing age, dementia severity, behavioral problems, wandering and falling, and hearing loss. You explain that his mother is considerably older than his uncle was at the time of diagnosis, and that this likely explains some of the difference. It would be nice to use the prognostic factors to further refine the chance of death in his mother. Her age is almost identical to the mean age of the cohort studied by Walsh and colleagues. However, her Mini-Mental State Examination score is quite low (indicating more severe dementia), and her behavioral problems also suggest that she is at higher risk than the average patient in the study by Walsh et al. Unfortunately, no table or formula was presented that allows you to combine all of these factors and estimate a risk of mortality that is specific for your patient. However, you can feel confident in telling her son that his moth-

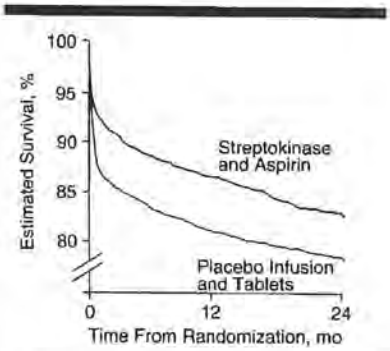



Fig 1.—Survival after myocardial infarction in patients treated with streptokinase and aspirin compared with placebo. Adapted from ISIS-2 (Second International Study of Infarct Survival) Collaborative Group. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17 187 cases of suspected acute myocardial infarction: ISIS-2. 13:349-360. © by The Lancet Ltd, 1988.

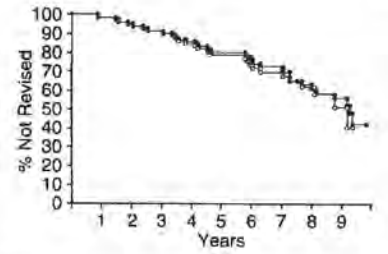



Fig 2.—Need for revision after total hip arthroplasty in two cohorts of patients in the same center (adapted from Dorey and Amstutz[14]).

er's chances of dying are at least 50% during the next 5 years, and probably greater.

The son might then ask whether his mother's chances of survival could change with time. Neither the absolute nor relative expressions of results address this question. For this answer we should turn to a survival curve, a graph of the number of events over time (or conversely, the chance of being free of these events over time).[12] The events must be discrete (eg, death, stroke, recurrence of cancer), and the time at which they occur must be precisely known. In most clinical situations the chance of an outcome changes with time. Figures 1 and 2 show two survival curves, one of survival after a myocardial infarction[13] and the other the results of hip replacement surgery.[14] Note that the chance of dying after a myocardial infarction is highest shortly after the event (reflected by an initially steep slope of the curve, which then flattens), while very few hip replacements require revision until much later (this curve starts out flat and then steepens). Walsh and colleagues provided a survival curve in Fig 1 of their article that suggests that the chance of dying is more or less constant during the first 7 years after referral to the clinic for dementia.

**How Precise Are the Estimates of Likelihood?**—Even when valid, a prognostic study provides only an estimate of the true risk. After determining the size of the risk, we should next examine the precision of the estimate, which is best done with a confidence interval (CI). Walsh and colleagues found that the 95% CI for survival 5 years after presentation was approximately 39% to 58% (extrapolated from Fig 1 in their article). Note that in most survival curves, the earlier follow-up periods usually include results from more patients than the later periods (because of unavailability for follow-up and because patients are not enrolled into the study at the same time). This means that the survival curves are more precise in the earlier periods, indicated by narrower confidence bands around the left-hand parts of the curve.

Walsh and colleagues also provided 95% CIs for the RR associated with each prognostic factor. For example, the RR associated with a behavioral problem was 1.5 with a 95% CI of 1.0 to 2.5. This means that the best estimate is that a patient with a behavioral problem is 1.5 times more likely to die than an individual without a behavioral problem. The probability that the true RR is between 1.0 (ie, no effect) and 2.5 is 95%.

## WILL THE RESULTS HELP ME IN CARING FOR MY PATIENTS?

**Were the Study Patients Similar to My Own?**—How well do the study results generalize to the patients in your practice? The authors should describe the study patients in enough detail to allow comparison with your patients. The article should list the patients' important clinical characteristics, along with the definitions used for these characteristics. The closer the match between the patient before you and those in the study, the more confident you can be in applying the study results to that patient. The characteristics of the study patients were quite similar to your patient.

**Will the Results Lead Directly to Selecting or Avoiding Therapy?**—Since there are no therapies for dementia that are routinely available and clearly effective, this guide does not directly apply to your patient. However, prognostic data often provide the basis for sensible decisions about therapy. Knowing the expected clinical course of your patient's condition can help you judge whether treatment should be offered at all. For example, warfarin markedly decreases the risk of stroke in patients with nonrheumatic atrial fibrillation and is indicated for many patients with this disorder.[15] However, in one study the frequency of stroke in patients with "lone" atrial fibrillation (patients 60 years of age or younger with no associated cardiopulmonary disorders) was 1.3% over 15 years.[16] The risks of long-term warfarin therapy in this group of patients probably outweigh the benefits.

**Are the Results Useful for Reassuring or Counseling Patients?**—Even if the prognostic result does not lead you to prescribe an effective therapy, it can still be clinically useful. A valid, precise, and generalizable result of uniformly good prognosis is very helpful to the clinician when reassuring a concerned patient or relative. Some conditions, such as asymptomatic hiatal hernia or asymptomatic colonic diverticula, have such a good overall prognosis that they have been termed "nondisease."[17] On the other hand, a prognostic result of uniformly bad prognosis provides the clinician with a starting place for a discussion with the patient and family, leading to counseling about end-of-life concerns.

In your patient, information on the likelihood of death will be useful to the son and his family as they plan the future care of his mother. Of course, other prognostic information about the rate of progression of the dementing process and the need for intensive nursing care would also be useful.[18,19]

We thank Malcolm Hing, MD, for his comments and Karen Weeks for secretarial assistance.

### References


1. Molloy DW, Alemayehu E, Roberts R. Reliability of a standardized Mini-Mental State Examination compared with the traditional Mini-Mental State Examination. *Am J Psychiatry*. 1991;148:102-105.
2. Walsh JS, Welch G, Larson EB. Survival of outpatients with Alzheimer-type dementia. *Ann Intern Med*. 1990;113:429-434.
3. Levine M, Walter S, Lee H, et al, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, IV: how to use an article about harm. *JAMA*. 1994;271:1615-1619.
4. Dawber TR, Kaunel WB, Lyell LP. An approach to longitudinal studies in a community: the Framingham Study. *Ann N Y Acad Sci*. 1963;107:539-556.
5. Bennett JC, for the Board on Health Sciences Policy of the Institute of Medicine. Inclusion of women in clinical trials: policies for population subgroups. *N Engl J Med*. 1993;329:288-292.
6. Ellenberg JH, Nelson KB. Sample selection and the natural history of disease: studies of febrile seizures. *JAMA*. 1980;243:1337-1340.
7. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. 4th ed. Washington DC; 1994:139-143.
8. Early Breast Cancer Trialists' Collaborative Group. Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy: 133 randomised trials involving 31 000 recurrences and 24 000 deaths among 75 000 women. *Lancet*. 1992; 339:1-15.
9. Wolf PA, Dawber TR, Thomas HE, Kannel WB. Epidemiologic assessment of chronic atrial fibrillation and risk of stroke: the Framingham Study. *Neurology*. 1978;28:973-977.
10. Pincus T, Brooks RH, Callahan LF. Prediction of long-term mortality in patients with rheumatoid arthritis according to simple questionnaire and joint count measures. *Ann Intern Med*. 1994;120:26-34.
11. Berger CJ, Murabito JM, Evans JC, Anderson KM, Levy D. Prognosis after first myocardial infarction: comparison of Q-wave and non–Q-wave myocardial infarction in the Framingham Heart Study. *JAMA*. 1992;228:1545-1551.
12. Katz MH, Hauck WW. Proportional hazards (Cox) regression. *J Gen Intern Med*. 1993;8:702-711.
13. ISIS-2 (Second International Study of Infarct Survival) Collaborative Group. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17 187 cases of suspected acute myocardial infarction: ISIS-2. *Lancet*. 1988;2:349-360.
14. Dorey F, Amstutz H. The validity of survivorship analysis in total joint arthroplasty. *J Bone Joint Surg Am*. 1989;71:544-548.
15. Laupacis A, Albers G, Dunn M, Feinberg W. Antithrombotic therapy in atrial fibrillation. *Chest*. 1992;102:426S-433S.
16. Kopecky SL. The natural history of lone atrial fibrillation: a population-based study over three decades. *N Engl J Med*. 1987;317:669-674.
17. Meador CK. The art and science of nondisease. *N Engl J Med*. 1965;272:92.
18. Stern Y, Mayeux R, Hauser WA, Bush T. Predictors of disease course in patients with probable Alzheimer's disease. *Neurology*. 1987;37:1649-1653.
19. Drachman DA, O'Donnell BF, Lew RA, Swearer JM. The prognosis in Alzheimer's disease: 'how far' rather than 'how fast' best predicts the course. *Arch Neurol*. 1990;47:851-856.

# Users' Guides to the Medical Literature

## VI. How to Use an Overview

Andrew D. Oxman, MD, MSc; Deborah J. Cook, MD, MSc; Gordon H. Guyatt, MD, MSc;
for the Evidence-Based Medicine Working Group

### CLINICAL SCENARIO

A 55-year-old man had his serum cholesterol level measured at a shopping mall 2 months ago. His cholesterol level was elevated and he comes to you, his primary care physician, for advice. He does not smoke, is not obese, and does not have hypertension, diabetes mellitus, or any first-order relatives with premature coronary heart disease (CHD). You repeat his cholesterol test and schedule a follow-up appointment. The test confirms an elevated cholesterol level (7.9 mmol/L [305 mg/dL]), but before deciding on a treatment recommendation, you elect to find out just how big a reduction in the risk of CHD this patient could expect from a cholesterol-lowering diet or drug therapy.

### THE SEARCH

There are a number of cholesterol-lowering trials, and instead of trying to find and review all of the original studies yourself, you use Grateful Med to find a recent overview. On the first subject line you select hypercholesterolemia or cholesterol from the list of Medical Subject Headings (MeSH) used to index articles. On the second subject line you use the MeSH term coronary disease, which you explode to capture articles that are indexed with more specific terms that come under coronary disease, such as myocardial infarction. You limit your search to English-language articles, and to find a quantitative review, you use the term meta-analysis on the line for publication type. Titles and abstracts suggest two of the nine references from this search are definitely on target, and you decide to examine both.[1,2]

From the Departments of Clinical Epidemiology and Biostatistics (Drs Oxman, Cook, and Guyatt), Medicine (Drs Cook and Guyatt), and Family Medicine (Dr Oxman), McMaster University, Hamilton, Ontario.

A complete list of members (with affiliations) of the Evidence-Based Medicine Working Group appears in the first article of this series (*JAMA*. 1993;270:2093-2095). The following members contributed to this article: Eric Bass, MD, MPH; Patrick Brill-Edwards, MD; George Browman, MD, MSc; Allan Detsky, MD, PhD; Michael Farkouh, MD; Hertzel Gerstein, MD, MSc; Ted Haines, MD, MSc; Brian Haynes, MD, MSc; Robert Hayward, MD, MPH; Anne Holbrook, MD, PharmD, MSc; Roman Jaeschke, MD, MSc; Elizabeth Juniper, MCSP, MSc; Andreas Laupacis, MD, MSc; Hui Lee, MD, MSc; Mitchell Levine, MD, MSc; Virginia Moyer, MD, MPH; David Naylor, MD, DPhil; Jim Nishikawa, MD; Ameen Patel, MD; John Philbrick, MD; Scott Richardson, MD; Stephane Sauve, MD, MSc; David Sackett, MD, MSc; Jack Sinclair, MD; Brian Strom, MD, MPH; K. S. Trout, FRCE; Sean Tunis, MD, MSc; Stephen Walter, PhD; John Williams, Jr, MD, MHS; and Mark Wilson, MD, MPH.

Reprint requests to Room 2C12, McMaster University Health Sciences Centre, 1200 Main St W, Hamilton, Ontario, Canada L8N 3Z5 (Dr Guyatt).

### INTRODUCTION

Systematic overviews of the medical literature that summarize scientific evidence (in contrast to unsystematic narrative reviews that mix together opinions and evidence) are becoming increasingly prevalent. These overviews address questions of treatment, causation, diagnosis, or prognosis. In each case, the rules for deciding whether the overviews are credible, and for interpreting their results, are similar. In this article, we provide guidelines for distinguishing a good overview from a bad one and for using the results. In doing so, we will ask the same key questions that we have suggested for original reports of research[3]: Are the results valid? If they are, what are the results, and will they be helpful in my patient care (Table 1)?

Authors sometimes use the terms "systematic review," "overview," and "meta-analysis" interchangeably. We use overview as a term for any summary of the medical literature and meta-analysis as a term for reviews that use quantitative methods to summarize the results. Investigators must make a host of decisions in preparing an overview, including determining the focus; identifying, selecting, and critically appraising the relevant studies (which we will call the "primary studies"); collecting and synthesizing (either quantitatively or nonquantitatively) the relevant information; and drawing conclusions. Avoiding errors in both meta-analyses and other overviews requires a systematic approach, and enabling users to assess the validity of an overview's results requires explicit reporting of the methods. A number of authors have recently examined issues pertaining to the validity of overviews.[4-7] In this article we will emphasize key points from the perspective of a clinician needing to make a decision about patient care.

You can use the first two validity guides in Table 1 to quickly screen out most published review articles.[7] The discrepancies between the results of systematic meta-analyses and the recommendations made by clinical experts in nonsystematic review articles[8] reflects the limited validity of most published review articles. Archie Cochrane pointed out the need for more systematic overviews when he wrote: "It is surely a great criticism of our profession that we have not organised a critical summary, by specialty or subspecialty, adapted periodically, of all relevant randomised controlled trials [RCTs]."[9] The Cochrane Collaboration, an international effort to prepare, maintain, and disseminate systematic reviews of the effects of health care, has evolved in response to this challenge.[10,11] As the Collaboration develops, you will find more and more systematic reviews of RCTs addressing important issues in patient management.

### ARE THE RESULTS OF THE OVERVIEW VALID?

#### Primary Guides

**Did the Overview Address a Focused Clinical Question?**—Unless an overview clearly states the question it addresses, you can only guess whether it is pertinent to your patient care. Most clinical questions can be formulated in terms of a simple relationship between the patient, some exposure (to a treatment, a diagnostic test, a potentially harmful agent, and the like), and one or more outcomes of interest. If the main question that an overview addresses is not clear from the title or abstract, it is probably a good idea to move on to the next article.

Users' Guides to the Medical Literature section editor: Drummond Rennie, MD, Deputy Editor (West), *JAMA*.

Table 1.—Users' Guides for How to Use Review Articles

| |
|---|
| **Are the results of the study valid?** |
| Primary guides: |
| Did the overview address a focused clinical question? |
| Were the criteria used to select articles for inclusion appropriate? |
| Secondary guides: |
| Is it unlikely that important, relevant studies were missed? |
| Was the validity of the included studies appraised? |
| Were assessments of studies reproducible? |
| Were the results similar from study to study? |
| **What are the results?** |
| What are the overall results of the review? |
| How precise were the results? |
| **Will the results help me in caring for my patients?** |
| Can the results be applied to my patient care? |
| Were all clinically important outcomes considered? |
| Are the benefits worth the harms and costs? |

Table 2.—Guides for Selecting Articles That Are Most Likely to Provide Valid Results*

| | |
|---|---|
| Therapy | • Was the assignment of patients to treatments randomized?<br>• Were all of the patients who entered the trial properly accounted for and attributed at its conclusion? |
| Diagnosis | • Was there an independent, blind comparison with a reference standard?<br>• Did the patient sample include an appropriate spectrum of the sort of patients to whom the diagnostic test will be applied in clinical practice? |
| Harm | • Were there clearly identified comparison groups that were similar with respect to important determinants of outcome, other than the one of interest?<br>• Were outcomes and exposures measured in the same way in the groups being compared? |
| Prognosis | • Was there a representative and well-defined sample of patients at a similar point in the course of disease?<br>• Was follow-up sufficiently long and complete? |

*From Oxman et al.[3]

Many overviews address a number of questions. For example, a review article or a chapter from a textbook might include sections on the etiology, diagnosis, prognosis, treatment, and prevention of asthma. While such broad reviews can provide a useful introduction to an area, they usually offer limited support for their conclusions. Typically, you will find only a declarative statement followed by one or more citations. You must then study the references in order to judge the validity of the authors' conclusions.

**Were the Criteria Used to Select Articles for Inclusion Appropriate?**—To determine if the investigators reviewed the appropriate research, the reader needs to know the criteria they used to select research. These criteria should specify the patients, exposures, and outcomes of interest. They should also specify the methodologic standards used to select studies, and these standards should be similar to the primary validity criteria we have described for original reports of research[3] (Table 2).

In looking at the effectiveness of lowering cholesterol on CHD, investigators might restrict themselves to studies of patients who did not have clinically manifest CHD at the beginning of the study (primary prevention), to studies of patients who already had symptomatic CHD (secondary prevention), or include both. They might include only trials of diet therapy, only trials of drug therapy, or both. They might consider several different outcomes, such as nonfatal CHD, CHD mortality, and total mortality. With respect to methodologic criteria, they might consider only RCTs or include observational studies.

Differences in the patients, exposures, and outcomes can lead to different results among overviews that appear to address the same clinical question.[12] The clinician must be sure the criteria used to select the studies correspond to the clinical question that led her to the article in the first place. The impact of cholesterol-lowering strategies, for instance, differs in studies of primary vs secondary prevention.[1,2]

If the authors state their inclusion criteria, it is less likely they will (as they are wont to do) preferentially cite studies that support their own prior conclusion. Bias in choosing articles to cite is a problem for both overviews and original reports of research (in which the discussion section often includes comparisons with the results of other studies). Gøtzsche, for example, reviewed citations in reports of trials of new nonsteroidal anti-inflammatory drugs in rheumatoid arthritis.[13] Among 77 articles where the authors could have referenced other trials with and without outcomes favoring the new drug, nearly 60% (44) cited a higher proportion of the trials with favorable outcomes. In 22 reports of controlled trials of cholesterol lowering, Ravnskov[14] found a similar bias toward citing positive studies.

## Secondary Guides

**Is It Unlikely That Important Relevant Studies Were Missed?**—It is important that authors conduct a thorough search for studies that meet their inclusion criteria. This should include the use of bibliographic databases, such as MEDLINE and EMBASE, checking the reference lists of the articles they retrieved, and personal contact with experts in the area. Unless the authors tell us what they did to locate relevant studies, it is difficult to know how likely it is that relevant studies were missed.

There are two important reasons why a review's authors should use personal contacts. The first is so they can identify published studies that might have been missed (including studies that are in press or not yet indexed or referenced). The second is so they can identify unpublished studies. Although the inclusion of unpublished studies is controversial,[15] their omission increases the chances of "publication bias"—a higher likelihood for studies with positive results to be published[16-19] and the attendant risk for the review to overestimate efficacy or adverse effects.

If investigators include unpublished studies in an overview, they should obtain full written reports and appraise the validity of both published and unpublished studies; they may also use statistical techniques to explore the possibility of publication bias.[20] Overviews based on a small number of small studies with weakly positive effects are the most susceptible to publication bias.

**Was the Validity of the Included Studies Appraised?**—Even if a review article includes only RCTs, it is important to know whether they were of good quality. Unfortunately, peer review does not guarantee the validity of published research.[21] For exactly the same reason that the guides for using original reports of research begin by asking if the results are valid, it is essential to consider the validity of research included in overviews.

Differences in study methods might explain important differences among the results.[22,23] For example, less rigorous studies tend to overestimate the effectiveness of therapeutic and preventive interventions.[24] Even if the results of different studies are consistent, it is still important to know how valid the studies are. Consistent results are less compelling if they come from weak studies than if they come from strong studies.

There is no one correct way to assess validity. Some investigators use long checklists to evaluate methodologic quality, while others focus on three or four key aspects of the study. You will remember that in our previous articles about therapy, diagnosis, and prognosis in the Users' Guides series, we asked the question, "Is the study valid?" and presented criteria to help you answer these questions. When considering whether to believe the results of an overview, you should check whether the authors examined criteria similar to those we have presented in deciding on the credibility of their primary studies (Table 2).

**Were Assessments of Studies Reproducible?**—As we have seen, authors of review articles must decide which studies to include, how valid they are, and which data to extract from them. Each of these decisions requires judgment by

Table 3.—Assessments of Overviews From the Clinical Scenario*

| Criterion | Davey Smith et al,[1] 1993 | Silberberg and Henry,[2] 1991 |
|---|---|---|
| **Are the results of the study valid?** | | |
| Did the overview address a focused clinical question? | Yes: to examine effects of cholesterol lowering on mortality in relationship to baseline risk of CHD death | Yes: to examine effects of drug treatment to lower cholesterol in primary and secondary prevention of CHD events |
| Were the criteria used to select articles for inclusion appropriate? | Yes, although inclusion of trials of estrogen and surgery can be questioned: single-factor (dietary interventions, lipid-lowering drugs [including estrogen] or surgery) RCTs of cholesterol lowering with ≥6 mo follow-up and at least 1 death—35 trials, 57 124 patients | Yes, although exclusion of nondrug trials could be questioned: single-factor RCTs of drug treatments (excluding trials of estrogen and thyroxine)—9 trials, 26 609 patients |
| Is it unlikely that important relevant studies were missed? | Yes: MEDLINE, previous overviews, and personal contact with investigators were used to identify studies | Can't tell: MEDLINE and previous overviews were used to identify studies; investigators were not contacted, non–English-language publications and unpublished data were not included |
| Was the validity of the included studies appraised? | No | No |
| Were assessments of studies reproducible? | Can't tell | Yes: data were extracted independently by two reviewers |
| Were the results similar from study to study? | Probably not (test of homogeneity not reported): baseline risk of CHD death and percent reduction in cholesterol levels hypothesized as explanation for variation in effect of treatment | Probably not (test of homogeneity not reported), but pooled ORs for primary and secondary prevention studies respectively were similar: baseline risk hypothesized as explanation for variation in absolute risk reduction |
| **What are the results?** | | |
| What are the overall results of the review? | For total mortality, the OR (and 95% CI) was 0.74 (0.60-0.92 for high-risk groups [>50 deaths/1000 person-years in the control group]), 0.96 (0.84-1.09) for medium-risk groups (10-50 deaths/1000 person-years), and 1.22 (1.06-1.42) in low-risk groups (<10 deaths/1000 person-years) | For CHD death, the OR (and 95% CI) was 0.85 (0.64-1.14) in primary prevention and 0.84 (0.75-0.95) in secondary prevention studies; the NNT to prevent one death from CHD was 675 and 38 in the primary and secondary trials, respectively |
| How precise were the results? | | |

*CHD indicates coronary heart disease; RCTs, randomized controlled trials; OR, odds ratio; CI, confidence interval; and NNT, number needed to treat.

the reviewers and each is subject to both mistakes (random errors) and bias (systematic errors). Having two or more people participate in each decision guards against errors, and if there is good agreement among the reviewers, the clinician can have more confidence in the results of the overview.

**Were the Results Similar From Study to Study?**—Despite restrictive inclusion criteria, most systematic overviews document important differences in patients, exposures, outcome measures, and research methods from study to study. Readers must decide when these factors are so different that it no longer makes sense to combine the study results.

One criterion for deciding to combine results quantitatively is whether the studies seem to be measuring the same underlying magnitude of effect. In meta-analyses, investigators can test the extent to which differences among the results of individual studies are greater than you would expect if all studies were measuring the same underlying effect and the observed differences were due only to chance. The statistical analyses that are used to do this are called "tests of homogeneity."

The more significant the test of homogeneity, the less likely it is that the observed differences in the size of the effect are due to chance alone. Both the "average" effect and the confidence interval (CI) around the average effect need to be interpreted cautiously when there is "statistically significant" heterogeneity (a low probability of the differences in results from study to study being due to chance alone, indicating that differences in patients, exposures, outcomes, or study design are responsible for the varying treatment effect).

Unfortunately, a nonsignificant test does not necessarily rule out important heterogeneity. Hence, clinically important differences between study results still dictate caution in interpreting the overall findings, despite a nonsignificant test of homogeneity. However, even when there are large differences between the results of different studies, a summary measure from all of the best available studies may provide the best estimate of the impact of the intervention or exposure.[25-27]

Neither of the two overviews identified in the scenario reported a test of homogeneity. However, both of them included graphic and tabular displays of the results of the primary studies that suggest differences in study results that are likely to be both clinically important and statistically significant. Both of the overviews suggest possible explanations for the observed heterogeneity (Table 3).

## WHAT ARE THE RESULTS?

**What Are the Overall Results of the Overview?**—In clinical research, investigators collect data from individual patients. Because of the limited capacity of the human mind to handle large amounts of data, investigators use statistical methods to summarize and analyze them. In overviews, investigators collect data from individual studies. These data must also be summarized, and increasingly, investigators are using quantitative methods to do so.

Simply comparing the number of positive studies with the number of negative studies is not an adequate way to summarize the results. With this sort of "vote counting," large and small studies are given equal weights, and (unlikely as it may seem) one investigator may interpret a study as positive, while another investigator interprets the same study as negative.[28] For example, a clinically important effect that is not statistically significant could be interpreted as positive in light of clinical importance and negative in light of statistical significance. There is a tendency to overlook small but clinically important effects if studies with statistically nonsignificant (but potentially clinically important) results are counted as negative.[29] Moreover, a reader cannot tell anything about the magnitude of an effect from a vote count even when studies are appropriately classified using additional categories for studies with a positive or negative trend.

Typically, meta-analysts weight studies according to their size, with larger studies receiving more weight. Thus, the overall results represent a weighted average of the results of the individual studies. Occasionally studies are also given more or less weight depending on their quality, or poorer quality studies might be given a weight of zero (excluded) either in the primary analysis or in a "sensitivity analysis" to see if this makes an important difference in the overall results.

Table 4.—Odds Ratio, Relative Risk, Risk Reduction, and Number Needed to Treat

| Treatment or Exposure | Adverse Outcome* | |
|---|---|---|
| | Positive | Negative |
| Positive | *A* | *B* |
| Negative | *C* | *D* |

*When the outcome is undesirable, a relative risk (RR) or odds ratio (OR) of <1.0 represents a beneficial treatment or exposure, with zero representing 100% effectiveness. An absolute risk reduction (ARR) of <0 represents a benefit, and 100% effectiveness would be equivalent to the risk observed in the control group. The OR can also be expressed as $(A/C)\div(B/D)$ (ie, the odds of a case having been exposed relative to the odds of a control having been exposed), and both of these expressions are equivalent to $(A \cdot D)\div(B \cdot C)$. From the two expressions, if *A* is small relative to *B* and *C* is small relative to *D*, the OR and the RR are approximately the same.
Thus,

OR=$(A/B)/(C/D)$
RR=$[A/(A+B)]/[C/(C+D)]$
RR reduction=1−RR
ARR=$[A/(A+B)]-[C/(C+D)]$
Number needed to treat=1/ARR

You should look to the overall results of an overview the same way you look to the results of primary studies. In our articles concerning therapy, we described the relative risk and the absolute risk reduction, and how they could be interpreted.[30] In the articles about diagnostic tests, we discussed likelihood ratios.[31] In overviews of treatment and etiologic and prognostic factors, you will often see the ratio of the odds of an adverse outcome occurring in those exposed (to a treatment or risk factor) to the odds of an adverse outcome in those not exposed. This odds ratio, illustrated in Table 4, has desirable statistical properties when combining results across studies. Whatever method of analysis the investigators used, you should look for a summary measure (such as the number needed to treat[32]) that clearly conveys the practical importance of the result.

Sometimes the outcome measures that are used in different studies are similar but not exactly the same. For example, different trials might measure functional status using different instruments. If the patients and the interventions are reasonably similar, it might still be worthwhile to estimate the average effect of the intervention on functional status. One way of doing this is to summarize the results of each study as an "effect size."[33] The effect size is the difference in outcomes between the intervention and control groups divided by the standard deviation (SD). The effect size summarizes the results of each study in terms of the number of SDs of difference between the intervention and control groups. Investigators can then calculate a weighted average of effect sizes from studies that measured an outcome in different ways.

You are likely to find it difficult to interpret the clinical importance of an effect size (if the weighted average effect is one half of an SD, is this effect clinically trivial, or is it large?). Once again, you should look for a presentation of the results that conveys their practical importance (for example, by translating the summary effect size back into natural units).[34] For instance, if clinicians have become familiar with the significance of differences in walk test scores in patients with chronic lung disease, the effect size of a treatment on a number of measures of functional status (such as the walk test and stair climbing) can be converted back into differences in walk test scores.

Although it is generally desirable to have a quantitative summary of the results of a review, it is not always appropriate. For example, there may be unexplained heterogeneity in study results or the studies may be of such poor quality that the overall results would be uninterpretable. In these cases investigators should still present tables or graphs that summarize the results of the primary studies, and their conclusions should be cautious.

**How Precise Were the Results?**— In the same way that it is possible to estimate the average effect across studies, it is possible to estimate a CI around that estimate; ie, a range of values with a specified probability (typically 95%) of including the true effect. A previous article in this series provides a guide for understanding CIs.[30]

## WILL THE RESULTS HELP ME IN CARING FOR MY PATIENTS?

**Can the Results Be Applied to My Patient Care?**—One of the advantages of an overview is that since it includes many studies, the results come from a very diverse range of patients. If the results are consistent across studies, they apply to this wide variety of patients. Even so, the clinician may still be left with doubts about the applicability of the results. Perhaps the patient is older than any of those included in the individual trials summarized by the overview. If studies using different members of a class of drug have been combined, one might question whether one of the drugs has a larger effect than the others.

These questions raise the issue of subgroup analysis. Detailed guides for deciding whether to believe subgroup analyses are available.[26,27] One of the most important guides is that conclusions that are drawn on the basis of between-study comparisons (comparing patients in one study with patients in another) should be viewed skeptically. For example, meta-analysis of the effectiveness of β-blockers after myocardial infarction found a statistically significant and clinically important difference in effect between trials of β-blockers with and without intrinsic sympathomimetic activity.[35] This resulted in clinical recommendations that only β-blockers without intrinsic sympathomimetic activity should be used. However, the addition of two subsequent trials eliminated this difference in the overall summary.[25] In fact, a large number of subgroup analyses exploring differences in either patients or the β-blocker regimen used suggest that any apparent differences are probably due to chance.[25]

Other criteria that make a hypothesized difference in subgroups more credible include a big difference in treatment effect; a highly statistically significant difference in treatment effect (the lower the *P* value on the comparison of the different effect sizes in the subgroups, the more credible the difference); a hypothesis that was made before the study began and was one of only a few hypotheses that were tested; consistency across studies; and indirect evidence in support of the difference ("biological plausibility"). If these criteria are not met, the results of a subgroup analysis are less likely to be trustworthy and you should assume that the overall effect across all patients and all treatments, rather than the subgroup effect, applies to the patient at hand and to the treatment under consideration.

**Were All Clinically Important Outcomes Considered?**—While it is a good idea to look for focused review articles because they are more likely to provide valid results, this does not mean that you should ignore outcomes that are not included in a review. For example, the potential benefits and harms of hormone replacement therapy include reduced risk of fractures and CHD and increased risk of breast cancer and endometrial cancer. Focused reviews of the evidence for individual outcomes are more likely to provide valid results, but a clinical decision requires considering all of them.

**Are the Benefits Worth the Harms and Costs?**—Finally, either explicitly or implicitly, when making a clinical de-

cision the expected benefits must be weighed against the potential harms and costs. While this is most obvious for deciding whether to use a therapeutic or preventive intervention, providing patients with information about causes of disease or prognosis can also have both benefits and harms. For example, informing a woman about potentially teratogenic exposures might result in her reducing her risk of exposure (with potential benefits), and also cause anxiety or loss of work. Informing an asymptomatic woman with newly detected cancer about her prognosis might help her to plan better, but also label her, cause anxiety, or increase the period during which she is "sick."

A valid review article provides the best possible basis for quantifying the expected outcomes, but these outcomes still must be considered in the context of your patient's values and concerns about the expected outcomes of a decision. In the next articles in this series we will address this issue in the context of decision analysis and clinical practice guidelines.

## RESOLUTION OF THE SCENARIO

The two meta-analyses identified in the clinical scenario at the beginning of this article had different objectives, which resulted in differences in the trials included and the analyses. Nonetheless, the results of the two overviews support similar conclusions. Both meta-analyses meet most validity criteria (Table 3), with some limitations. Neither one included explicit assessments of the validity of the primary studies, one does not state whether more than one of the reviewers assessed the studies independently,[1] and there appears (in graphic and tabular displays of the results) to be clinically important heterogeneity among the studies included in both of the overviews. Of more concern than any bias in the overviews, however, is how we should interpret the data on noncardiac and total mortality, an issue that is also the focus of other published meta-analyses of these data,[14,36,37] including one that was published but not yet indexed at the time of the scenario.[38] Despite the uncertainty, it is still possible to draw some reasonable conclusions.

The benefit we can expect from interventions to lower cholesterol depends on the baseline risk of death from CHD and, possibly, whether we are considering dietary or drug interventions.[1,2,38,39] The higher the risk of dying of CHD, the greater the likelihood of benefit. Drug, but not dietary therapy, may be associated with a greater likelihood of death from causes other than CHD.[1,38] The patient described in the scenario has a risk of dying of CHD of approximately 1.0% over the next decade. You would have to treat approximately 1000 such patients for 10 years with a dietary intervention to save a life. If you were to treat such patients with drug therapy, it is not certain you would reduce total mortality. The results suggest that drug therapy should be restricted to those at high risk, such as individuals with known coronary artery disease, and whether diet therapy is worthwhile in low-risk individuals (such as the patient in the scenario) is uncertain.

### References


1. Davey Smith G, Song F, Sheldon TA. Cholesterol lowering and mortality: the importance of considering initial level of risk. *BMJ*. 1993;306:1367-1373.
2. Silberberg JS, Henry DA. The benefits of reducing cholesterol levels: the need to distinguish primary from secondary prevention, I: a meta-analysis of cholesterol-lowering trials. *Med J Aust*. 1991; 155:665-666, 669-670.
3. Oxman AD, Sackett DL, Guyatt GH, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, I: how to get started. *JAMA*. 1993;270:2093-2095.
4. L'Abbe KA, Detsky AS, O'Rourke K. Meta-analysis in clinical research. *Ann Intern Med*. 1987;107: 224-233.
5. Sacks HS, Berries J, Reitman D, Arcona-Berk VA, Chalmers TC. Meta-analyses of randomized controlled trials. *N Engl J Med*. 1987;316:450-455.
6. Oxman AD, Guyatt GH. Guidelines for reading literature reviews. *Can Med Assoc J*. 1988;138:697-703.
7. Mulrow CD. The medical review article: state of the science. *Ann Intern Med*. 1987;106:485-488.
8. Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts: treatments for myocardial infarction. *JAMA*. 1992;268:240-248.
9. Cochrane AL. 1931-1971: a critical review, with particular reference to the medical profession. In: *Medicines for the Year 2000*. London, England: Office of Health Economics; 1979:2-12.
10. *The Cochrane Collaboration*. Oxford, England: UK Cochrane Centre, National Health Service Research and Development Programme; 1994. Brochure.
11. Enkin MW, Keirse MJNC, Renfrew MJ, Neilson JP, eds. Cochrane Pregnancy and Childbirth Database [derived from the Cochrane Database of Systematic Reviews], Cochrane Updates on Disk, Update Software. Oxford, England: UK Cochrane Centre; 1993: disk issue 2.
12. Chalmers TC, Berrier J, Sacks HS, et al. Meta-analysis of clinical trials as a scientific discipline, II: replicate variability and comparison of studies that agree and disagree. *Stat Med*. 1987;6:733-744.
13. Gøtzsche PC. Reference bias in reports of drug trials. *BMJ*. 1987;295:654-656.
14. Ravnskov U. Cholesterol lowering trials in coronary heart disease: frequency of citation and outcome. *BMJ*. 1992;305:15-19.
15. Cook DJ, Guyatt GH, Ryan G, et al. Should unpublished data be included in meta-analyses? current convictions and controversies. *JAMA*. 1993; 269:2749-2753.
16. Dickersin K. The existence of publication bias and risk factors for its occurrence. *JAMA*. 1990; 263:1385-1389.
17. Dickersin K, Min Y-I, Meinert CL. Factors influencing publication of research results: follow-up of applications submitted to two institutional review boards. *JAMA*. 1992;267:374-378.
18. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet*. 1991;337:867-872.
19. Dickersin K, Min Y-I. NIH clinical trials and publication bias. *Online J Curr Clin Trials [serial online]*. April 28, 1993;50.
20. Begg CB, Berlin JA. Publication bias: a problem in interpreting medical data. *J R Stat Soc*. 1988;151:445-463.
21. Williamson JW, Goldschmidt PG, Colton T. The quality of medical literature: analysis of validation assessments. In: Bailar JC, Mosteller F, eds. *Medical Uses of Statistics*. Waltham, Mass: NEJM Books; 1986:370-391.
22. Horwitz RI. Complexity and contradiction in clinical trial research. *Am J Med*. 1987;82:498-510.
23. Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbe A. Incorporating variations in the quality of individual randomized trials into meta-analysis. *J Clin Epidemiol*. 1992;45:255-265.
24. Chalmers TC, Celano P, Sacks HS, Smith H. Bias in treatment assignment in controlled trials. *N Engl J Med*. 1983;309:1358-1361.
25. Peto R. Why do we need systematic overviews of randomized trials? *Stat Med*. 1987;6:223-240.
26. Yusuf S, Wiftes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA*. 1991;266:93-98.
27. Oxman AD, Guyatt GH. A consumer's guide to subgroup analysis. *Ann Intern Med*. 1992;116:78-84.
28. Glass GV, McGaw B, Smith ML. *Meta-analysis in Social Research*. Newbury Park, Calif: Sage; 1981:18-20.
29. Cooper RM, Rosenthal R. Statistical versus traditional procedures for summarizing research findings. *Psychol Bull*. 1980;87:442-449.
30. Guyatt GH, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, II: how to use an article about therapy or prevention, B: what were the results and will they help me in caring for my patients? *JAMA*. 1994;271:59-63.
31. Jaeschke R, Guyatt GH, Sackett DL, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, III: how to use an article about a diagnostic test, B: what are the results and will they help me in caring for my patients? *JAMA*. 1994;271:703-707.
32. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *N Engl J Med*. 1988;318: 1728-1733.
33. Rosenthal R. *Meta-analytic Procedures for Social Research*. 2nd ed. Newbury Park, Calif: Sage; 1991.
34. Smith K, Cook DJ, Guyatt GH, Mudhaven J, Oxman AD. Respiratory muscle training in chronic airflow limitation: a meta-analysis. *Am Rev Respir Dis*. 1992;145:533-539.
35. Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Prog Cardiovasc Dis*. 1985;27:335-371.
36. Muldoon MF, Manuck SB, Matthews KA. Lowering cholesterol concentrations and mortality: a quantitative review of primary prevention trials. *BMJ*. 1990;301:309-314.
37. Davey Smith G, Pekkanen J. Should there be a moratorium on the use of cholesterol lowering drugs? *BMJ*. 1992;304:431-434.
38. Law MR, Thompson SG, Wald NJ. Assessing possible hazards of reducing serum cholesterol. *BMJ*. 1994;308:373-379.
39. Law MR, Wald NJ, Thompson SG. By how much and how quickly does reduction in serum cholesterol concentration lower risk of ischaemic heart disease? *BMJ*. 1994;308:367-373.

# Users' Guides to the Medical Literature

## VII. How to Use a Clinical Decision Analysis

## A. Are the Results of the Study Valid?

W. Scott Richardson, MD; Allan S. Detsky, MD, PhD; for the Evidence-Based Medicine Working Group

### CLINICAL SCENARIO

You are the attending physician on an inpatient service where a 51-year-old man is admitted with congestive heart failure of recent onset. You find he has a dilated cardiomyopathy, the cause of which remains unknown after a thorough evaluation. He is in sinus rhythm. The team's resident asks you whether the patient should be anticoagulated with warfarin, enough to keep his international normalized ratio from 2.0 to 3.0, in order to prevent systemic emboli, even though his echocardiogram does not show left ventricular thrombus. You are not sure about the evidence concerning this issue, so you admit your shared knowledge gap and resolve to search together for the relevant information.

### THE SEARCH

In the hospital's library, the two of you search the MEDLINE system using several search terms, such as "cardiomyopathy, dilated," "cardiomyopathy, congestive," and "heart failure, congestive" crossed with "warfarin," "anticoagulation," and "thromboembolism." Despite several attempts, you retrieve no randomized trials of warfarin used for this purpose. Even after enlisting the help of the librarian, you are unable to locate any clinical trials about this question. You do come across an editorial calling for a clinical trial of your question.[1] You also retrieve two review articles, one that recommends anticoagulation for such patients,[2] and the other that recommends no anticoagulation.[3] The latter review cites a decision analysis on this issue,[4] which you retrieve, hoping to find further guidance for your decision.

From the Department of Medicine, University of Rochester (NY) School of Medicine and Dentistry (Dr Richardson), and the Departments of Health Administration and Medicine, University of Toronto (Ontario), and the Division of General Internal Medicine and Clinical Epidemiology, The Toronto (Ontario) Hospital (Dr Detsky).

A complete list of members (with affiliations) of the Evidence-Based Medicine Working Group appears in the first article of this series (*JAMA*. 1993;270:2093-2095). The following members contributed to this article: Gordon Guyatt, MD, MSc (chair); Deborah Cook, MD, MSc; Hertzel Gerstein, MD, MSc; Robert Hayward, MD, MPH; Anne Holbrook, MD, PharmD, MSc; Roman Jaeschke, MD, MSc; Elizabeth Juniper, MCSP, MSc; Mitchell Levine, MD, MSc; David Naylor, MD, DPhil; Andrew Oxman, MD, MSc, FACPM; David Sackett, MD, MSc; Sean Tunis, MD, MSc; Stephen Walter, PhD; John Williams, Jr, MD, MHS; and Mark Wilson, MD, MPH.

Reprint requests to Room 2C12, McMaster University Health Sciences Centre, 1200 Main St W, Hamilton, Ontario, Canada L8N 3Z5 (Gordon Guyatt, MD, MSc).

### INTRODUCTION

Decision making involves choosing an action after weighing the risks and benefits of the alternatives. While all clinical decisions are made under conditions of uncertainty, the degree of uncertainty decreases when the medical literature includes directly relevant, valid evidence. When the published evidence is scant, or less valid, uncertainty increases.

Decision analysis is the application of explicit, quantitative methods to analyze decisions under conditions of uncertainty. Decision analysis allows clinicians to compare the expected consequences of pursuing different strategies. The process of decision analysis makes fully explicit all of the elements of the decision, so that they are open for debate and modification. While a decision analysis will not solve your clinical problems, it can help you explore the decision.[5-7]

We will use the term "clinical decision analyses" to include studies that analyze decisions faced by clinicians in the course of patient care, such as deciding whether to screen for a condition, choosing a testing strategy, or selecting a treatment. While such analyses can be undertaken to inform a decision for an individual patient ("Should I recommend warfarin to this 51-year-old man with idiopathic dilated cardiomyopathy?"), they are more widely undertaken to help inform a decision about clinical policy[8] ("Should I routinely recommend warfarin to patients in my practice with dilated cardiomyopathy?"). The study retrieved by the search for our scenario is an example of this latter type, while an example of the former is the analysis by Wong et al[9] of whether to recommend cardiac surgery for an elderly woman with aortic stenosis.

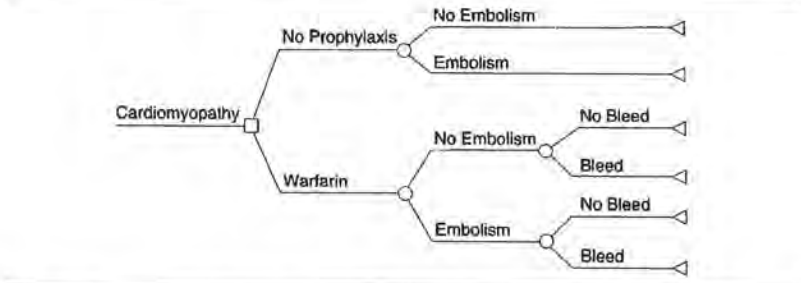
Decision analysis can also be applied to more global questions of health care policy, analyzed from the perspective of society or a national health authority. Examples include analyses of whether or not to screen for prostate cancer[10] and comparing different policies for cholesterol screening and treatment.[11] While decision analyses in health services research share many attributes with clinical analyses,[12] they are sufficiently different that they are beyond the scope of these articles.

In helping you understand decision analysis, we will review some of the "anatomy and physiology" of decision models. This is not meant to be an ar-

**Are the results valid?**
- Were all important strategies and outcomes included?
- Was an explicit and sensible process used to identify, select, and combine the evidence into probabilities?
- Were the utilities obtained in an explicit and sensible way from credible sources?
- Was the potential impact of any uncertainty in the evidence determined?

**What are the results?**
- In the baseline analysis, does one strategy result in a clinically important gain for patients? If not, is the result a toss-up?
- How strong is the evidence used in the analysis?
- Could the uncertainty in the evidence change the result?

**Will the results help me in caring for my patients?**
- Do the probability estimates fit my patients' clinical features?
- Do the utilities reflect how my patients would value the outcomes of the decision?




Structure of a decision tree. Square indicates decision node; circles, chance nodes; triangles, outcome nodes; and lines, strategy pathways. Numbers (when present) by lines indicate probabilities, and by triangles, utilities.

ticle on how to perform decision analysis; if you wish to read about that, you should look elsewhere.[13,14]

## FRAMEWORK FOR THE USERS' GUIDES

We will approach articles on clinical decision analysis using the same framework introduced in earlier articles in this series, as follows:

### Are the Results Valid?

This question addresses whether the strategy recommended by the analysis is truly likely to be the better one for patients. Just as with other types of studies, the validity of a decision analysis is largely determined by the strength of the methods used.

### What Are the Results?

The users' guides under this second question consider the size of the expected net benefit from the recommended strategy and our confidence in this estimate of net benefit.

### Will the Results Help Me in Caring for My Patients?

If the decision analysis yields valid and important results, you should examine whether these results can be generalized to the patients in your practice.

The Table summarizes the specific guides you should use when addressing these three questions. We will explore the guides by applying them to the study we found in our search. This article will deal with the validity guides, while the next in the series will address the results and applicability.

## ARE THE RESULTS VALID?

### Were All Important Strategies and Outcomes Included?

At issue here is how well the structure of the model fits the clinical decision you face. Most clinical decision analyses are built as decision trees, and the articles will usually include one or more diagrams showing the structure of the decision tree used for the analysis. Reviewing these diagrams will help you understand the model. You must then judge whether the model fits the clinical problem well enough to be valid.

The Figure shows a diagram of a much simplified version of the decision tree for the anticoagulation problem. The clinician has two options for patients with cardiomyopathy, either to offer no prophylaxis or to prescribe warfarin. Either way, patients may or may not develop embolic events. Prophylaxis lowers the chance of embolism but can cause bleeding in some patients. As seen in the Figure, decision trees are displayed graphically, oriented from left to right, with the decision to be analyzed on the left, the compared strategies in the center, and the clinical outcomes on the right. The decision is diagrammed by a square, termed a "decision node." The lines emanating from the decision node represent the clinical strategies being compared. Chance events are diagrammed with circles, called "chance nodes," and outcome states are shown as triangles or as rectangles.

To explore more fully how the model's structure affects its validity, we will highlight two aspects here.

**Were All of the Realistic Clinical Strategies Compared?**—In a decision analysis, a strategy is defined as a sequence of actions and decisions that are contingent on each other. For instance, the strategy of anticoagulant therapy for a patient includes not only the prescription and the monitoring, but also the adjustment of the warfarin dose for changes in prothrombin time. The authors should specify which decision strategies are being compared (at least two, otherwise there's no decision). Further, the clinical strategies included should be described in enough detail to recognize them as separate and realistic choices. You should satisfy yourself that the clinical strategies you consider important are included in the analysis.

For example, in a decision analysis of the management of suspected herpes encephalitis, the authors included the three strategies available to clinicians then: brain biopsy, empirical vidarabine, or neither.[15] At that time, this model represented the clinical decision well. Since then, however, acyclovir has become available and has been widely used for this disorder. Because the original model did not include an acyclovir strategy, it would no longer accurately portray the decision.

In the anticoagulation example, the analysts studied two clinical strategies, warfarin and no warfarin. This fits quite well the clinical decision you face in the scenario. Note that the decision model does not include a third strategy of using aspirin instead of warfarin. If, when considering the treatment options for this patient, you would seriously consider the use of aspirin instead of warfarin, then you would judge this model as incomplete.

**Were All Clinically Relevant Outcomes Considered?**—To be useful to clinicians and patients, the decision model should include the outcomes of the disease that matter to patients. Generally speaking, these include not only the quantity of life but also its quality, in measures of disease and disability. Obviously, the specific disorder in question determines which outcomes are clinically relevant. For an analysis of an acute, life-threatening condition, life expectancy might be appropriate as the main outcome measure. But in an analysis of diagnostic strategies for a nonfatal disorder, more relevant outcomes would be discomfort from testing or days of disability avoided. By examining the outcomes used in the analysis, you can discover the viewpoint from which the analyst built the decision model. Clinical decision analyses should be built from the perspective of the patient, that is, should include all the clinical benefits

and risks of importance to patients (they can include other considerations as well).

Also, by comparing the outcomes between strategies, you can discover the trade-offs built into the model. Most clinical dilemmas are dilemmas because they include trade-offs between competing benefits and competing risks. For instance, when deciding how best to manage small abdominal aortic aneurysms, one must weigh reducing the risk of aneurysm rupture against the chance of unnecessary surgery in patients who would have died from other causes before rupture.[16] For a decision analysis to be worth doing, ie, for the clinical decision to be difficult enough, the choice of strategies should be balanced on one or more of such trade-offs. You should satisfy yourself that these important trade-offs are represented well in the model's structure.

For the anticoagulation example, the authors' decision model includes all of the clinical events of interest to patients (stroke, other emboli, hemorrhage, and the like). The outcomes are measured as "quality-adjusted life expectancy," a scale that combines information about both the quantity and the quality of life. This metric fits your clinical decision well, for you can expect that warfarin might affect both the quantity and quality of life. By reviewing the tree diagram, you can see that the authors have included the principal trade-off in the decision: the warfarin strategy offers the benefits of preventing systemic arterial embolism causing stroke and preventing pulmonary embolism, while it could cause the harm of bleeding.

### Was an Explicit and Sensible Process Used to Identify, Select, and Combine the Evidence Into Probabilities?

To assemble the large amount of information necessary for a decision analysis, the analyst searches the published literature and interviews experts and patients. Just as with other integrative studies like overviews,[17] authors of clinical decision analyses should search and select the literature in an explicit and unbiased way, and then appraise the validity, effect size, and homogeneity of the studies in a reproducible fashion. Ideally, they would judge study quality by applying criteria akin to those in the other articles in this series, whether for primary studies of therapy,[18,19] diagnosis,[20,21] harm,[22] prognosis,[23] or for other integrative studies, such as overviews.[17] In other words, the authors should perform as comprehensive a literature review as is required for a meta-analysis.

Once gathered, the information must be transformed into quantitative estimates of the likelihood of events, or probabilities. The scale for probability estimates ranges from 0 (impossible) to 1.0 (absolutely certain). Probabilities must be assigned to each branch emanating from a chance node, and for each chance node, the sum of probabilities must add to 1.0.

For example, looking at the Figure, note that the no-anticoagulation strategy (the upper branch coming from the decision node) has one chance node, at which two possible events could occur, either an embolism or no embolism (labeled "no embolism"). To assign a probability to these two branches from the chance node, the analyst tracks down all relevant evidence about the rates of systemic emboli in patients with cardiomyopathy. If the best estimate of the rate were found to be 5%, then the analyst would assign 0.05 to the embolism branch and 0.95 to the no-embolism branch.

Usually, rates from clinical studies can be directly translated into probabilities, as in this example. In other instances, the data must be transformed first, such as when analysts must adjust 5-year survival data to fit an analysis concerned with only the first 3 years. Analysts should report which data were used and how the data were transformed.

In the anticoagulation example, the authors describe vigorous efforts to obtain the correct values for probabilities from the published literature and from experts, although they don't provide the search terms they used. The authors do highlight the limited data available and the data's methodological limits. Also, they tabulate the evidence they use and mention the transformations needed for the model.

### Were the Utilities Obtained in an Explicit and Sensible Way From Credible Sources?

Utilities represent quantitative measurements of the value to the decision maker of the various outcomes of the decision. Several methods are available to measure these values directly,[5,7,24,25] and which method is best remains controversial. Different methods use different scales; a commonly used utility scale ranges from 0 (worst outcome, usually death) to 1.0 (excellent health). Whatever the measurement method used, the authors should report the source of the ratings. In a decision analysis built for an individual patient, the most (and probably only) credible ratings are those measured directly from that patient. For analyses built to inform clinical policy, credible ratings could come from three sources: (1) direct measurements from a large group of patients with the disorder in question and to whom results of the decision analysis could be applied; (2) from published studies of quality-of-life ratings by such patients, as was done in a recent analysis of strategies for chronic atrial fibrillation[26]; or (3) from an equally large group of people representing the general public. Whoever provides the rating must understand the outcomes they are asked to rate; the more the raters know about the condition, the more credible are their utility ratings.

The authors of the anticoagulation example obtained values from several internists familiar with the clinical disorder and with the treatments. While physician raters were undoubtedly familiar with the outcomes of systemic emboli and major hemorrhage, only a small number of physicians made ratings, and their values may not represent those of either patients or the general public.

### Was the Potential Impact of Any Uncertainty in the Evidence Determined?

Much of the uncertainty in clinical decision making arises from the lack of valid evidence in the literature. This lack of data hampers both clinical decision making and formal decision analysis. Even when it is present, published evidence is often imprecise, with wide confidence intervals around estimates for important variables. For instance, in a decision analysis concerning the management of polymyalgia rheumatica, the analysts searched the literature for the test sensitivity of temporal artery biopsy for giant cell arteritis.[27] The reported test sensitivity ranged from about 60% to 100%. In the decision analysis, these analysts set the baseline value equal to 83%, but repeated the analysis for values between 60% and 100%.

Decision analysts use this systematic exploration of the uncertainty in the data, known as "sensitivity analysis," to see what effect varying estimates for risks, benefits, and values have on the expected clinical outcomes, and therefore on the choice of clinical strategies. Sensitivity analysis asks the question: is the conclusion generated by the decision analysis affected by the uncertainties in our estimates of the likelihood or value of the outcomes? Estimates can be varied one at a time, termed "one-way" sensitivity analyses, or two or three at a time, known as "multi-way" sensitivity analyses. You should look for a table listing which variables were included in the sensitivity analyses, what range of values were used for each variable, and which variables, if any, altered the choice of strategies. Sat-

isfy yourself that all of the clinically important variables were examined.

Generally, all of the probability estimates should be tested using sensitivity analyses. The range over which they should be tested will depend on the source of the data. If the estimates come from large, high-quality randomized trials with narrow confidence limits, the range of estimates tested can be narrow. The less valid the methods, or the less precise the estimates, the wider the range that must be included in the sensitivity analyses.

Utility values should also be tested with sensitivity analyses, with the range of values again determined by the source of the data. If large numbers of patients or knowledgeable and representative members of the general public gave very similar ratings to the outcome states, a narrow range of utility values can be used in the sensitivity analyses. If the ratings came from a small group of raters, or if individuals varied widely in their values, then investigators should use a wider range of utility values in the sensitivity analyses.

In the anticoagulation example, the authors responded to the poor quality of their evidence by varying all of the important variables over wide ranges. They report the results from several, although not all, of these sensitivity analyses, including the effect of higher bleeding risk while taking warfarin.

In the next article on clinical decision analysis, we will show you how to determine what the results are and how to use them in your practice.

Dr Detsky is supported in part by a National Health Research Scholar Award from Health and Welfare Canada.

**References**


1. Falk RH. A plea for a clinical trial of anticoagulation in dilated cardiomyopathy. *Am J Cardiol.* 1990;65:914-915.
2. Dec GW, Fuster V. Idiopathic dilated cardiomyopathy. *N Engl J Med.* 1994;331:1564-1575.
3. Baker DW, Wright RF. Management of heart failure, IV: anticoagulation for patients with heart failure due to left ventricular systolic dysfunction. *JAMA.* 1994;272:1614-1618.
4. Tsevat J, Eckman MH, McNutt RA, Pauker SG. Warfarin for dilated cardiomyopathy: a bloody tough pill to swallow? *Med Decis Making.* 1989;9:162-169.
5. Keeney RL. Decision analysis: an overview. *Operations Res.* 1982;30:803-838.
6. Eckman MH, Levine HJ, Pauker SG. Decision analytic and cost-effectiveness issues concerning anticoagulant prophylaxis in heart disease. *Chest.* 1992;102(suppl 4):538S-549S.
7. Kassirer JP, Moskowitz AJ, Lau J, Pauker SG. Decision analysis: a progress report. *Ann Intern Med.* 1987;106:275-291.
8. Eddy DM. Designing a practice policy: standards, guidelines, and options. *JAMA.* 1990;263:3077, 3081, 3084.
9. Wong JB, Salem DN, Pauker SG. You're never too old. *N Engl J Med.* 1993;328:971-975.
10. Krahn MD, Mahoney JE, Eckman MH, Trachtenberg J, Pauker SG, Detsky AS. Screening for prostate cancer: a decision analytic view. *JAMA.* 1994;272:773-780.
11. Krahn MD, Naylor CD, Basinski AS, et al. Comparison of an aggressive (U.S.) and a less aggressive (Canadian) policy for cholesterol screening and treatment. *Ann Intern Med.* 1991;115:248-255.
12. Goel V. Decision analysis: applications and limitations. *Can Med Assoc J.* 1992;147:413-417.
13. Weinstein MC, Fineberg HV, et al. *Clinical Decision Analysis.* Philadelphia, Pa: WB Saunders; 1980.
14. Sox HC, Blatt MA, Higgins MC, Marton KI. *Medical Decision Making.* Boston, Mass: Butterworth-Heinemann; 1988.
15. Barza M, Pauker SG. The decision to biopsy, treat, or wait in suspected herpes encephalitis. *Ann Intern Med.* 1980;92:641-649.
16. Katz DA, Littenberg B, Cronenwett JL. Management of small abdominal aortic aneurysms: early surgery vs watchful waiting. *JAMA.* 1992;268:2678-2686.
17. Oxman AD, Cook DJ, Guyatt GH, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VI: how to use an overview. *JAMA.* 1994;272;1367-1371.
18. Guyatt GH, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, II: how to use an article about therapy or prevention, A: are the results of the study valid? *JAMA.* 1993;270:2598-2601.
19. Guyatt GH, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, II: how to use an article about therapy or prevention, B: what were the results and will they help me in caring for my patients? *JAMA.* 1994;271:59-63.
20. Jaeschke R, Guyatt GH, Sackett DL, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, III: how to use an article about a diagnostic test, A: are the results of the study valid? *JAMA.* 1994;271:389-391.
21. Jaeschke R, Guyatt GH, Sackett DL, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, III: how to use an article about a diagnostic test, B: what are the results and will they help me in caring for my patients? *JAMA.* 1994;271:703-707.
22. Levine MS, Walter SS, Lee HN, Haines T, Holbrook A, Moyer V, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, IV: how to use an article about harm. *JAMA.* 1994;271:1615-1619.
23. Laupacis A, Wells G, Richardson WS, Tugwell P, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, V: how to use an article about prognosis. *JAMA.* 1994;272:234-237.
24. Llewellyn-Thomas H, Sutherland HJ, Tibshirani R, et al. The measurement of patient values in medicine. *Med Decis Making.* 1982;2:449-462.
25. Dolan JG, Isselhardt BJ, Cappuccio JD. The analytic hierarchy process in medical decision making: a tutorial. *Med Decis Making.* 1989;9:40-50.
26. Disch DL, Greenberg ML, Holzberger PT, et al. Managing chronic atrial fibrillation: a Markov decision analysis comparing warfarin, quinidine, and low-dose amiodarone. *Ann Intern Med.* 1994;120:443-457.
27. Buchbinder R, Detsky AS. Management of suspected giant cell arteritis: a decision analysis. *J Rheumatol.* 1992;19:1220-1228.

# Users' Guides to the Medical Literature

## VII. How to Use a Clinical Decision Analysis

## B. What Are the Results and Will They Help Me in Caring for My Patients?

W. Scott Richardson, MD, Allan S. Detsky, MD, PhD, for the Evidence-Based Medicine Working Group

YOU RECALL from the first of our two articles concerning clinical decision analysis[1] that your patient is a middle-aged man with heart failure from an idiopathic dilated cardiomyopathy. You are trying to decide whether to recommend anticoagulation with warfarin to prevent systemic or pulmonary thromboembolism. Your literature search showed that no randomized clinical trials of warfarin for this use have been published. The search did discover a clinical decision analysis,[2] and in the first article, we showed you how to evaluate its validity. In this article, we will show you how to interpret the results and generalizability of a clinical decision analysis (Table).

From the Department of Medicine, University of Rochester (NY) School of Medicine and Dentistry (Dr Richardson), and the Departments of Health Administration and Medicine, University of Toronto (Ontario), and the Division of General Internal Medicine and Clinical Epidemiology, The Toronto (Ontario) Hospital (Dr Detsky).

A complete list of members (with affiliations) of the Evidence-Based Medicine Working Group appears in the first article of this series (*JAMA*. 1993;270:2093-2095). The following members contributed to this article: Gordon Guyatt, MD, MSc (chair), Deborah Cook, MD, MSc; Hertzel Gerstein, MD, MSc; Robert Hayward, MD, MPH; Anne Holbrook, MD, PharmD; Roman Jaeschke, MD, MSc; Elizabeth Juniper, MCSP, MSc; Mitchell Levine, MD, MSc; David Naylor, MD, DPhil; Andrew Oxman, MD, MSc; David Sackett, MD, MSc; Sean Tunis, MD, MSc; Stephen Walter, PhD; John Williams, Jr, MD, MHS; and Mark Wilson, MD, MPH.

Reprint requests to Room 2C12, McMaster University Health Sciences Centre, 1200 Main St W, Hamilton, Ontario, Canada L8N 3Z5 (Gordon Guyatt, MD, MSc).

As shown in the Figure, decision trees are displayed graphically, oriented from left to right, with the decision to be analyzed on the left, the compared strategies in the center, and the clinical outcomes on the right. The square box, termed a "decision node," represents the decision to be made, and the lines emanating from this decision node represent the clinical strategies being compared. Circles, or "chance nodes," represent chance events and outcome states are shown as triangles on the far right. Numbers beside the strategies (if they were present) would be "probabilities," or the likelihood of events, while the numbers by the outcome states would be "utilities," or the value of these events.[3,4]

### In the Baseline Analysis, Does One Strategy Result in a Clinically Important Gain for Patients? If Not, Is the Result a Toss-up?

For a clinical decision analysis that compares two clinical strategies, there are three possible results: the first strategy is better than the second, the second strategy is better than the first, or both strategies are equally good (or equally bad), a result known as a "toss-up" or a "close call."[5] For instance, in an analysis of the management of solitary pulmonary nodules, the analysts found the choice of strategies to be a close call in terms of expected gains in life expectancy.[6] The larger the number of strategies compared in an analysis, the larger the number of possible results, but always with the same idea: any one strategy can "win" or two or more strategies could "tie." The terms "baseline" or "base case" refer to the set of numbers for probability that the analyst believes are closest to the actual state of affairs.

One chooses between strategies in a decision tree by comparing the overall benefits expected from pursuing each strategy, termed its "expected utility," and then selecting the strategy with the highest value of expected utility. Some controversy remains as to when exceptions to this rule are legitimate or desirable.[7] To calculate expected utility, one starts at the rightmost branches of the tree, multiplies the probability for each by its utility, and sums these products for each chance node. One repeats this calculation moving leftward, a process known as "folding back," until one has calculated the expected utility value for each strategy.

For example, consider the topmost chance node in the Figure, with its two branches. Imagine that the "no-embolism" and "embolism" branches have probabilities of 0.95 and 0.05 and utilities of 1.0 and .9, respectively. The expected utility for this chance node would be the sum of the

Users' Guides to the Medical Literature section editor: Drummond Rennie, MD, Deputy Editor (West), *JAMA*.

**Are the results valid?**
- Were all important strategies and outcomes included?
- Was an explicit and sensible process used to identify, select, and combine the evidence into probabilities?
- Were the utilities obtained in an explicit and sensible way from credible sources?
- Was the potential impact of any uncertainty in the evidence determined?

**What are the results?**
- In the baseline analysis, does one strategy result in a clinically important gain for patients?
  - If not, is the result a toss-up?
- How strong is the evidence used in the analysis?
- Could the uncertainty in the evidence change the result?

**Will the results help me in caring for my patients?**
- Do the probability estimates fit my patients' clinical features?
- Do the utilities reflect how my patients would value the outcomes of the decision?

product of each of the probabilities times the utilities, in this case $(0.95 \times 1.0) + (0.05 \times 0.9)$, which equals 0.995.

The decision analyst chooses the scale on which these expected utilities are measured to fit the clinical problem. For instance, in an analysis of strategies that could reduce death, the analyst might choose to measure utility as the number of lives saved or the average gain in remaining life expectancy, both measures of the quantity of life. Other utility scales can be used to report on the quality of life. Both quantity and quality can be combined into a single measure, such as quality-adjusted life years[8] or healthy-years equivalence.[9] For instance, suppose one strategy in a decision analysis yielded an average remaining life expectancy of 5 years, but that all five years were lived in a state of health rated by patients to have a utility value of 0.8. The quality-adjusted life expectancy would be $5 \times 0.8$ or 4 years.

Now that you understand where the results of the decision analysis come from, you must decide if any difference between strategies is clinically important. In making this judgment, consider that the differences presented will be average differences rather than differences that you can expect for every patient. Some patients will gain considerably more, while others will gain considerably less. This is no different than interpreting average differences between groups in randomized trials. You may not, however, be familiar with differences in life expectancy, the output of many decision analyses. Keep in mind that a gain in life expectancy does not occur just at the end of a person's life—it may occur at the beginning or be spread over the course of time.[10]

How large must a gain in remaining life expectancy be to be important? Probably smaller than you might think, although the answer to this question depends on judgments about several variables, and this controversial area has not yet been fully addressed by empirical research. In some recent studies, decision analysts have "translated" the results of clinical trials into life expectancy gains, for various widely accepted clinical interventions.[10,11] These studies suggest that a gain in life expectancy or quality-adjusted life expectancy of 2 or more months ought to be considered an important gain, while a gain of a few days would represent a toss-up.

In the anticoagulation for dilated cardiomyopathy example, the decision analysis finds warfarin to be the preferred strategy for all patients 35 to 75 years of age. The average gain in quality-adjusted life expectancy for 55-year-old patients (similar to your 51-year-old patient) is 115 days, or almost 3 months. From the above, you can see that this gain in life expectancy is probably important. Since the analysts explicitly considered both the reduction of emboli and the risk of bleeding, this 3-month gain in life expectancy represents the *net* clinical benefit you could expect from recommending anticoagulation to your patient.

## How Strong Is the Evidence Used in the Analysis?

The probabilities used in clinical decision analyses are estimates, taken mostly from the published literature, and while they may represent the best available evidence, they are nonetheless subject to potential error. The best defense against such error is for the analysts to base probability estimates on studies of high methodological quality, after a thorough and unbiased search for all relevant studies. The analysts should explain how they judged the quality of these primary studies. One way to do this would be to judge study quality by applying criteria akin to those in the other articles in this series, whether for primary studies of therapy,[12,13] diagnosis,[14,15] harm,[16] prognosis,[17] or for integrative studies, such as overviews.[18]

As with other integrative studies, the overall strength of the result of a clinical decision analysis depends on the strength of inference possible from the primary studies. Ideally, every probability estimate at every node in the tree is supported by precise estimates from primary and integrative studies of high methodological quality, but such idealized analyses are rare. Good decision analyses can still be performed with some imprecise or ambiguous data, as long as most of the data are of good quality and the analysts explain any limitations and plan their sensitivity analyses accordingly. The fewer the probabilities that can be precisely estimated from high quality primary studies, ie, the weaker the evidence used in the analysis, the weaker the overall inference one can make from the results.

In the anticoagulation example, the authors describe vigorous efforts to obtain the correct values for probabilities from the published literature and from experts. They highlight the limited methodological quality of the primary literature and acknowledge the weakened inference. In particular, there are no randomized trials to tell you whether patients with cardiomyopathy will live longer or have fewer morbid events if given anticoagulants.

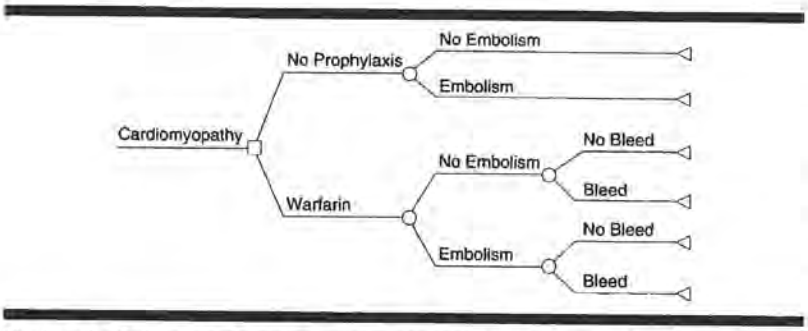
## Could the Uncertainty in the Evidence Change the Result?

For any clinical variable such as the probability of bleeding, or the value that patients place on avoiding a stroke, the decision analyst can calculate the value, or "threshold," above which the results favor one strategy, and below which the results favor another strategy. For multiway sensitivity analyses, the analyst can show two-dimensional graphs of the variables, with the thresholds displayed as a line (two-way analyses) or a series of lines (three-way analyses) separating zones of strategy preference. While this may be daunting at first, these tables and graphs provide the most clinically useful information from a decision analysis.

If the result of the analysis (one strategy is preferred or a toss-up is found) would change by choosing different values for one of the variables, the result is said to be "sensitive" to that variable. On the other hand, if changing the variable throughout its plausible range of values doesn't change the result, the analysis result is said to be "robust" to the sensitivity analysis. As you might guess, the more robust the result is, the more confident you can be that the recommended strategy should indeed be preferred. If the result was a toss-up and that indifference proves robust to sensitivity analyses, you can be confident that the strategies are equivalent.

The analysts of the anticoagulation example found the preference for the warfarin strategy to be robust to the sensitivity analyses they completed, with two exceptions (we will return to one of these, the bleeding risk for patients taking warfarin).

For the other exception, the analysts assumed in the base case that patients' quality of life was not impaired by the inconvenience and anxiety associated with taking warfarin (ie, a utility value of 1.0 on a 0 to 1.0 scale). When testing




Structure of a decision tree. Square indicates decision node; circles, chance nodes; triangles, outcome nodes; and lines, strategy pathways. Numbers (when present) by lines indicate probabilities, and by triangles, utilities.

this assumption, by adjusting downward the utility rating for quality of life while taking warfarin, the analysts discovered that the choice of strategies would change substantially. For 55-year-old patients, the threshold utility value was 0.92. In other words, if patients rated their quality of remaining life while taking warfarin as 0.93 or greater, then anticoagulation would be preferred. For a utility rating of exactly 0.92, the two strategies would be equally preferred, while for utility ratings below 0.92, no anticoagulation would be preferred.

To put this result in perspective, recall that utility represents the value to the patient of remaining expected life, and that a rating of 0.92 is 8% less than normal. In other words, a utility threshold of 0.92 means that your patient feels he would be willing to sacrifice 8% of his remaining life to avoid taking warfarin. On a time scale, this means that a year taking warfarin would have to be worth only approximately 11 months of life not taking warfarin, in order for him to choose not to take it.

## WILL THE RESULTS HELP ME IN CARING FOR MY PATIENTS?

### Do the Probability Estimates Fit My Patients' Clinical Features?

This first issue of applicability concerns whether the clinical characteristics of patients for whom the analysis was intended are similar to your patients. For a decision analysis built for an individual patient, look for the description of that patient's condition; if the patient is well described, you should be readily able to judge how closely your patient resembles her or him. An article reporting a decision analysis built for a group of patients should have an analogous portion of the text, detailing the clinical characteristics of patients to whom the results of the analysis are to be applied. You should satisfy yourself that your patient would be included in this group.

You could be confident that the probabilities fit your patients if the estimates were taken from one or more rigorous clinical studies in which patient samples included patients similar to yours. If the authors don't describe the samples, you could track down the references and review the inclusion and exclusion criteria to see whether your patient would fit.

If the analysis was intended for patients different from yours, review the results of the sensitivity analyses. The clinical variables used for these analyses should be detailed enough for you to locate where your patient would fit, and thus what net benefit your patient might expect from the clinical strategies. If you still can't tell, ask yourself whether the clinical characteristics of the intended patients are so different from yours that you should discard the results. If not, you can proceed, with some caution, to use them.

In the anticoagulation example, most of the probabilities fit your dilated cardiomyopathy patient, including the rates of systemic and pulmonary emboli and the estimated mortality. The baseline average annual risk of major hemorrhage on warfarin was estimated to be 4.5%. If you worried that your patient's risk of bleeding while taking warfarin could be higher than average, you should examine the sensitivity analyses for this variable. These sensitivity analyses show that anticoagulation with warfarin remains the preferred strategy until the annual bleeding risk reaches 15%, more than triple the baseline estimate. Above this value, no anticoagulation became the preferred strategy.

When a clinical decision analysis shows that the preferred strategy is sensitive to a given variable, you will need to gauge where your patient fits on the scale of that variable. Thus, when deciding how to use the results of the anticoagulation decision analysis for your particular patient, you will need to estimate his annual risk of bleeding while undergoing warfarin therapy. While a full discussion of estimating the bleeding risk is beyond the scope of this article, we offer a few suggestions.

First, look in the text for the authors' description of their systematic review of the literature. Ideally, they will have found one or more original articles or systematic reviews of high methodological quality from which they obtained their baseline estimate, and from which you could obtain an individualized estimate for your patient. Alternatively, you could do your own search for this information, using the tactics introduced in the first article in this series.[19]

If you did so you would find a systematic review of this topic,[20] wherein the authors cite the average annual frequencies of fatal and major hemorrhage in patients taking warfarin as 0.6% and 3.0%, respectively. You might also find a study of warfarin use in atrial fibrillation,[21] wherein the incidence of major or fatal bleeding was 2.5%. If these numbers are close to the truth, then by using somewhat higher figures in the anticoagulation decision analysis, the analysts would have overestimated the risk of harm and might have obscured a net benefit. Despite this, the warfarin strategy still resulted in a clinically important expected gain in life expectancy, suggesting that the true net benefit might be somewhat larger than reported. Note also that these published estimates are substantially lower than the 15% threshold value for annual bleeding risk, above which the no-warfarin strategy would be preferred.

Your search would also turn up a retrospective analysis of thromboembolism rates in two randomized trials of other treatments (not anticoagulants) for heart failure.[22] During the approximately 2.5 years average follow-up, the trials showed thromboembolic events occurred in 4.7% and 5.2% of patients. After transformation to comparable event rates, these results may be a little over half of the values used in the anticoagulation decision analysis. By using somewhat higher estimates, the analysts could have overestimated the benefit of warfarin.

### Do the Utilities Reflect How My Patients Would Value the Outcomes of the Decision?

Since the utility ratings for the value of outcomes has a strong influence on the choice of strategies, you must consider whether your patient's values are similar to those used in the decision analysis. In a decision analysis built for an individual patient, the utilities are usually measured directly from that patient, so while those values should be quite believable for that patient, they may not necessarily fit your patient. Alternatively, utilities measured from a

large group of patients or members of the general public would probably include a set of values similar to those of your patient, but the range of values might be so broad that you are left uncertain as to which values to use. If you encounter such difficulties, you should examine the one-way and multiway sensitivity analyses that use a wide range of utility estimates to see how your patient's values will affect the final decision.

If you were to ask your patient to rate the outcome states using the rating instrument in the article, you would know exactly what utility values to use. However, most clinicians won't have the time or inclination to do this. Fortunately, you can still make some judgment about this question by asking your patient about values in nonquantitative terms. For instance, one patient may be extremely averse to regular monitoring, while another may not mind. Disabling stroke might devastate one patient, whereas another might be more resilient.

As mentioned above in the anticoagulation example, the utility rating for life while taking warfarin had a substantial influence on the preference of strategies. The authors highlight the importance of this variable and urge that investigators examine patients' reactions to taking warfarin and undergoing monitoring, so that subsequent recommendations about anticoagulation can be better informed.

## RESOLUTION OF THE SCENARIO

Without a randomized trial of anticoagulation in patients with dilated cardiomyopathy in sinus rhythm, your overall confidence in a decision to anticoagulate your patient will be limited. In the absence of trial data, experts have recommended that the decision to use warfarin in this setting be made on an individual basis.[23-25] How are you to individualize the treatment decision for your middle-aged man with dilated cardiomyopathy? The anticoagulation decision analysis suggests that if he has a low or moderate bleeding risk and a ready acceptance of anticoagulation monitoring, he is likely to be better off taking warfarin. Thus, the decision analysis identifies the few clinical variables on which the decision depends, and estimates the size and likelihood of net clinical benefit you could expect from the alternative courses of action. While the better therapy may still be unproved, you should now be much more informed about the choice and better prepared to decide with the patient what is to be done.

Dr Detsky is supported in part by a National Health Research Scholar Award from Health and Welfare Canada.

### References


1. Richardson WS, Detsky AS, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VII: how to use an article on clinical decision analysis, A: are the results of the study valid? *JAMA.* 1995;273:1292-1295.
2. Tsevat J, Eckman MH, McNutt RA, Pauker SG. Warfarin for dilated cardiomyopathy: a bloody tough pill to swallow? *Med Decis Making.* 1989;9:162-169.
3. Weinstein MC, Fineberg HV. *Clinical Decision Analysis.* Philadelphia, Pa: WB Saunders; 1980.
4. Sox HC, Blatt MA, Higgins, MC, Marton KI. *Medical Decision Making.* Boston, Mass: Butterworth-Heinemann; 1988.
5. Kassirer JP, Pauker SG. The toss-up. *N Engl J Med.* 1981;305:1467-1469.
6. Cummings SR, Lillington GA, Richard RJ. Managing solitary pulmonary nodules: the choice of strategy is a 'close call.' *Am Rev Respir Dis.* 1986;134:453-460.
7. Deber RB, Goel V. Using explicit decision rules to manage issues of justice, risk and ethics in decision analysis: when is it not rational to maximize expected utility? *Med Decis Making.* 1990;10:181-194.
8. Torrance GW, Feeny D. Utilities and quality-adjusted life years. *Int J Technol Assess Health Care.* 1989;5:559-579.
9. Mehrez A, Gafni A. Quality-adjusted life years, utility theory and healthy-years equivalence. *Med Decis Making.* 1989;9:142-149.
10. Naimark DM, Naglie G, Detsky AS. The meaning of life expectancy: what is a clinically significant gain? *J Gen Intern Med.* 1994;9:702-707.
11. Tsevat J, Weinstein MC, Williams LW, et al. Expected gains in life expectancy for various coronary heart disease risk factor modifications. *Circulation.* 1991;83:1194-1201.
12. Guyatt GH, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, II: how to use an article about therapy or prevention, A: are the results of the study valid? *JAMA.* 1993;270:2598-2601.
13. Guyatt GH, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, II: how to use an article about therapy or prevention, B: what were the results and will they help me in caring for my patients? *JAMA.* 1994;271:59-63.
14. Jaeschke R, Guyatt GH, Sackett DL, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, III: how to use an article about a diagnostic test, A: are the results of the study valid? *JAMA.* 1994;271:389-391.
15. Jaeschke R, Guyatt GH, Sackett DL, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, III: how to use an article about a diagnostic test, B: what are the results and will they help me in caring for my patients? *JAMA.* 1994;271:703-707.
16. Levine MS, Walter SS, Lee HN, Haines T, Holbrook A, Moyer V, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, IV: how to use an article about harm. *JAMA.* 1994;271:1615-1619.
17. Laupacis A, Wells G, Richardson WS, Tugwell P, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, V: how to use an article about prognosis. *JAMA.* 1994;272:234-237.
18. Oxman AD, Cook DJ, Guyatt GH, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VI: how to use an overview. *JAMA.* 1994;272:1367-1371.
19. Oxman AD, Sackett DL, Guyatt GH, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, I: how to get started. *JAMA.* 1993;270:2093-2095.
20. Landefeld CS, Beyth RJ. Anticoagulant-related bleeding: clinical epidemiology, prediction and prevention. *Am J Med.* 1993;95:315-328.
21. Connolly SJ, Laupacis A, Gent M, Roberts RS, Cairns JA, Joyner C, for the CAFA Study Coinvestigators. Canadian atrial fibrillation anticoagulation (CAFA) study. *J Am Coll Cardiol.* 1991;18:349-355.
22. Dunkman WB, Johnson GR, Carson PE, et al. Incidence of thromboembolic events in congestive heart failure. *Circulation.* 1993;87[suppl VI]:VI94-VI101.
23. Dec GW, Fuster V. Idiopathic dilated cardiomyopathy. *N Engl J Med.* 1994;331:1564-1575.
24. Baker DW, Wright RF. Management of heart failure, IV: anticoagulation for patients with heart failure due to left ventricular systolic dysfunction. *JAMA.* 1994;272:1614-1618.
25. Kubo SH, Cohn JN. Approach to treatment of the patient with heart failure in 1994. *Adv Intern Med.* 1994;39:485-515.

# Users' Guides to the Medical Literature

## VIII. How to Use Clinical Practice Guidelines

## A. Are the Recommendations Valid?

Robert S. A. Hayward, MD, MPH; Mark C. Wilson, MD, MPH; Sean R. Tunis, MD, MSc; Eric B. Bass, MD, MPH; Gordon Guyatt, MD, MSc; for the Evidence-Based Medicine Working Group

### CLINICAL SCENARIO

You are relieved to find that the last patient in your busy primary care clinic is a previously well 48-year-old woman with acute dysuria. There has been no polydipsia, fever, or hematuria; the physical examination reveals suprapubic tenderness; and urinalysis shows pyuria but no casts. You arrange cultures and antibiotic treatment for a lower urinary tract infection. On her way out the door, your patient observes that her friend has just started taking "female hormones," and she wonders whether she should too. Her menstrual periods stopped 6 months ago and she has never had cervical, ovarian, uterine, breast, or cardiovascular problems, but her mother had a mastectomy at age 57 for postmenopausal breast cancer. You give the same general advice you have offered similar patients in the past, but suggest that the matter be discussed at greater length when she returns after completing the antibiotic treatment. Later, as you lament doorknob consults, you are irritated when a colleague asserts that your primary advice about prophylactic hormone replacement therapy (HRT) was wrong and that you should have recommended exactly the opposite. You resolve to revisit this disagreement, armed with the best evidence.

### THE SEARCH

You begin by using Grateful Med to look for a recent overview because many articles about prophylactic HRT have appeared recently, your time is short, and your patient would want to know about all significant benefits and harms associated with HRT. On the first subject line of the Grateful Med search, you select "estrogen replacement therapy" by marking this as a major concept in the list of Medical Subject Headings (MeSH) that Grateful Med associates with the term "estrogen." After limiting your search to English-language reviews (Publication Type="review"), you still have 131 articles to consider. A quick scan of the first 25 titles reveals diverse topics, including the effect of HRT on lipid profiles, bone density, fracture rates, and the incidence of endometrial, cervical, and breast cancer. Knowing that "practice guideline" is among the publication types listed by Grateful Med, you reason that clinical practice guidelines might address multiple HRT-related outcomes at one time, and thus provide you with the most efficient access to the best summary or summaries of the available data. A repeat search with the new publication type yields five citations. Two of these are "technical bulletins" of the American College of Obstetricians and Gynecologists,[1,2] one is written for surgeons,[3] one is a recent guideline from the American College of Physicians (ACP),[4] and the last is a commentary on the ACP guideline.[5] Observing that the ACP guideline is published together with a systematic overview of the evidence supporting its recommendations,[6] you begin your review of issues in HRT decision making with the ACP guideline.

### INTRODUCTION

Clinicians serve patients by addressing each individual's health care needs. This includes recognizing important health problems, considering sensible options for managing each problem, interpreting evidence about the outcomes of each option, and ascertaining patient

From the Departments of Medicine (Drs Hayward and Guyatt) and Clinical Epidemiology and Biostatistics (Drs Hayward and Guyatt), McMaster University, Hamilton, Ontario; Division of Internal Medicine, Johns Hopkins University School of Medicine, Baltimore, Md (Dr Bass); Health Program, Office of Technology Assessment, US Congress, Washington, DC (Dr Tunis); and the Department of Medicine, Bowman Gray School of Medicine of Wake Forest University, Winston-Salem, NC (Dr Wilson).

A complete list of members (with affiliations) of the Evidence-Based Medicine Working Group appears in the first article of this series (*JAMA*. 1993;270:2093-2095). The following members contributed to this article: Deborah Cook, MD, MSc; Brian Haynes, MD, MSc, PhD; Roman Jaeschke, MD, MSc; Andreas Laupacis, MD, MSc; Virginia Moyer, MD, MPH; David Naylor, MD, DPhil; John Philbrick, MD; W. Scott Richardson, MD; David Sackett, MD, MSc; and Stephen Walter, PhD.

Reprint requests to Room 2C12, McMaster University Health Sciences Centre, 1200 Main St W, Hamilton, Ontario, Canada L8N 3Z5 (Dr Guyatt).

Users' Guides to the Medical Literature section editor: Drummond Rennie, MD, Deputy Editor (West), *JAMA*.

preferences for each outcome. Increasingly, clinicians must also consider the resource implications of their decisions. This involves detecting, treating, palliating, and preventing health problems in a way that maximizes the public good achieved with available resources.

To meet patients' expectations, individually and in aggregate, clinicians face intimidating tasks of information management. Overviews can help by systematically gathering, selecting, and combining evidence that links options to outcomes. Clinical decision analyses can help by refining questions and exploring the trade-offs between competing benefits and harms. Economic analyses can help by tallying the costs associated with different options. While useful, these approaches do not always synthesize information in a way that directly supports specific clinical recommendations.

Clinical practice guidelines, which have been defined as "systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances,"[7] represent an attempt to distill a large body of medical knowledge into a convenient, readily usable format.[8] Like overviews, they gather, appraise, and combine evidence. Guidelines, however, go beyond most overviews in attempting to address all the issues relevant to a clinical decision and all the values that might sway a clinical recommendation. Like decision analyses, guidelines refine clinical questions and balance trade-offs. Guidelines differ from decision analyses in relying more on qualitative reasoning and in emphasizing a particular clinical context.

Guidelines make explicit recommendations, often on behalf of health organizations, with a definite intent to influence what clinicians do. These suggestions about what should be done go beyond a simple presentation of evidence, costs, or decision models. They reflect value judgments about the relative importance of various health and economic outcomes in specific clinical situations. As a result, they should be required to pass unique tests about how matters of opinion, in addition to matters of science, are handled.

When appraising a consultant's counsel, we are impressed if she states and explains her suggestions clearly, discusses alternatives, and acknowledges possible biases and extenuating circumstances. We can use this common-sense approach to assess the validity, importance, and applicability of clinical practice guidelines. In this article, we offer suggestions for deciding whether to use a clinical practice guideline in formulating one's own clinical policies (Table). Our focus is on evaluation of interventions—including prevention, diagnosis, and therapy—that are designed to improve important patient outcomes. For prevention and diagnosis, this involves looking beyond the accuracy of the test to the ultimate consequences of choosing a diagnostic strategy on patients' morbidity, mortality, and health-related quality of life.

Guidelines for How to Use Articles Describing Clinical Practice Guidelines

**Are the recommendations valid?**
- Primary guides:
  - Were all important options and outcomes clearly specified?
  - Was an explicit and sensible process used to identify, select, and combine evidence?
- Secondary guides:
  - Was an explicit and sensible process used to consider the relative value of different outcomes?
  - Is the guideline likely to account for important recent developments?
  - Has the guideline been subject to peer review and testing?

**What are the recommendations?**
- Are practical, clinically important, recommendations made?
- How strong are the recommendations?
- What is the impact of uncertainty associated with the evidence and values used in the guidelines?

**Will the recommendations help you in caring for your patients?**
- Is the primary objective of the guideline consistent with your objective?
- Are the recommendations applicable to your patients?

We use the same basic questions as the users' guides for original research articles, overviews, and decision analyses. Are the recommendations valid? If they are, what are the recommendations and will they be helpful in patient care? To answer these questions, we draw on an emerging literature about practice guideline development and evaluation[9-15] (and S. H. Woolf, unpublished data, 1991), while emphasizing the perspective of practitioners who must adopt, adapt, or reject recommendations. Busy clinicians might hope that criteria for appraising practice guidelines would obviate the need for reviewing how the guideline developers have brought together the evidence, and how they have chosen the values reflected in their recommendations. Unfortunately, any shortcuts that bypass at least a cursory look at evidence and values will leave the clinician open to being misled by guidelines that may be based on a biased selection of evidence, a skewed interpretation of that evidence, or an idiosyncratic set of values. Shortcuts that do not highlight health conditions and interventions, patients and practitioners, and benefits and harms will leave the clinician open to misapplication of guidelines in clinical practice.

## ARE THE RECOMMENDATIONS VALID?

### Primary Guides

You need to determine whether guideline developers used appropriate methods and adduced evidence that support the recommendations made. If developers do not include—either in their policy statement or in a supporting article—information about how they chose options and outcomes, selected evidence, and decided on values, you might suspect that these steps were not done systematically.[16] In any case, you cannot evaluate such guidelines, and their recommendations probably should not influence your decision making.

**Were All Important Options and Outcomes Considered?**—Guidelines pertain to decisions and decisions involve choices and consequences. To appreciate why a particular practice is recommended, you should check to see that guideline developers have considered all reasonable practice options and all important potential outcomes.

Whether developers present guidelines for prevention, diagnosis, therapy, or rehabilitation, they should specify both the interventions of interest and sensible alternative practices. For example, in a guideline based on a careful systematic literature review,[17] the ACP offers recommendations about medical interventions for preventing strokes.[18] While carotid endarterectomy is mentioned as a possible surgical intervention in the preamble to the guideline, the procedure is not considered in the recommendations themselves. This guideline could have been strengthened if medical interventions for transient ischemic attacks had been placed in a management context that included the highly effective surgical procedure.[19]

In its HRT guideline, the ACP makes recommendations about counseling women who are postmenopausal and are considering HRT to prevent disease and prolong life.[4] The interventions they considered were (1) long-term daily prophylaxis (10 to 20 years) with 0.625 mg of oral conjugated estrogen, (2) daily estrogen and medroxyprogesterone acetate (2.5 mg orally per day or 5 to 10 mg on days 10 to 14 of the month), (3) short-term HRT therapy (1 to 5 years), or (4)

no prophylactic hormone use. The guideline did not consider calcium supplementation, newer estrogen delivery systems, or other approaches to the prevention of osteoporosis-related fractures.

Guideline developers must consider not only all the best management options, but all the important consequences of the options. As a clinician looking after individual patients, you look for information on morbidity, mortality, and quality of life and you must decide if the guideline ignores outcomes that your patients would care about. As a practitioner interested in using resources efficiently, you must also mind economic outcomes. Whether developers examine economic outcomes at all—and if they do, whether they look at costs from the patients', insurers', or health care system perspective, or consider broader issues such as the consequences of time lost from work—can strongly influence final recommendations.[20] The majority of published guidelines do not include formal cost analyses, those that do use a variety of analytic techniques, and it will be difficult for you to determine whether actual cost estimates are valid or applicable for your practice setting. You can gain a better understanding of the potential importance of these issues by seeing if the economic projections are subjected to sensitivity analysis. If so, you can gauge the extent to which guideline recommendations might change if assumptions about costs change. You can also check to see if the guideline developers offer clinically relevant comparisons. For example, the average cost of preventing one cardiovascular-related death by means of HRT might be compared with the cost of doing the same by means of cholesterol reduction, blood pressure control, or smoking cessation counseling.

In its HRT guideline, the ACP used lifetime probability of developing endometrial cancer, breast cancer, hip fracture, coronary heart disease, and stroke, and median life expectancy to estimate risks and benefits for subgroups of women. They acknowledged possible HRT effects on serum lipoproteins, uterine bleeding, sexual and urinary function, and the need for endometrial surveillance by biopsy, but did not include these considerations in the model used to synthesize evidence. The effects of HRT on costs and quality of life, which could have a major impact on patient choices, were not explicitly considered.

**Was an Explicit and Sensible Process Used to Identify, Select, and Combine Evidence?**—Having specified options and outcomes, the next task in decision making is to estimate the likelihood that each outcome will occur. In effect, one has a series of specific questions. For HRT, what is the effect of the alternative approaches on hip fracture incidence, on myocardial infarction and coronary death, or on breast and endometrial cancer incidence? Guideline developers must bring together all the relevant evidence, and then combine that evidence in an appropriate manner. In carrying out this task, they must avoid bias that will distort the results. In effect, they must have access to, or conduct, a systematic overview of the evidence bearing on each question they address.

The users' guide on overviews includes criteria that can be used to judge whether guideline developers have done an adequate job in accumulating and synthesizing evidence.[21] Developers should specify a focused question, define appropriate evidence using explicit inclusion and exclusion criteria, conduct a comprehensive search, and examine the validity of the results in a reproducible fashion.

The best guidelines define admissible evidence, report how it was selected and combined, make key data available for your review, and report that they found randomized trials that link the interventions to the outcomes. Such randomized trials may, however, be unavailable, and guideline developers are in a different position from the authors of overviews who may abandon their project if there are not any high-quality studies to summarize. Many important clinical problems are technically, economically, or ethically difficult to address with randomized clinical trials. Because guideline developers must deal with inadequate evidence, they may have to consider a variety of studies as well as reports of expert and consumer experience. They must formulate recommendations, but they should be candid about the type and quantity of evidence on which those recommendations are based.

The nature and appropriate use of expertise is one of the most hotly debated areas in guideline development. Sometimes "experts" have preeminent knowledge of the basic science, pathophysiology, and natural history of a health condition. They may also be distinguished by extensive direct clinical experience. Persons who have witnessed and understood the limitations of clinical trials in the clinical domain offer another dimension of expertise. For some guidelines, extra emphasis may be placed on the expertise of generalists who can gauge the practical implications of interventions applied to large groups. Although the RAND Corporation and others have developed protocols for recording and quantifying expert assessments of the appropriateness of health interventions,[22,23] guideline developers must decide what type of expert opinion to solicit and how to incorporate it into the evidential foundation for guideline development. You are unlikely to find systematic methods for selecting, capturing, and grading relevant expertise in today's guidelines, but you should try to determine whether and how expert opinion was used to fill in gaps in the evidence from clinical trials.

A quality-of-evidence scale can be used to rate different categories of evidence (eg, expert opinion or clinical investigation) and methods for producing it (eg, blinded or nonblinded outcome assessment) according to the likelihood that the source or design will yield biased results.[24] Developers working on a different problem with a different supporting literature may devise an evidence-filtering instrument that stratifies case-control studies into categories of differing quality.[25] The prospective development and application of a systematic approach to appraising and classifying evidence is important because this means that the strength of the evidence in support of the recommendations can be reported. Strategies for summarizing the strength of both evidence and recommendations will be addressed in the second of our articles about using practice guidelines, which deals with interpreting and applying the results.

The ACP HRT guideline developers searched MEDLINE (1970 to 1991) and citations from articles, and conferred with expert consultants to identify studies published in English about the treatment options and outcomes. They conducted formal overviews, including meta-analysis, and derived summary estimates of relative risks and lifetime probabilities of the principal outcomes with and without HRT for subgroups of women. These subgroups included women without risk factors; women at increased risk for coronary disease, hip fracture, or breast cancer; and women who had a hysterectomy. Their overviews met the validity criteria we have suggested. In most cases, randomized trials had not been conducted, and the investigators relied on observational studies. Therefore, they appropriately conducted sensitivity analyses to determine the implications if the results of observational studies represented overestimates or underestimates of the true effect of the interventions on the relevant outcomes.

## Secondary Guides

**Was an Explicit and Sensible Process Used to Consider the Relative Value of Different Outcomes?**—Link-

ing treatment options to outcomes is largely a question of fact and a matter of science. In contrast, assigning preferences to outcomes is largely a question of opinion and a matter of value. The extent to which HRT increases the incidence of breast cancer or decreases death rates from myocardial infarction can be ascertained from the evidence. The relative importance placed on avoiding breast cancer or cardiovascular disease depends on what patients care about most. Consequently, it is important that guideline developers report the sources of their value judgments and the method by which consensus was sought.

You should look for information about who was *explicitly* involved in assigning values to outcomes, or who, by influencing recommendations, was *implicitly* involved in assigning values. Expert panels and consensus groups are often used to determine what a guideline will say. You need to know who the panel members are, bearing in mind that panels dominated by members of specialty groups may be subject to intellectual, territorial, and even financial biases (some organizations screen potential panel members for conflicts of interest, others do not). By identifying the agencies that have sponsored and funded guideline development, you can decide whether their interests or delegates are overrepresented on the consensus committee. Panels that include a balance of research methodologists, practicing generalists and specialists, and public representatives are more likely to have considered diverse views in their deliberations.

Even with broad representation, the actual process of deliberation can influence recommendations. You should therefore look for a report of methods used to synthesize preferences from multiple sources. Informal and unstructured processes for arbitrating values may be vulnerable to undue influence by individual panel members, particularly the panel chair. Appropriate structured processes increase the likelihood that all important values are duly considered.[26]

It is particularly important to know how patient preferences were considered. Health interventions have beneficial and harmful effects along with associated costs, and recommendations may differ depending on our relative emphasis on specific benefits, harms, and costs. What is the relative importance of an uncertain risk for increases in breast cancer vs a fairly clear expectation of decreased incidence of heart attacks and strokes? Many guideline reports, by their silence on the matter of patient preferences, assume that guideline developers adequately represent patients' interests. Methods for directly assessing patient and societal values exist but are rarely used by guideline developers. You may be limited to gauging whether the values implicit in the guideline appear to favor patient, third-party (eg, reimbursement agencies), or societal priorities.[27] You can also consider which ethical principles—such as patient autonomy (the patient's control over decisions about her health), nonmaleficience (avoiding harm), or distributive justice (the fair distribution of health care resources)—prevailed in guiding decisions about the value of alternative interventions. For guidelines based on formal risk-benefit and cost-benefit analyses, declarations of acceptable levels of risks and costs per benefit achieved can help you make comparisons across guidelines.

Variation (disagreement) and uncertainty (ambivalence) in values could affect summary recommendations and so should be recorded and reported by guideline developers. The clinical problems for which practice guidelines are most needed often involve complex trade-offs between competing benefits, harms, and costs, usually under conditions of uncertainty. Even in the presence of strong *evidence* from randomized clinical trials, the effect size of an intervention may be marginal or the intervention may be associated with costs, discomforts, or impracticalities that lead to disagreement or ambivalence among guideline developers about what to *recommend*. Explicit strategies for documenting, describing, and dealing with dissent among judges, or frank reports of the degree of consensus attained, can help you decide whether to adopt or adapt recommendations. Unfortunately, until guideline development methods mature, you will rarely find this information.

An example of the implicit, and perhaps questionable, value judgments guideline developers make comes from the ACP recommendations for medical therapies to prevent stroke.[17] This guideline recommended that aspirin be considered the drug of choice in patients with transient ischemic attacks, and suggested that ticlopidine be reserved for patients who do not tolerate aspirin. The best estimate of the effect of ticlopidine relative to aspirin in patients with transient ischemic attacks is a 15% reduction in relative risk, a benefit that would translate into preventing one stroke for every 70 patients treated in a group of patients with a 10% risk of stroke. The ACP presumably makes their recommendation that aspirin, not ticlopidine, be the drug of choice for patients with transient ischemic attack on the basis of the increased cost of ticlopidine, and the need for checking the white blood cell count in patients receiving ticlopidine. This implicit value judgment could be questioned, and the guideline would be strengthened if the authors had made the values that underlie their judgment explicit.

In the case of the ACP HRT guideline, the developers gave priority to outcomes that are major contributors to morbidity and mortality in North America (eg, the effect of long-term estrogen use on risk of death from myocardial infarction, osteoporosis-related fractures, and endometrial cancer), but acknowledged that other considerations may be as important as preventing disease and death for some women (eg, resumption of menses, changes in mood, and sexual function). The task of assigning relative value to different types of morbidity or causes of mortality is left to patients and their clinicians.

**Is the Guideline Likely to Account for Important Recent Developments?**—Guidelines often concern controversial health problems about which new knowledge is actively sought in ongoing studies. Because of the time required to assemble and review evidence and achieve consensus about recommendations, the guideline may be out of date by the time you see it. You should look for two important dates: the publication date of the most recent evidence considered and the date on which the final recommendations were made. Some authorities also identify important studies in progress and new information that could change the guideline. Ideally, these considerations may be used to qualify guidelines as "temporary" or "provisional," to specify dates for expiration or review, or to identify key research priorities. For most guidelines, however, you must scan the bibliography to get an impression of how current a particular guideline may be. The ACP HRT guideline gives dates for evidence considered (1970 through 1991) and final approval (March 1992). The guideline acknowledged that its advice about use of estrogen in combination with a progestin was limited by uncertainty about whether the progestin neutralizes the beneficial effects of estrogen on risk factors for unwanted cardiovascular outcomes. The guideline did not alert readers to watch for results from the Postmenopausal Estrogen/Progestin Interventions (PEPI) trial, initiated in 1988, which would directly address that uncertainty. An early report from the PEPI group concludes that estrogen alone or in combination with a progestin improves lipoprotein levels and lowers fibrinogen levels without detectable effects on insulin or blood pressure.[28]

**Has the Guideline Been Subjected to Peer Review and Testing?**—People may interpret evidence differently and their values may differ, and guidelines are subject to both sorts of differences. Your confidence in the validity of a guideline increases if external reviewers have judged the conclusions reasonable, and clinicians have found the guidelines applicable in practice. If the guidelines differ from those adduced by others, you should look for an explanation. On the other hand, if the guidelines meet the first four validity criteria and the underlying evidence is strong, rejection by clinicians or peer reviewers may have more to do with their biases than to any limitation in the validity of the guidelines.

If the underlying evidence is weak, no matter what the degree of consensus or peer review, the clinicians' confidence in the validity of the guideline will be limited. In the second part of our users' guide for practice guidelines, we will describe explicit frameworks for judging the strength of recommendations. The weaker the underlying evidence, the greater the argument for actually testing the guideline to determine whether its application improves patient outcomes.[29] The question for any such test would be: are patient outcomes better, or are outcomes equivalent at decreased cost, when clinicians operate on the basis of the practice guidelines?

Weingarten and colleagues[30] conducted such an investigation examining the impact of implementation of a practice guideline suggesting that low-risk patients admitted to coronary care units should receive early discharge.[30] On alternate months over the period of a year, clinicians either received or did not receive a reminder of the guideline recommendations. During the months in which the intervention was in effect, hospital stay for coronary care unit patients was approximately a day shorter and the average cost of the stay was over $1000 less. Mortality and health status at 1 month were similar in the two groups. The investigators concluded that the guideline reminder reduced hospital stay and associated costs without adversely affecting measured patient outcomes. Although in this case the authors used alternate-month allocation, which makes the study weaker than a true randomized trial, a study of this type helps to validate the predicted consequences of guideline implementation for defined outcomes.

Once you are confident that the clinical practice guideline addresses your clinical question and is based on a rigorous up-to-date assessment of the relevant evidence, you can review the recommendations to determine how useful they will be in your practice. While not pristine, the ACP guidelines on HRT do a good job at meeting the primary criteria for using a practice guideline. We will describe how to interpret and apply the results in the next article of this series.

We offer special thanks to Deborah Maddock who has provided outstanding administrative support and coordination for the activities of the Evidence-Based Medicine Working Group.

**References**

1. American College of Obstetricians and Gynecologists. Hormone replacement therapy. *Int J Gynaecol Obstet.* 1993;41:291-297. (ACOG Technical Bulletin No. 166—April 1992 [replaces No. 93, June 1986]).
2. American College of Obstetricians and Gynecologists. Report of Task Force on Routine Cancer Screening. In: *Standards for Obstetric-Gynecologic Services.* Washington, DC: American College of Obstetricians and Gynecologists; 1989:97-104.
3. Wallace WA. HRT and the surgeon: guidelines from the Royal College of Surgeons of Edinburgh (November 1992). *J R Coll Surg Edinb.* 1993;38:58-61.
4. American College of Physicians. Guidelines for counseling postmenopausal women about preventive hormone therapy. *Ann Intern Med.* 1992;117:1038-1041.
5. Moy JG, Realini JP. Guidelines for postmenopausal preventive hormone therapy: a policy review: American College of Physicians. *J Am Board Fam Pract.* 1993;6:153-162.
6. Grady D, Rubin SM, Petitti DB, et al. Hormone therapy to prevent disease and prolong life in postmenopausal women. *Ann Intern Med.* 1992;117:1016-1037.
7. Institute of Medicine. *Clinical Practice Guidelines: Directions for a New Program.* Washington, DC: National Academy Press; 1990.
8. Eddy DM. The challenge. *JAMA.* 1990;263:287-290.
9. American Medical Association. *Attributes to Guide the Development of Practice Parameters.* Chicago, Ill: American Medical Association; 1990.
10. American College of Physicians. *Clinical Efficacy Assessment Project: Procedural Manual.* Philadelphia, Pa: American College of Physicians; 1986.
11. Gottlieb LK, Margolis CZ, Schoenbaum SC. Clinical practice guidelines at an HMO: development and implementation in a quality improvement model. *QRB.* 1990;16:80-86.
12. Institute of Medicine. *Guidelines for Clinical Practice: From Development to Use.* Washington, DC: National Academy Press; 1992.
13. Eddy DM. *A Manual For Assessing Health Practices and Designing Practice Policies: The Explicit Approach.* Philadelphia, Pa: American College of Physicians; 1992.
14. Park RE, Fink A, Brook RH, et al. Physicians' rating of appropriate indications for six medical and surgical procedures. *Am J Public Health.* 1986;76:766-772.
15. Hayward RSA, Laupacis A. Initiating, conducting and maintaining guideline development programs. *Can Med Assoc J.* 1993;148:507-512.
16. Hayward RSA, Tunis SR, Wilson MC, Bass EB, Rubin HR, Haynes RB. More informative abstracts of articles describing clinical practice guidelines. *Ann Intern Med.* 1993;118:731-737.
17. Matchar DB, McCory DC, Barnett HJM, Feussner JR. Medical treatment for stroke prevention. *Ann Intern Med.* 1994;121:41-53.
18. American College of Physicians. Guidelines for medical treatment for stroke prevention. *Ann Intern Med.* 1994;121:54-55.
19. North American Symptomatic Carotid Endarterectomy Trial Collaborators. Beneficial effect of carotid endarterectomy in symptomatic patients with high-grade carotid stenosis. *N Engl J Med.* 1991;325:445-453.
20. Hayward RSA, Steinberg EP, Ford DE, Roizen MF, Roach K. Preventive care guidelines: 1991. *Ann Intern Med.* 1991;114:758-783.
21. Oxman AD, Cook DJ, Guyatt GH, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VI: how to use an overview. *JAMA.* 1994;272:1367-1371.
22. Kanouse DE, Winkler JD, Kosecoff J, et al. *Changing Medical Practice Through Technology Assessment: An Evaluation of the NIH Consensus Development Program.* Ann Arbor, Mich: Health Administration Press; 1989.
23. Merrick JN, Fink A, Brook RH, et al. Carotid endarterectomy. In: *Indications for Selected Medical and Surgical Procedures: A Literature Review and Ratings of Appropriateness.* Santa Monica, Calif: The Rand Corporation; 1986:41-47.
24. Braunwald E, Mark DB, Jones RH, et al. *Unstable Angina: Diagnosis and Management: Clinical Practice Guideline Number 10.* Rockville, Md: US Dept of Health and Human Services; 1994.
25. Cataract Management Guideline Panel. *Cataract in Adults: Management of Functional Impairment: Clinical Practice Guideline Number 4.* Rockville, Md: Agency for Health Care Policy and Research; 1993.
26. Sackman H. *Delphi Critique: Expert Opinion, Forecasting, and Group Process.* Lexington, Mass: Lexington Books; 1975.
27. Diamond GA, Denton TA. Alternative perspectives on the biased foundations of medical technology assessment. *Ann Intern Med.* 1993;118:455-464.
28. The Writing Group for the PEPI Trial. Effects of estrogen or estrogen/progestin regimens on heart disease risk factors in postmenopausal women: the Postmenopausal Estrogen/Progestin Interventions (PEPI) trial. *JAMA.* 1995;273:199-208.
29. Health Services Research Group. Standards, guidelines and clinical policies. *Can Med Assoc J.* 1992;146:833-837.
30. Weingarten SR, Reidinger MS, Conner L, et al. Practice guidelines and reminders to reduce duration of hospital stay for patients with chest pain. *Ann Intern Med.* 1994;120:257-263.

# Users' Guides to the Medical Literature

## VIII. How to Use Clinical Practice Guidelines

## B. What Are the Recommendations and Will They Help You in Caring for Your Patients?

Mark C. Wilson, MD, MPH; Robert S. A. Hayward, MD, MPH; Sean R. Tunis, MD, MSc; Eric B. Bass, MD, MPH; Gordon Guyatt, MD, MSc; for the Evidence-Based Medicine Working Group

### CLINICAL SCENARIO

At the conclusion of our first article on practice guidelines[1] in this series, we left you examining the full text of a practice guideline[2] that could help you marshal a convincing response to a colleague who disagrees with your approach to hormone replacement therapy (HRT) in postmenopausal women. Later that day, chatting with another colleague, you mention the disagreement. He shrugs, and avows, "It's entirely a matter of personal preference, the evidence doesn't support either of you." You return to the guideline, looking for how particular recommendations may be justified and adapted to your patient's circumstances.

### WHAT ARE THE RECOMMENDATIONS?

#### Are Practical, Clinically Important, Recommendations Made?

To be useful, recommendations should give practical, unambiguous advice about a specific health problem. For guidelines about managing health conditions, you should determine if the intent is to prevent, screen for, diagnose, treat, or palliate the disorder. For guidelines about the appropriate uses of health interventions, the recommendations should include a definition of the intervention and its optimal role in patient management. In the American College of Physicians (ACP) guideline on HRT,[2] recommendations are divided into general observations that can help the clinician discuss with patients the effects of therapy, and specific management recommendations concerning what should be done in patient evaluation, risk assessment, hormone administration, and follow-up to achieve the outcomes predicted by the available evidence.

To be clinically important, a practice guideline should convince you that the benefits of following the recommendations are worth the expected harms and costs. You should consider both the relative and absolute changes in outcomes. A 25% reduction in relative risk of death from a disease is much more compelling if it involves a reduction in the proportion of deaths from 40 of 100 to 30 of 100 (an absolute risk reduction of 10 in 100), than if it involves a reduction in the proportion of deaths from four of 100 to three of 100 (an absolute risk reduction of one in 100).[3]

The ACP guideline cites extensive and consistent observational data to show that unopposed estrogen therapy (ET) reduces the lifetime risk of developing coronary heart disease (CHD) by about 35% (for 50-year-old women with no extraordinary CHD risks, about 12 of 100 would be spared CHD in their lifetimes) and hip fractures by about 15% (two to three of 100 avoid hip fracture because of ET use). In women who have a uterus and take unopposed ET, the risk of developing endometrial cancer increases up to eightfold (approximately 17 women of 100 who take ET and would not otherwise have developed endometrial cancer will develop the disease) and the risk for breast cancer may increase as much as 25% (absolute increase of about three of 100 women). Clearly, the relative increases or decreases in outcomes can be misleading if baseline risks and absolute changes in outcomes are not reported. Addition of progestin maintains hip fracture risk reduction and removes the increased risk of endometrial cancer, but has uncertain effects on risks for breast cancer and cardiovascular disease. Hormone replacement therapy can increase life expectancy by 10 months to 2 years, depending on the presence of risk factors, a gain similar to that achieved by treatment of hypertension. The guideline did not consider personal or societal costs associated with HRT.

#### How Strong Are the Recommendations?

The "strength," "grade," "confidence," or "force" of a recommendation should be informed by multiple considerations: the quality of the investigations that provide the evidence for the recommendations, the magnitude and consistency of positive outcomes relative to negative outcomes (adverse effects, burdens to the patient and the health care system, costs), and the relative value placed on different outcomes. Even in the presence of strong *evidence* from randomized clinical trials, the effect size of an intervention may be marginal. The intervention may be associated with costs, discomforts, or impracticalities that downgrade the strength of a summary *recommendation* about what practicing clinicians should do. It is important to consider this distinction and to scrutinize a guideline document for what, in addition to evidence, determines the wording of actual recommendations. These factors are key to understanding conflicts among guidelines on similar topics from different organizations.[4]

From the Division of Internal Medicine, Johns Hopkins University School of Medicine, Baltimore, Md (Dr Bass); the Departments of Medicine (Drs Hayward and Guyatt) and Clinical Epidemiology and Biostatistics (Drs Hayward and Guyatt), McMaster University, Hamilton, Ontario; Health Program, Office of Technology Assessment, US Congress, Washington, DC (Dr Tunis); and the Department of Medicine, Bowman Gray School of Medicine of Wake Forest University, Winston-Salem, NC (Dr Wilson).

A complete list of members (with affiliations) of the Evidence-Based Medicine Working Group appears in the first article of this series (*JAMA*. 1993;270:2093-2095). The following members contributed to this article: Deborah Cook, MD, MSc; Brian Haynes, MD, MSc, PhD; Roman Jaeschke, MD, MSc; Andreas Laupacis, MD, MSc; Virginia Moyer, MD, MPH; David Naylor, MD, DPhil; John Philbrick, MD; Scott Richardson, MD; David Sackett, MD, MSc; and Stephen Walter, PhD.

Reprint requests to Room 2C12, McMaster University Health Sciences Centre, 1200 Main St W, Hamilton, Ontario, Canada L8N 3Z5 (Dr Guyatt).

Users' Guides to the Medical Literature section editor: Drummond Rennie, MD, Deputy Editor (West), *JAMA*.

In our first article about using practice guidelines,[1] we pointed out that the best available evidence about the effects of health interventions may come from sources as diverse as, on the one hand, well-conducted randomized trials and, on the other, expert opinion. Thus, users of practice guidelines will find tremendous variability in strength of the evidence linking options and outcomes. Among guidelines developed by different groups about the same health condition or intervention, there should be little variability in estimates of the strength of evidence as long as the supporting overviews considered the same body of literature.[5-7] Here, differences in recommendations probably reflect differences in the relative value placed on various health and economic outcomes.[8] Unfortunately, these considerations are rarely exposed in guideline documents and there is no commonly accepted approach for grading evidence or recommendations.[9-12]

Formal taxonomies of "levels of evidence" and "grades of recommendations" were first popularized by the Canadian Task Force on the Periodic Health Examination,[13] and later revised in cooperation with the United States Preventive Services Task Force.[9] Like previous articles in this series,[14] these guideline developers emphasized that the strongest evidence comes from rigorous randomized controlled trials and weaker evidence from observational studies using cohort or case-control designs. Inferring strength of evidence from study design alone, however, may overlook other determinants of the quality of evidence, such as sample size, recruitment bias, losses to follow-up, unmasked outcome assessment, atypical patient groups, unreproducible interventions, impractical clinical settings, and other threats to internal and external validity. Moreover, results from a single randomized controlled trial with a small sample size are not necessarily more convincing than consistent results with high precision from a large number of high-quality trials of nonrandomized design conducted in a variety of places and times. Recent proposals for summarizing strength of evidence have emphasized the need for overviews to filter out studies with major design flaws, and meta-analyses to consider the precision, magnitude, and heterogeneity of study results.[11] The United States Preventive Services Task Force now supplements its "study design categories" with prose descriptions of flaws in the published evidence.[15]

Another approach to categorizing evidence from multiple studies offers a hierarchy from overviews of observational studies with inconsistent results to overviews of randomized controlled trials with consistent results (Table).[16] Since inferences about the health effects of interventions are weakened when there are unexplained major differences in effects in different studies, guidelines based on randomized controlled trials are stronger when the results of individual studies are similar, and weaker when major differences between studies (heterogeneity) are present. If the evidence linking interventions and outcomes came from overviews of articles, you could apply the criteria for a valid overview and the schema in the Table to decide on the strength of evidence supporting recommendations.

This approach is constrained by its focus on only one major outcome (for HRT we are interested in many outcomes), but it exemplifies how the strength of evidence and the strength of recommendations could be integrated on a common scale. It considers study design, heterogeneity, effect size, confidence intervals (CIs) around the effect sizes, and threshold effect sizes over which negative outcomes outweigh the benefits. The threshold effect size presumes value judgments about the relative importance of various outcomes resulting from the health intervention have been applied. In principle, strong recommendations are warranted when the smallest effect compatible with the data (the lower boundary of the CI) is still greater than the threshold below which the negative outcomes outweigh the benefits. (In an upcoming article[16] in this series, we describe this approach to levels of recommendation in much more detail.)

If the guidelines are developed on the basis of observational studies or if the estimate of the treatment effect is imprecise, the user should not expect strong recommendations unless major harms and costs are associated with the intervention or a catastrophic outcome (eg, death) may be prevented by a low-risk, low-cost intervention of probable efficacy. Guideline developers could compensate for weak evidence by testing the effect of their guideline on patient outcomes in a real-world clinical situation.[17] Such a study, if methodologically strong, could enhance the strength of the recommendations in the absence of strong evidence from original studies.

While the ACP HRT guideline does not grade its recommendations, the guideline does cross-reference recommendations to discussions about evidence and effect sizes in the associated overview. Because the guideline is based largely on observational studies, the recommendations are relatively weak, and would be categorized as C1 in the schema in the Table.

Grades of Recommendations for a Specified Level of Baseline Risk*

| | |
|---|---|
| A1 | RCTs, no heterogeneity, CIs all on one side of threshold NNT |
| A2 | RCTs, no heterogeneity, CIs overlap threshold NNT |
| B1 | RCTs, heterogeneity, CIs all on one side of threshold NNT |
| B2 | RCTs, heterogeneity, CIs overlap threshold NNT |
| C1 | Observational studies, CIs all on one side of threshold NNT |
| C2 | Observational studies, CIs overlap threshold NNT |

*RCT indicates randomized controlled trial; CI, confidence interval; and NNT, number needed to treat to avoid one unwanted outcome.

### What Is the Impact of Uncertainty Associated With the Evidence and Values Used in the Guidelines?

Guideline developers should consider the possibility that the effect of a management option on an outcome, or the relative value of different outcomes, is much greater, or much less, than their best estimate. We have discussed how to examine this possibility, a process we call sensitivity analysis, in the users' guide for decision analysis.[18] The weaker the evidence linking intervention and outcome, and the greater the possible range of competing values, the greater the need for a sensitivity analysis. For example, the range of plausible estimates of the impact of HRT on breast cancer is very wide, and guideline developers should test how their recommendations would differ across the range of possible effects. When the evidence is of the weakest sort, arising from expert opinion, sensitivity analysis is essential.

The authors of the HRT guideline acknowledge that the observational design of the studies may introduce bias, and they alert us to areas where the evidence is particularly weak (such as the effect of combined estrogen and progestins on breast cancer). They don't, however, provide a formal sensitivity analysis. Such a sensitivity analysis might have been useful in highlighting the uncertainty of many of the estimates on which the recommendations are based, particularly those relating to life expectancy.

## WILL THE RECOMMENDATIONS HELP YOU IN CARING FOR YOUR PATIENTS?

### Is the Primary Objective of the Guideline Consistent With Your Objectives?

You should try to anticipate how a guideline will be used. Guidelines may be disseminated to assist physicians with clinical decision making (for example, clinical algorithms and reminders), to enable evaluation of physician practices (eg, uti-

lization review, quality assurance), or to set limits on physician choices (eg, recertification, reimbursement). Guidelines may be directed at different practitioners. Some guidelines about detection and treatment of depression have, for example, aimed to guide primary care providers and others to guide psychiatrists.[19] You should ensure the purpose of the guideline meets the use you intend for it.

### Are the Recommendations Applicable to Your Patients?

To be really useful, guidelines should describe interventions well enough for their exact duplication. You must determine whether your patients are the intended target of a particular guideline. If your patients have a different prevalence of disease or risk factors, for instance, the guidelines may not apply.

The flexibility of the guideline may be indicated by patient or practice characteristics that require individualizing recommendations or that justify departures from the recommendations. For example, the American College of Cardiology, the American Heart Association, and the ACP advise against using electrocardiograms to screen asymptomatic adults, but they acknowledge that this advice may not be valid for persons who smoke; are male and of "increased age"; have a family history of coronary artery disease; have hypertension, diabetes, or other cardiovascular risk factors; are sedentary; or whose occupation affects public safety.[20-24] The caveats reflect reluctance to make recommendations in the absence of good evidence. They also exclude groups of patients who, in total, may account for a majority of an internist's patients!

You should look for information that must be obtained from and provided to patients and for patient preferences that should be considered. It is important to consider whether the values assigned (implicitly or explicitly) to outcomes could differ enough from your patients' preferences to change a decision about whether to adopt a recommendation.

When you review the HRT guidelines, you may begin to understand why your colleague in the scenario with which this article began felt that recommendations regarding HRT must be different for every patient. In its HRT guideline, the ACP offers separate recommendations for women at increased risk for CHD, hip fracture and breast cancer, and for women who have had a hysterectomy. These different recommendations reflect the fact that different women are at varying risk of adverse outcomes, and the impact of HRT on them will therefore differ. The most vivid example is women who have had hysterectomies: since they are not at risk of endometrial cancer, unopposed estrogen is much more likely to be the right treatment choice.

## RESOLUTION OF THE SCENARIO

The ACP recommends that all women consider taking preventive hormone therapy, while admitting that no evidence supports strong advice except for some women who are at increased risk for some outcomes. The guidelines suggest that women at increased risk for CHD are likely to achieve longevity gains from HRT, but that conclusion needs to be confirmed by randomized controlled trials. Hormone replacement therapy is likely to decrease the risk of hip, vertebral, and wrist fractures, but, without a progestin, risks for endometrial cancer increase up to eightfold. Women who have had a hysterectomy should take ET alone; others should add a progestin or comply with careful endometrial monitoring. The effect of estrogen on breast cancer appears to be small, but the evidence is weak and many women may not be willing to "take a chance," particularly if they bear low or average risks for CHD. Clinicians should assess risks, estimate benefits and harms, educate patients, and facilitate individualized decision making for all postmenopausal patients.

There is certainly much more to making decisions about HRT than perhaps you or your colleague had at first appreciated. There are many options, multiple outcomes, and significant trade-offs in benefits and harms. A good guideline, based on solid scientific evidence and an explicit process for judging the value of alternative practices, allows you to review, at one sitting, links between multiple options and outcomes. Unfortunately, well-developed and usefully summarized guidelines are still rare in the clinical literature. We hope that more consistent reporting of guideline development methods will prevail, making the guidelines literature more accessible to and useful for prospective guideline users.[25]

We offer special thanks to Deborah Maddock who has provided outstanding administrative support and coordination for the activities of the Evidence-Based Medicine Working Group.

#### References

1. Hayward RSA, Wilson MC, Tunis SR, Bass EB, Guyatt G, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VIII: how to use clinical practice guidelines, A: are the recommendations valid? *JAMA.* 1995;274:570-574.
2. American College of Physicians. Guidelines for counseling postmenopausal women about preventive hormone therapy. *Ann Intern Med.* 1992;117: 1038-1041.
3. Laupacis A, Naylor CD, Sackett DL. How should the results of clinical trials be presented to clinicians [editorial]? *ACP J Club.* May/June 1992:A12-A14. *Ann Intern Med.* Vol 116, suppl 3).
4. Eddy DM. Resolving conflicts in practice policies. *JAMA.* 1990;264:389-391.
5. Whelton PK. Reflections on the U.S. Preventive Services Task Force recommendations for screening for hypertension and hypercholesterolemia. *J Gen Intern Med.* 1990;5:S17-S19.
6. American College of Physicians. Screening low risk, asymptomatic adults for cardiac risk factors: serum serum cholesterol and triglycerides. In: Eddy DM, ed. *Common Screening Tests.* Philadelphia, Pa: American College of Physicians; 1991.
7. Canadian Task Force on the Periodic Health Examination. Periodic health examination, 1991 update: lowering blood cholesterol to prevent coronary heart disease. *Can Med Assoc J.* 1993;148:521-538.
8. Kottke TE, Brekke ML. Cholesterol policy: what should we do? how should we decide? *J Clin Epidemiol.* 1990;43:1023-1027.
9. Woolf SH, Battista RN, Anderson GM, Logan AG, Wang E. Assessing the clinical effectiveness of preventive maneuvers: analytic principles and systematic methods in reviewing evidence and developing clinical practice recommendations. *J Clin Epidemiol.* 1990;43:891-905.
10. Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. *Chest.* 1986;89(suppl 2):2S-3S.
11. Cook DJ, Guyatt GH, Laupacis A, Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. *Chest.* 1992;102 (suppl 4):305S-311S.
12. Acute Pain Management Panel. *Acute Pain Management: Operative or Medical Procedures and Trauma: Clinical Practice Guideline.* Rockville, Md: Agency for Health Care Policy and Research, Public Health Service, US Dept of Health and Human Services; 1992.
13. Canadian Task Force on the Periodic Health Examination. The periodic health examination. *Can Med Assoc J.* 1979;121:1193-1254.
14. Jaeschke R, Guyatt G, Sackett DL, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, III: how to use an article about a diagnostic test, A: are the results of the study valid? *JAMA.* 1994;271:389-391.
15. Battista RN, Fletcher SW. Making recommendations on preventive practices: methodological issues. *Am J Prev Med.* 1988;4S:53-67.
16. Guyatt GH, Sackett DL, Sinclair J, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, IX: a method for grading health care recommendations. *JAMA.* In press.
17. Weingarten SR, Reidinger MS, Conner L, et al. Practice guidelines and reminders to reduce duration of hospital stay for patients with chest pain. *Ann Intern Med.* 1994;120:257-263.
18. Richardson WS, Detsky AS, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VII: how to use a clinical decision analysis, B: what are the results and will they help me in caring for my patients? *JAMA.* 1995;273:1610-1613.
19. Rush AJ. Depression in primary care: detection, diagnosis and treatment. *Am Fam Physician.* 1993;47:1776-1788.
20. Sox HC Jr, Littenberg B, Garber AM. The role of exercise testing in screening for coronary artery disease. *Ann Intern Med.* 1989;110:456-469.
21. Sox HC Jr, Garber AM, Littenberg B. The resting electrocardiogram as a screening test: a clinical analysis. *Ann Intern Med.* 1989;111:489-502.
22. American College of Cardiology/American Heart Association Task Force on Assessment of Cardiovascular Procedures. Guidelines for exercise testing. *J Am Coll Cardiol.* 1986;8:725-738.
23. American College of Physicians. Screening for asymptomatic coronary artery disease: the resting electrocardiogram In: Eddy DM, ed. *Common Screening Tests.* Philadelphia, Pa: American College of Physicians; 1991.
24. American College of Physicians. Screening for asymptomatic coronary artery disease: exercise stress testing. In: Eddy DM, ed. *Common Screening Tests.* Philadelphia, Pa: American College of Physicians; 1991.
25. Hayward RSA, Tunis SR, Wilson MC, Bass EB, Rubin HR, Haynes RB. More informative abstracts of articles describing clinical practice guidelines. *Ann Intern Med.* 1993;118:731-737.

# Users' Guides to the Medical Literature

## IX. A Method for Grading Health Care Recommendations

Gordon H. Guyatt, MD; David L. Sackett, MD; John C. Sinclair, MD; Robert Hayward, MD; Deborah J. Cook, MD; Richard J. Cook, PhD; for the Evidence-Based Medicine Working Group

THE ULTIMATE PURPOSE of applied health research is to improve health care. Summarizing the literature to adduce recommendations for clinical practice is an important part of the process. Recently, the health sciences community has reduced the bias and imprecision of traditional literature summaries and their associated recommendations through the development of rigorous criteria for both literature overviews[1-3] and practice guidelines.[4,5] Even when recommendations come from such rigorous approaches, however, it is important to differentiate between those based on weak vs strong evidence. Recommendations based on inadequate evidence often require reversal when sufficient data become available,[6] while timely implementation of recommendations based on strong evidence can save lives.[6] In this article, we suggest an approach to classifying strength of recommendations. We direct our discussion primarily at clinicians who make treatment recommendations that they hope their colleagues will follow. However, we believe that any clinician who attends to such recommendations would benefit from the increased understanding they will gain through reading this article.

From the Departments of Medicine (Drs Guyatt, Hayward, and D. J. Cook) and Clinical Epidemiology and Biostatistics (Drs Guyatt, Hayward, D. J. Cook, and Sinclair), and the Department of Pediatrics (Dr Sinclair), McMaster University, McMaster University Faculty of Health Sciences, Hamilton, Ontario; the Centre for Evidence-Based Medicine, Nuffield Department of Medicine, University of Oxford (England) (Dr Sackett); and the Department of Statistics and Actuarial Sciences, Faculty of Mathematics, University of Waterloo (Ontario) (Dr R. J. Cook).

A complete list of members (with affiliations) appears in the first article of this series (*JAMA*. 1993;270:2093-2095). The following members contributed to this article: Eric Bass, MD, MPH; Hertzel Gerstein, MD, MSc; Brian Haynes, MD, MSc; Anne Holbrook, MD, PharmD, MSc; Roman Jaeschke, MD, MSc; Andreas Laupacis, MD, MSc; Virginia Moyer, MD, MPH; and Mark Wilson, MD, MPH.

Reprint requests to Room 2C12, McMaster University Health Sciences Centre, 1200 Main St W, Hamilton, Ontario, Canada L8N 3Z5 (Dr Guyatt).

### GRADING HEALTH CARE RECOMMENDATIONS: PREVIOUS CRITERIA

In 1979, the Canadian Task Force on the Periodic Health Examination made one of the first efforts to specify the strength of practice recommendations.[7] This group classified the quality of the evidence regarding the benefit of interventions into one of four categories based on the quality of the individual study designs. Their classification of the strength of their recommendations was considerably less explicit, only labeling evidence as "good," "fair," or "poor." The original Canadian Task Force approach, with minor modifications, has been reaffirmed by the Canadian Task Force[8] and endorsed by the US Preventive Services Task Force.[9] Both task forces contributed to progress in developing ways of grading the strength of health care recommendations that enhance both their interpretability and validity.

### ADVANCES IN METHODOLOGY

The classification system we present in this article is driven by four advances in translating evidence from original studies into clinical recommendations. First, methodologists have developed standardized approaches to the scientific conduct of literature reviews, and reviewers are increasingly using these approaches. This methodology includes systematic procedures and statistical techniques for combining results from different studies to minimize bias and increase precision.[10] Second, we have distinguished between clinical importance and statistical significance and realize that an intervention may be beneficial, but the effect too small to make the intervention worth administering.[11] The third advance is the more explicit acknowledgment that the strength of health care recommendations should depend on the precision of the estimated intervention effects: in general, the greater the sample size, the more precise our estimates of intervention effects, the narrower the confidence interval (CI) around our estimate of those effects, and the greater our ability to make strong recommendations. Finally, we are more aware that we may serve individual patients or groups of patients best if we withhold treatment for those at very low risk of clinical events while at the same time recommending treatment to those at higher risk.[12-15]

The Canadian and US Task Force criteria do not incorporate these advances. Members of our group have previously developed and modified criteria that addressed systematic overviews, but we failed either to clearly separate study design from the magnitude of the intervention effect, or to consider the impact of degree of patients' risk on treatment recommendations.[16,17] The approach we present in this article builds on the extensive work undertaken to date. We will focus on situations where investigations provide data regarding the effect of interventions on clinically important outcomes, whether the interventions are therapeutic, preventative, or diagnostic.

Our approach begins with the identification of a systematic overview of the existing evidence. By "systematic" we mean one that meets the following standards: the overview (1) addresses a focused clinical question; (2) uses appropriate criteria to select studies for inclusion; (3) conducts a comprehensive search; and (4) appraises the validity of the individual studies in a reproducible fashion. These standards are the same as those we recommend that clinicians use to identify an overview that is likely to yield an unbiased estimate of treatment effect.[18] Recommendations intended to influence clinical practice should be based on a current overview that meets these criteria.

### COMPONENTS OF THE APPROACH TO GRADES OF RECOMMENDATION

In our framework, making a recommendation about a health care interven-

Users' Guides to the Medical Literature section editor: Drummond Rennie, MD, Deputy Editor (West), *JAMA*.

Table 1.—Grades of Recommendations for a Specified Level of Baseline Risk*

| | |
|---|---|
| A1 | RCTs, no heterogeneity, CIs all on one side of threshold NNT |
| A2 | RCTs, no heterogeneity, CIs overlap threshold NNT |
| B1 | RCTs, heterogeneity, CIs all on one side of threshold NNT |
| B2 | RCTs, heterogeneity, CIs overlap threshold NNT |
| C1 | Observational studies, CIs all on one side of threshold NNT |
| C2 | Observational studies, CIs overlap threshold NNT |

*RCT indicates randomized controlled trial; CI, confidence interval; and NNT, number needed to treat to avoid one unwanted outcome.

Table 2.—Number Needed to Treat

| | Bleeding Risk if Untreated, U | Relative Risk Reduction, (U−T)/U | Bleeding Risk if Treated, T | Absolute Risk Reduction, U−T | No. Needed to Treat, 1/(U−T), to Prevent a Bleed |
|---|---|---|---|---|---|
| Critically ill patient receiving mechanical ventilation and/or has a coagulopathy | 0.037 | 58% | 0.0155 | 0.0215 | 45 |
| Critically ill patient breathing spontaneously without a coagulopathy | 0.0014 | 58% | 0.0006 | 0.0008 | 1250 |

tion requires the integration of three elements: the strength of the evidence presented in the overview; the threshold or magnitude of intervention effect at which benefit exceeds the risks of therapy, including both adverse effects and costs; and the relationships between the estimate of the magnitude of the intervention effect, the precision of that estimate, and the threshold. We will deal with each of these components in turn. In describing results of studies, we will consider the effect of the intervention on the clinical event that it is designed to prevent, which we will call the "target event." We will focus on the following: (1) the relative risk (RR), which is the ratio of the risk of target events in treated patients to the risk of target events in the untreated patients, and the RR reduction, or (1 − RR)[19]; (2) the absolute risk reduction, which is the difference in the absolute risk of the target event between treatment and control groups; and (3) the number needed to treat (NNT), which is the number of patients one needs to treat to prevent one target event (arithmetically, the inverse of the absolute risk reduction).[20]

## Component 1: The Strength of the Evidence

**Randomized Controlled Trials.**—Because no other study design can provide the safeguards against bias associated with randomization, randomized controlled trials (RCTs) yield stronger evidence than other study designs. Overviews of RCTs, therefore, provide far stronger evidence than do overviews of cohort and case-control studies. The strength of evidence from an otherwise systematic overview of RCTs will, however, depend on the consistency of the results from study to study. When different studies in the same overview yield very different estimates of treatment effect (a situation we refer to as "heterogeneity" of study results), one must question why. Possibilities include differences in patient populations, the way the interventions were administered, the way the outcomes were measured, the way the studies were conducted, or the play of chance.[21,22] A statistical test of the homogeneity of the intervention effect asks the question, "Are the differences in treatment effect from study to study greater than one would expect simply as a result of chance"?

If investigators conducting an overview conclude that treatment has a different effect depending on the population or the way the intervention is administered, they may conduct separate overviews for the different populations or treatments.[21,22] When differences in treatment effect across studies are greater than one would expect by chance alone, and varying populations, interventions, outcomes, or study methods cannot explain the differences, inferences become weaker. We therefore rank the strength of evidence from overviews of RCTs according to the presence or absence of unexplained differences in results from study to study (Table 1). We rank overviews with significant and important heterogeneity (level B) lower than those without significant and important heterogeneity (level A).

Before concluding that recommendations be classified as level B rather than level A, we should be confident that the degree of heterogeneity is clinically important. Heterogeneity can be considered clinically important if there is a large difference in RR reduction across studies. If the estimates from the individual studies are imprecise, however, an apparent large difference may be due to the play of chance. We propose the following criteria for clinically important heterogeneity:

1. The difference in the estimate of RR reduction between the two most disparate studies is greater than 20% (for instance an RR reduction of 40% in one study and less than 20% in another).

2. The difference between the boundaries of the CIs between the two most disparate studies is greater than 5% (for instance, the lower boundary [the smallest RR reduction compatible with the data] in the first study is 30% and the upper boundary of the CI [the largest RR reduction compatible with the data] in the second study is less than 25%).

Before heterogeneity bears on the strength of treatment recommendations, it must be both clinically important and statistically significant ($P<.05$).

**Observational Studies.**—Because the potential for bias is much greater in cohort and case-control studies than in RCTs, recommendations from overviews combining observational studies will be much weaker.[23,24] Thus, we classify observational studies as providing weaker evidence than RCTs (Table 1).

## Component 2: How Big an Impact of Treatment Warrants Its Use?

Any decision about initiating a preventive or therapeutic regimen represents a trade-off between patient or public benefits, on the one hand, and toxicity, cost, and administrative burden to patients and providers on the other. Clinicians do not, therefore, administer all effective treatments (effective in that they have a positive effect on some important outcome) to all potentially eligible patients. For example, $H_2$ receptor antagonists reduce the RR of serious bleeding in critically ill patients by approximately 58%.[25] However, a patient who is breathing spontaneously without a coagulopathy has a risk of serious bleeding of only 0.14% without treatment.[26] This baseline risk is so low that most clinicians would not consider it worth treating to lower the RR by another 58% (to 0.06%).

For administration of $H_2$ receptor antagonists to critically ill patients, and indeed for any treatment of any condition, it is useful to think of a threshold effect, above which one would treat and below which one would not. Moreover, it is informative to think of the number of patients one would need to treat to prevent a single serious gastrointestinal bleed.[27,28] Consider a group of critically ill patients who are receiving mechanical ventilation or who have a coagulopathy and whose risk of bleeding is therefore increased to 3.7%.[25,26] Treating such patients with $H_2$ receptor antagonists, one reduces their RR by 58%, to 1.55%. In absolute terms, their risk has fallen 2.15% (Table 2). The reciprocal of this absolute risk reduction is the NNT. In this case, 45 patients must receive prophylaxis to prevent an episode of serious bleeding.

Table 3.—How to Calculate the Threshold Number Needed to Treat

This table outlines how we calculate the threshold number needed to treat (NNT), a complete description of which will appear in an article we are preparing for publication. In describing how to calculate a threshold NNT, we will use the following notation:

T-NNT: the threshold number needed to treat

$Cost_{treatment}$: the cost of treating one patient

$Cost_{target}$: the cost of treating one target event

$Cost_{AE}$: the cost of treating one adverse event, with a further subscript 1 or 2 denoting the first and second adverse effects

$Rate_{AE}$: the proportion of treated patients who suffer an adverse event (again, subscripts 1 and 2 denoting the two adverse events)

$Value_{target}$: the dollar value we assign to preventing one target event

$Value_{AE}$: the dollar value we assign to preventing one adverse event (again, subscripts 1 and 2 denoting the two adverse events)

The general approach for generating the threshold NNT is based on the concept that at this threshold the value of treatment inputs equals the value of treatment outputs; that is, the net cost of treating the number of patients one needs to treat to prevent one patient having the target event equals the net value of the adverse events prevented or caused by treating that number of patients. The value of the treatment inputs includes the following:

the cost of treating the number of patients that will comprise the threshold NNT: $(Cost_{treatment})(T\text{-}NNT)$

*plus*

the cost of treating the adverse events attributable to treatment in the number of patients that will comprise the threshold NNT: $(Cost_{AE})(Rate_{AE})(T\text{-}NNT)$

*minus*

the cost of treating one target event: $Cost_{target}$

The value of the outputs includes the following:

the dollar value assigned to the one target event prevented: $Value_{target}$

*minus*

the dollar value assigned to adverse events attributable to treatment: $(Value_{AE})(Rate_{AE})(T\text{-}NNT)$

Thus, we have:

$$[(Cost_{treatment})(T\text{-}NNT)] + [(Cost_{AE})(Rate_{AE})(T\text{-}NNT)] - Cost_{target} = Value_{target} - [(Value_{AE})(Rate_{AE})(T\text{-}NNT)]$$

Rearranging:

$$T\text{-}NNT\,[Cost_{treatment} + (Cost_{AE})(Rate_{AE})] - Cost_{target} = Value_{target} - (T\text{-}NNT)[(Value_{AE})(Rate_{AE})]$$

And solving for threshold NNT:

$$T\text{-}NNT = (Cost_{target} + Value_{target})/[Cost_{treatment} + (Cost_{AE})(Rate_{AE})] + [(Value_{AE})(Rate_{AE})]$$

In the example we have used in the body of the article concerning the prevention of gastrointestinal bleeding, there are two adverse effects attributable to treatment that we must consider. The equation therefore becomes the following:

$$T\text{-}NNT = (Cost_{target} + Value_{target})/[Cost_{treatment} + (Cost_{AE1})(Rate_{AE1}) + (Cost_{AE2})(Rate_{AE2})] + [(Value_{AE1}) + (Rate_{AE1})\,(Value_{AE2})(Rate_{AE2})]$$

Substituting the figures from the body of the article yields the following:

$$T\text{-}NNT = (12\,000+3000)/[65+(10\,000)(0.0006)+(500)(0.015)]+[(3000)(0.0006)+(300)(0.015)]$$

$$T\text{-}NNT = 15\,000/[65 + 6 + 7.5] + [18 + 4.5]$$

$$T\text{-}NNT = 15\,000/101$$

$$T\text{-}NNT = 148.5$$

We believe that it is important to consider costs in deciding on the threshold NNT. Some clinicians may be uncomfortable with including costs. A model for calculating the threshold NNT that neglects costs would use the following formula:

$$T\text{-}NNT = 1/[(Value_{AE1})(Rate_{AE1}) + (Value_{AE2})(Rate_{AE2})]$$

In this equation the value of the adverse events is not the dollar value as in the model that includes costs, but the value of the adverse event in terms of the target event. That is, if we decided that the negative consequences of an adverse event was only one-tenth as great as the negative consequences of the target event, the value of that adverse event would be 0.1.

Consider again the first group of critically ill patients we've mentioned, those who are breathing spontaneously and who don't have a coagulopathy. Their risk of bleeding without treatment, which we call the "baseline" risk, is 0.14%, their risk with treatment is 0.06%, and one must treat 1250 such patients to prevent a serious bleed (Table 2).

Should we treat either, or both, of these patients? This decision involves generating a threshold NNT. If the patients' risk without treatment is high enough, and the NNT is below the threshold, we administer treatment. If the patient's risk without treatment is low enough, and the NNT is therefore above the threshold, we would not treat.

Generating the threshold NNT involves three steps. In the first step, we identify two sorts of undesirable events. One is the target event, and the other is the adverse effects attributable to treatment. To generate the threshold NNT, we must specify the costs we incur when we treat patients, the costs we save when we prevent the occurrence of the target event, costs that we might incur as a result of preventing the target event, and the costs we incur when we look after patients who suffer adverse events associated with treatment.

In considering the decision whether to administer prophylaxis for gastrointestinal bleeding, some of the costs we specify below are based on a detailed economic analysis from a hospital's point of view (D. Heyland, A. Gafni, D. Cook, G. H. Guyatt, unpublished data, 1995), while others are much more approximate estimates. In this case, the cost of administering ranitidine during a patient's 10-day stay in the intensive care unit (calculated, as are all our costs, based on Canadian data) is approximately $65 (including drug costs and costs of administering the treatment) and the cost of treating a gastrointestinal bleed is $12 000. Adverse effects of the $H_2$ receptor antagonist ranitidine include hepatitis with hepatic failure (an incidence of 0.06%,[29] with a treatment cost of $10 000 per episode) and central nervous system toxicity (an incidence of 1.5%,[30] with a cost of $500 per episode).

The second step in generating the threshold NNT is assigning relative values to the outcomes and relating them to dollar costs. These values may come from health workers, administrators, patients, or a large random sample of the general public and might use one of a number of approaches (such as individual interviews or a group consensus process) to assess utility.[31] While there is no consensus about either who should be deciding values or the best method of establishing that group's values, we would recommend individual interviews with either patients or the general public. Whatever population and approach to eliciting values one chooses, the process would involve (in this case) determining the degree of satisfaction, distress, or desirability that people associate with having an episode of gastrointestinal bleeding relative to an episode of liver toxicity or central nervous system toxicity. The process then involves deciding how much money should be allocated to prevent a single episode of gastrointestinal bleeding, which in turn sets the money we would be willing to spend to avoid the adverse events attributable to treatment.[32]

For purposes of the present discussion, we have not actually obtained values from a random sample of the population, but have guessed at what the population might say. In this case, we would be willing to spend $3000 to prevent one gastrointestinal bleed. We have equated one episode of liver toxicity and 10 episodes of central nervous system toxicity to a serious gastrointestinal bleed. Thus, we would be willing to spend $3000 to avoid one episode of liver toxicity and $300 to avoid one episode of central nervous system toxicity. We explain the algebra involved in calculating the threshold NNT in Table 3; as it turns out, the figures above generate a threshold NNT of approximately 150.

Figure 1 presents the relationship between the treatment NNT, the thresh-

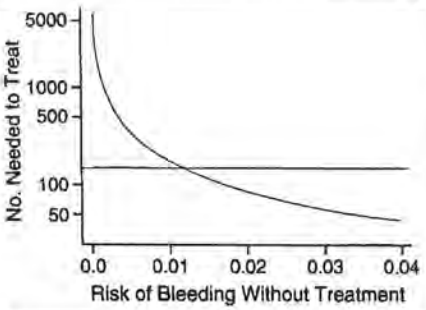



Figure 1.—Relationship between number needed to treat (NNT) associated with treatment, threshold NNT (horizontal line), and risk of bleeding without treatment for critically ill patients.

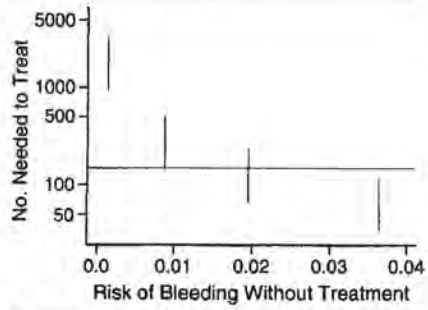



Figure 2.—Levels of baseline risk and threshold number needed to treat (NNT). Vertical lines represent the 95% confidence intervals around the treatment NNT at baseline risks of 0.14%, 0.9%, 2%, and 3.7%.

old NNT, and the risk of bleeding without treatment for critically ill patients. In constructing Figure 1, we have used the RR reduction we can expect with administration of $H_2$ receptor antagonists (58%), and the threshold NNT of 150 that we have generated. The horizontal line at an NNT of 150 represents this threshold NNT. The decreasing curve represents the NNT for any given risk of bleeding without treatment, which we will call the "treatment NNT line." Points on this line include the groups of patients from Table 2: patients with a risk of serious bleeding without treatment of 3.7%, for whom the NNT is 45, and patients with a risk of serious bleeding without treatment of 0.14%, for whom the NNT is 1250. The treatment NNT line crosses the threshold NNT at a risk without treatment of 1.15%. Therefore, our judgment is that treatment is warranted in patients whose risk of serious bleeding without treatment is greater than 1.15%, and not warranted for those whose risk is less than 1.15%.

The threshold NNT will vary depending on the values the clinician and patient place on its components. Some clinicians may be uncomfortable including costs as a consideration in the decision to treat. The strength of the threshold approach is that those recommending policy can, in generating a threshold NNT, make explicit the values they place on avoiding clinical events, adverse effects, and costs incurred or avoided, or omit costs from the consideration. In Table 3, we provide a method of calculating the threshold NNT without considering costs. Clinicians can examine the basis for the decision regarding threshold NNT, and the implications of differences in values, and the lower or higher threshold generated as a result of different values.

### Component 3: How Much Does the Treatment Work?

A meta-analysis is a quantitative overview that yields the best estimate of the treatment effect by pooling results from different trials. This estimate is called a "point estimate" to remind us that although the true value lies somewhere in its neighborhood, it is unlikely to be exactly correct. Confidence intervals tell us the range within which the true treatment effect likely lies.[33,34] We usually (though arbitrarily) use the 95% CI, which can be interpreted as defining the range that would include the true treatment effect 95% of the time on repetition of the experiment.

Given a specified risk of a clinical event without treatment, we can use the reduction in RR of clinical events with treatment and the CI around that reduction in RR, to calculate not only the NNT, but also the CI around the NNT. The relationship between that CI and the threshold NNT will have a profound effect on the strength of any recommendation to treat or not to treat. There are four possible relationships between the threshold NNT, the point estimate of the treatment effect, and the CI around the point estimate. We will examine each of these four in turn.

Consider critically ill patients who are receiving mechanical ventilation or have a coagulopathy. We have already decided that since their NNT lies below the threshold, they should be treated with $H_2$ receptor antagonists (or some equivalent treatment) (Table 2, Figure 1). We must remember, however, the upper boundary of the CI around the NNT. This boundary represents the smallest reduction in risk and thus the largest NNT, which is likely to be consistent with the data. In this case, the 95% CI around the RR reduction of 58% ranges from 79% to 21%, and the corresponding CI around the NNT, given the risk without treatment of 3.7%, ranges from 34 to 129. Here, the boundary of the CI that represents the highest NNT consistent with the data is still less than the threshold NNT of 150. We can be confident that the treatment for patients whose risk of bleeding is 3.7% does more good than harm, on average, given the relative values and costs we have specified.

Consider critically ill patients who are neither receiving mechanical ventilation nor have a coagulopathy and whose risk of bleeding is therefore 0.14%. Given the 58% RR reduction, we must treat 1250 such patients to prevent a bleed (Table 2). The 95% CI around this NNT ranges from 904 to 3401. The boundary of the CI that represents the largest plausible treatment effect, and thus the smallest NNT (904), is greater than the threshold NNT of 150. We can therefore be confident that the risks and costs of treatment outweigh the benefits.

If the risk of bleeding without treatment is intermediate, the recommendation is less clear. Take, for instance, a critically ill patient with a bleeding risk of 2%. Given an RR reduction of 58%, we must treat 86 such patients to prevent a bleed. Given the range of the 95% CI around the RR reduction (79% to 21%), the true NNT may lie between 63 and 238. The boundary of the 95% CI that represents the smallest plausible treatment effect, and thus the greatest NNT, 238, is greater than the threshold NNT. While the overall recommendation will still be to treat patients with this level of risk of bleeding, our strength of inferences will be weaker.

Similarly, if one considers a patient with a risk of serious bleeding without treatment of 0.9%, the most likely NNT is 192, but the 95% CI ranges from 141 to 529. Since the most likely NNT is above the threshold, the recommendation will be to withhold treatment, but because the 95% CI overlaps the threshold NNT of 150, the strength of inference is relatively weak.

We present results from all four levels of baseline risk (0.14%, 0.9%, 2%, and 3.7%) together with the threshold NNT in Figure 2.

## THE FINAL PRODUCT: RECOMMENDATIONS

If one combines the strength and heterogeneity of the primary studies with the magnitude and precision of the treatment effect as it relates to the threshold NNT, one can decide on the strength of the recommendation to treat or not to treat (Table 1). As we have demonstrated, the recommendation may change from "offer the intervention" when the baseline risk is high, to "don't offer the intervention" when the baseline risk is low. We believe that within RCTs, whether the CI on the NNT overlaps the threshold NNT is more impor-

tant than the presence of heterogeneity. However, RCT evidence is always stronger than evidence from observational studies. Thus, for any given baseline risk, A1 and B1 designate the strongest recommendations, A2 and B2 represent intermediate-strength recommendations, and C1 and C2 are the weakest recommendations.

## COMMENT

There are many issues in arriving at recommendations that remain to be fully explored. The .05 threshold for deciding whether or not heterogeneity is statistically significant, the proposed criteria for deciding whether heterogeneity is clinically important, and the choice of 95% for the CI around the treatment NNT are all arbitrary. Our choice of the 95% CI is based on tradition. Less stringent values would lead to narrower CIs (and thus more level 1 recommendations) and may ultimately be judged more appropriate.

The decision regarding the threshold NNT requires data both on costs and on the relative values we place on varying outcomes, data that will often not be available. Limitations in the data will emphasize the need to conduct additional rigorous studies. In the meanwhile, we must make treatment decisions and these decisions imply estimates of costs and values. Making these estimates explicit is worthwhile, even if we acknowledge their imprecision. We can examine the treatment implications of varying assumptions about costs and values (and thus, varying threshold NNTs). This emphasizes the absolute requirement to be explicit about what drives our decisions, particularly the underlying values.

The decision about the threshold NNT may vary in different practice settings and from patient to patient. We suggest that those making recommendations for clinical practice be explicit about how they arrive at their threshold NNT. They must consider all major toxicity, annoyance or inconvenience for the patient, the administrative burden on the health care system, and the cost of treatment, and describe how they have valued each component. If clinicians disagree with the values underlying a particular threshold NNT or work in a setting in which a particular threshold NNT does not apply, they can generate a new threshold NNT consistent with their values or practice setting. They could still use the overview evidence and the treatment NNT and quickly generate recommendations.

Our approach represents one in a series of steps along the road to optimal categorization of treatment recommendations and will likely require modification. Nevertheless, four elements of the approach presented here should help us move forward in the search for better ways of framing treatment recommendations. First, recommendations must be based on systematic overviews of methodologically sound primary studies. Second, those making recommendations must specify a threshold level of impact that warrants recommendation for applying the intervention. Third, recommendations will almost certainly vary when the magnitude of risk without treatment varies. Finally, recommendations must be based on two clearly separated components, the design and heterogeneity of the primary studies, on the one hand, and the magnitude and precision of the estimates of the treatment effects on the other. We hope that clinicians and policymakers find these insights useful in future development of treatment recommendations.

We would like to thank John Simes for his insightful comments, which helped us address a number of important conceptual issues. Kjell Apslund, Iain Chalmers, Peter Gotzsche, Chris Silagy, and Salim Yusuf all provided helpful comments on earlier drafts of the manuscript. We offer special thanks to Deborah Maddock who has provided outstanding administrative support and coordination for the activities of the Evidence-Based Medicine Working Group.

### References

1. Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analyses of randomized controlled trials. *N Engl J Med.* 1987;316:450-455.
2. Oxman AD, Cook DJ, Guyatt GH, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VI: how to use an overview. *JAMA.* 1994;272:1367-1371.
3. L'Abbe KA, Detsky AS, O'Rourke K. Meta-analysis in clinical research. *Ann Intern Med.* 1987;107: 224-233.
4. Hayward RSA, Wilson MC, Tunis SR, Bass EB, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VIII: how to use clinical practice guidelines, A: are the recommendations valid? *JAMA.* 1995;274:570-574.
5. Wilson MC, Hayward RSA, Tunis SR, Bass EB, Guyatt G, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VIII: how to use clinical practice guidelines, B: what are the recommendations and will they help you in caring for your patients? *JAMA.* 1995;274: 1630-1632.
6. Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts: treatments for myocardial infarction. *JAMA.* 1992;268:240-248.
7. Canadian Task Force on the Periodic Health Examination. The periodic health examination. *Can Med Assoc J.* 1979;121:1193-1254.
8. Woolf SH, Battista RN, Anderson GM, et al. Assessing the clinical effectiveness of preventative meneuvers: analytic principles and systematic methods in reviewing evidence and developing clinical practice recommendations. *J Clin Epidemiol.* 1990; 43:891-905.
9. US Preventive Services Task Force. Screening for adolescent idiopathic scoliosis: review article. *JAMA.* 1993;269:2667-2672.
10. Hedges L, Olkin I. *Statistical Methods for Meta-analysis.* New York, NY: Academic Press Inc; 1985.
11. Diamond GA, Denton TA. Alternative perspectives on the biased foundations of medical technology assessment. *Ann Intern Med.* 1993;118:455-464.
12. Jackson R, Barham P, Bills J, et al. Management of raised blood pressure in New Zealand: a discussion document. *BMJ.* 1993;307:107-110.
13. Smith GD, Egger M. Who benefits from medical interventions? *BMJ.* 1993;308:72-74.
14. Glasziou P, Irwig L. Generalizing the results of clinical trials. Presented at the Second Cochrane Colloquium; October 2, 1994; Hamilton, Ontario.
15. Lubsen J, Tijssen JGP. Large trials with simple protocols: indications and contraindications. *Control Clin Trials.* 1989;10:151S-160S.
16. Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. *Chest.* 1986;89(suppl 2):2S-3S.
17. Cook DJ, Guyatt GH, Laupacis A, Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. *Chest.* 1992;102 (suppl 4):305S-311S.
18. Oxman AD, Sackett DL, Guyatt GH, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VI: how to get started. *JAMA.* 1993;270:2093-2095.
19. Sinclair JC, Bracken MB. Clinically useful measures of effect in binary analyses of randomized trials. *J Clin Epidemiol.* 1994;47:881-889.
20. Jaeschke R, Guyatt GH, Shannon H, et al. Basic statistics for clinicians, III: assessing the effects of treatment: measures of association. *Can Med Assoc J.* 1995;152:351-357.
21. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA.* 1991;266:93-98.
22. Oxman AD, Guyatt GH. A consumer's guide to subgroup analysis. *Ann Intern Med.* 1992;116:78-84.
23. Sacks HS, Chalmers TC, Smith H Jr. Sensitivity and specificity of clinical trials: randomized v historical controls. *Arch Intern Med.* 1983;143:753-755.
24. Chalmers TC, Celano P, Sacks HS, Smith H Jr. Bias in treatment assignment in controlled clinical trials. *N Engl J Med.* 1983;309:1358-1361.
25. Cook DJ, Reeve BK, Guyatt GH, Griffith LE, Heyland DK, Tryba M. Stress ulcer prophylaxis in the critically ill: resolving discordant meta-analyses. *JAMA.* In press.
26. Cook DJ, Fuller HD, Guyatt GH, et al. Risk factors for gastrointestinal bleeding in critically ill patients. *N Engl J Med.* 1994;330:377-381.
27. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *N Engl J Med.* 1988;318: 1728-1733.
28. Guyatt GH, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, II: how to use an article about therapy or prevention, B: what were the results and will they help me in caring for my patients? *JAMA.* 1994;271:59-63.
29. Dobbs JH, Muir JG, Smith RN. H2-antagonists and hepatitis. *Ann Intern Med.* 1986;105:803.
30. Vial T, Goubier C, Begeret A, et al. Side effects of ranitidine. *Drug Saf.* 1991;6:94-117.
31. Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life: basic sciences review. *Ann Intern Med.* 1993;70:225-230.
32. Torrance GW. Measurement of health state utilities for economic appraisal. *J Health Econ.* 1986; 5:1-30.
33. Guyatt G, Jaeschke R, Cook DJ, Shannon H, Heddle N, Walter S. Basic statistics for clinicians, 2: interpreting study results: confidence intervals. *Can Med Assoc J.* 1995;152:169-173.
34. Altman DG, Gore SM, Gardner MJ, Pocock SJ. Statistical guidelines for contributors to medical journals. In: Gardner MJ, Altman DG, eds. *Statistics With Confidence: Confidence Intervals and Statistical Guidelines.* London, England: British Medical Journal; 1989:83-100.

family discussed your values and their need to know the extent of treatment you would find acceptable if seriously ill? Have you discussed this topic with your physician?"

This questionnaire can serve 2 functions, both toward the same goal. First, it can become a communications tool and stimulate an opening dialogue. This can lead to the development of a health care discussion and true decision partnership. Second, if the patient elects, it can become the basis of what I would like to call a "caring covenant" between patient and physician. If this caring covenant is used up front to define the understanding between physician and patient, it will help ensure the results the patient wants while at the same time relieving the physician of an enormous burden.

We patients still expect a great deal of physicians. We used to expect them to perform medical miracles—now we expect them to perform miracles of conscience. We need to take more charge of the process, for their sake as well as our own.

Mary S. Strong
Westwood, Mass

Ms Strong is chair emerita of American Health Decisions.

1. The SUPPORT Principal Investigators. A controlled trial to improve care for seriously ill hospitalized patients: the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT). *JAMA.* 1995;274:1591-1598. Correction: *JAMA.* 1996;275:1232.

### The Symbol of Medicine: Aesculapius or Caduceus?

*To the Editor.*—Since antiquity, the Aesculapian staff has served as a symbol of medicine and the art of healing. The emblem is depicted by a single serpent coiled around a rough, knotty staff. For various reasons, the symbol of medicine is often misrepresented as the caduceus, a winged staff with 2 intertwined serpents. Though similar in appearance and today often used interchangeably, the staffs have neither the same meaning nor the same origin.

The origin of the Aesculapian staff is subject to broad speculation. However, there are 3 accounts that suggest its origin. The first is from Greek mythology.[1] In this myth, Aesculapius, the Greek god of medicine, observes a serpent placing a magical herb in the mouth of a dead snake. The herb brings the snake back to life.

The second account, also from mythology, maintains that the god Aesculapius was imported to Rome (circa 295 BC) in the form of a serpent to cure a stubborn pestilence.[2] Ovid recounts the words of Aesculapius: "Fear not! I shall come and leave my images. Only be sure to note this snake that twines about my staff and mark it well to fix it in your mind. To this snake I shall change."[3] After successfully ridding the city of the pestilence, temples were built throughout the Roman Empire in honor of Aesculapius. Soon thereafter, the Aesculapian staff became associated with medicine.

The third account explains the origins of the emblem on purely medical grounds. In a procedure done to extract the parasitic guinea worm from its victim, the worm is removed by carefully winding it around a thin stick and pulling it from the tissue.[4]

Although different, each account suggests a medically based origin for the staff. In contrast, the caduceus is actually a hybrid between the staff of Hermes, messenger of the gods, and 2 serpents in the coital position symbolizing fertility. In early history, this emblem was a symbol of commerce and gainful trade.[2]

Confusion between the staffs is thought to have originated in the 16th century when printer Johannes Froben used the caduceus on the cover of various medical texts.[4] This mistaken association was compounded during the 19th century when the US Marine Hospital Service, the Public Health Service, and the US Army Medical Corps each adopted the caduceus as its symbol to designate a neutral, noncombatant status.[1]

Despite the deeply rooted confusion between the staffs, only the Aesculapian staff has consistently represented medicine. More importantly, the Aesculapian staff, because of its direct association to the Greek god of medicine, is historically valuable. It is an important link between medicine today and its origin. For these reasons, the Aesculapian staff should be used as the true emblem of medicine.

Rade Nicholas Pejic
Tulane University
New Orleans, La

1. Rakel RE. One snake or two? *JAMA.* 1985;253:2369.
2. Rutkow IM. *Surgery: An Illustrated History.* St Louis, Mo: Mosby–Year Book Co; 1993.
3. Ovid. *Metamorphoses Book XV.* New York, NY: Oxford University Press; 1986.
4. Schouten J. *The Rod and the Serpent of Asklepios.* New York, NY: Elsevier Science Publishing Co Inc; 1967.

## CORRECTIONS

**Correction in Authorship.**—In the Original Contribution entitled "A Controlled Trial to Improve Care for Seriously Ill Hospitalized Patients: The Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT)," published in the November 22/29, 1995, issue of THE JOURNAL (1995;274:1591-1598), the byline should read as follows: The Writing Group for the SUPPORT Investigators.

The Writing Group comprises the following members: Alfred F. Connors, Jr, MD, and Neal V. Dawson, MD, MetroHealth Medical Center, Cleveland, Ohio; Norman A. Desbiens, MD, Marshfield (Wis) Medical Research Foundation; William J. Fulkerson, Jr, Duke University Medical Center, Durham, NC; Lee Goldman, MD, MPH, Beth Israel Hospital, Boston, Mass; Frank E. Harrell, Jr, PhD, Duke University Medical Center, Durham, NC; William A. Knaus, MD, George Washington University Medical Center, Washington, DC; Joanne Lynn, MD, Dartmouth Medical School, Hanover, NH; Robert K. Oye, MD, University of California at Los Angeles Medical Center; Russell S. Phillips, MD, Beth Israel Hospital, Boston, Mass; Joan Teno, MD, Dartmouth Medical School, Hanover, NH; and Neil S. Wenger, MD, MPH, University of California at Los Angeles Medical Center.

**Incorrect Reference.**—In the Clinical Crossroads entitled "A 50-Year-Old Woman With Disabling Spinal Stenosis," published in the December 27, 1995, issue of THE JOURNAL (1995;274:1949-1954), an incorrect reference appeared. Reference 41 should read as follows: Haselkorn JK, Ciol MA, Rapp S, Elam K, Deyo RA. Epidural steroid injections in the management of sciatica. *Arch Phys Med Rehabil.* 1995;76:1037.

**Error in Equation.**—In The Medical Literature article entitled "Users' Guides to the Medical Literature, IX: A Method for Grading Health Care Recommendations," published in the December 13, 1995, issue of THE JOURNAL (1995;274:1800-1804), there is an error in an equation in Table 3 on page 1802. In the 30th line, the last part of the equation should read as follows:

$$[(\text{Value}_{AE1})(\text{Rate}_{AE1}) + (\text{Value}_{AE2})(\text{Rate}_{AE2})].$$

# Users' Guides to the Medical Literature

## X. How to Use an Article Reporting Variations in the Outcomes of Health Services

C. David Naylor, MD, DPhil; Gordon H. Guyatt, MD, MSc; for the Evidence-Based Medicine Working Group

### CASE SCENARIO

Your patient, a 78-year-old retired internist, has been complaining of increasing symptoms of benign prostatic hyperplasia. He has long-standing hypertension and coronary artery disease, with remote anterolateral myocardial infarction and bypass surgery 10 years ago. His left ventricular ejection fraction was recently documented at 20%, and he has been started on an angiotensin-converting enzyme inhibitor. Rectal examination confirms a moderately enlarged prostate, without irregularities, nodularity, or tenderness. As you discuss management options, your patient insists that transurethral prostate surgery is dangerous and that international studies of thousands of patients have proved that, as he puts it, "old-fashioned open prostatectomy is safer than that keyhole surgery." You prescribe a trial of an α-blocker, terazosin, and arrange to see him again. However, the retired internist sounds so convinced that you also resolve to look into the evidence about the two forms of prostatectomy.

### THE SEARCH

Later, you sit down in the hospital library, using a program that contains the MEDLINE database from January 1990 to October 1994. You start from "Explode Prostatic Hyperplasia," limit the search to English-language articles on human subjects, and then combine the resulting set with "transurethral" and "mortality" as text words. This yields 27 citations. Browsing through the resulting abstracts, two appear to address your patient's concern. One, by a Danish group,[1] addresses the long-term outcomes of transurethral vs "open" (suprapubic or transvesical) prostatectomy using hospitalization data linked to vital status data for the entire Danish male population from 1977 to 1985. The study relies on administrative data and massive population-based numbers (38 067 men) and shows excessive mortality among patients undergoing transurethral resection of the prostate (TURP). The other report, by Concato et al,[2] offers long-term outcomes data on only 252 patients who underwent either procedure at a Yale teaching hospital in New Haven, Conn, between 1979 and 1981. However, a detailed chart audit was undertaken, and the results suggested that patients undergoing the more extensive open procedure had lower long-term mortality because they were healthier at the outset.

### INTRODUCTION

Over the last decade, changes in health care delivery have broadened the range of groups interested in the outcomes of medical care. Concern with costs and with dramatic interregional or international differences in practice among clinicians and institutions have focused the attention of administrators and politicians on the interplay between the processes and outcomes of health services. The evolution of managed care has sharpened interest in measuring and managing the quality of care delivered by individual practitioners, hospitals, and other institutions.

Implicitly, the questions about quality of care and the best way of delivering health services are issues of optimal treatment. For example, once a patient's problem is identified, the primary care physician first determines what intervention, if any, should be undertaken, and may then face the quality-related issue of choosing a specialist or institu-

From the Institute for Clinical Evaluative Sciences, Ontario, North York, the Clinical Epidemiology and Health Care Research Program, Sunnybrook Unit, and the Departments of Medicine and Surgery, University of Toronto (Ontario) (Dr Naylor); and the Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario (Dr Guyatt).

A complete list of members (with affiliations) of the Evidence-Based Medicine Working Group appears in the first article of this series (*JAMA*. 1993;270:2093-2095). The following members contributed to this article: Eric Bass, MD, MPH; Hertzel Gerstein, MD, MSc; Daren Heyland, MD, MSc; Ann Holbrook, MD, PharmD, MSc; Virginia Moyer, MD, MPH; Tom Newman, MD; Andrew Oxman, MD, MSc; W. Scott Richardson, MD; Peter Tugwell, MD, MSc; and John Williams, Jr, MD, MHS.

Corresponding author: C. David Naylor, MD, DPhil, Institute for Clinical Evaluative Sciences in Ontario, G-106, Sunnybrook Health Science Centre, 2075 Bayview Ave, North York, Ontario, Canada M4N 3M5.

Reprint requests to Room 2C12, McMaster University Health Sciences Centre, 1200 Main St W, Hamilton, Ontario, Canada L8N 3Z5 (Dr Guyatt).

Users' Guides to the Medical Literature section editor: Drummond Rennie, MD, Deputy Editor (West), *JAMA*.

tion to offer that service. From a prior Users' Guide[3] you've learned that decisions about what treatment to provide are best made in light of evidence from randomized studies with complete follow-up. However, investigators are generally not going to be able to randomize patients to different practitioners or hospitals, and focusing on the outcomes associated with these differences in care will require strategies other than randomized trials. Increasingly, investigators have looked to large administrative or other observational databases to examine the outcomes of care associated with different procedures, practitioners, or institutions. Under what circumstances should you believe the inferences made on the basis of such studies?

There is a parallel here with studies assessing potential harm to patients: it is impossible to randomize people to smoke or not, or to various levels of air pollution, and so observational studies or "natural experiments" are used as sources of insight. In a previous Users' Guide[4] we provided criteria for validity for the observational studies that investigators must use when exploring issues of harm. The challenges are fundamentally the same for comparing outcomes of two or more sets of health care practitioners or delivery systems. However, observational studies using administrative databases are growing in scope and importance and have their own particular challenges. Therefore, we devote this Users' Guide to these issues. Table 1 revisits our criteria for assessing an article about harm, modified here for examining associations between variations in processes and outcomes of health care in the real-world setting.

Table 1.—Three Core Questions to Ask About a Study Using an Observational Design to Examine Sources of Difference in Patients' Outcomes

| |
|---|
| Are the outcome measures accurate and comprehensive? |
| Were there clearly identified, sensible comparison groups? |
| Were the comparison groups similar with respect to important determinants of outcome, other than the one of interest? |

## ARE THE OUTCOME MEASURES ACCURATE AND COMPREHENSIVE?

A randomized therapeutic trial must have valid and reliable outcome measures; so must any observational study assessing patients' outcomes. The easiest outcomes for health researchers to measure are those that are defined objectively and usually captured in large insurance databases or computerized hospital administrative data, eg, death, in-hospital complications of surgery that are routinely coded, or readmissions to the hospital. Linkage to vital status registries is also performed to track out-of-hospital deaths. However, other outcomes, eg, disability, discomfort, distress, and dissatisfaction[5] are important to patients. Functional status and quality-of-life measures are needed to capture these burdens, but these measures are not applied in routine clinical care, and if applied, their results are not incorporated into administrative databases. Incorporating these measures into routine care and administrative databases, moreover, may generate more questions than answers. Researchers have begun to understand some of the factors that predict, for example, increased risk of mortality after various types of elective surgery. However, there is no similar understanding of the factors that predict functional status and quality of life.

In sum, many large databases are not designed for clinical research and may either mismeasure patients' outcomes or fail to capture outcomes that are important to patients and their physicians. Researchers should therefore report on the quality and comprehensiveness of the data source. Ideally there should be independent cross-checks to ensure that the same outcomes are measured consistently and completely for whatever unit of comparison is used, eg, verifying that data on ascertainment or cause of death are accurate or confirming hospital readmission rates after a specific surgical procedure in a quality-of-care study.

How did our two studies of prostate surgery perform in these respects? Andersen et al[1] used vital status data for the entire population of Denmark, and therefore mortality was measured in a reliable and unbiased fashion across all groups for comparison. Concato et al[2] reported on all-cause mortality data within 5 years of the procedure obtained by hospital chart review and, where those data were inconclusive, from the national vital status registry.

The complete resection attained by open prostatectomy obviously eliminates the need for repeat procedures as occasionally occurs with TURPs. However, neither study compared the two procedures with respect to various outcomes of interest to patients and physicians, eg, effectiveness in relieving obstructive or irritative symptoms of benign prostatic hyperplasia, overall recovery time, rates of complications such as impotence or incontinence, and so forth. Careful prospective data collection is necessary to capture these outcomes and provide a more complete tally of the burdens and benefits of the two treatments being compared. Even with those data, moreover, there would be uncertainty about the weights that patients would give to diverse benefits and harms, and a major challenge in determining how different outcomes related to each other and to patients' pretreatment characteristics.

Table 2.—Factors That May Systematically Affect Outcomes

| |
|---|
| What service was provided*<br>For example, variations among two or more management strategies with respect to use of drugs, doses, devices, type of procedure, and the like |
| Who provided the service<br>For example, variations among procedural specialists; nurse practitioners vs family physicians; by level of experience (house staff vs qualified specialists); by volume of service delivered (high-caseload vs low-caseload practitioners) |
| Where the service was provided<br>For example, variations among hospitals or clinics; between wards in a hospital; between a step-down unit and a conventional intensive care unit; home vs hospital care; by city; by county; by region or nation |
| When the service was provided<br>For example, variations in timing of service (eg, day or evening, weekend vs weeks, the July phenomenon for house staff effects); according to length of stay in hospital; across months (seasonal effects) or years (broad temporal trends) |

*These questions are best addressed using randomized trial methods; see Guyatt et al.[9]

## WERE THE COMPARISON GROUPS SIMILAR WITH RESPECT TO IMPORTANT DETERMINANTS OF OUTCOME OTHER THAN THE ONE OF INTEREST, AND WERE RESIDUAL DIFFERENCES ADJUSTED FOR IN THE ANALYSIS?

Clinicians and health care managers are interested in a variety of determinants of outcome, the major categories of which are shown in Table 2. One type of comparison examines differences that may be due to variations in quality of care across individual practitioners or institutions providing care in a specific city or region. State agencies now publish some provider- or institution-specific outcomes, and researchers sometimes relate these outcomes to the provider- or institution-specific volume of the services under scrutiny. This reflects a belief that "practice makes perfect"—all things being equal, centers (and by, inference, physicians or surgeons) with a higher caseload will generally achieve better outcomes than lower-volume centers. For example, various studies suggest that in-hospital postoperative mortality after aortic aneurysm surgery,[6] percutaneous transluminal coronary angioplasty,[7] and coronary artery bypass graft surgery[8,9] is lower for centers or surgeons managing more patients. On the other hand, large tertiary care centers often treat the sickest patients and therefore may have worse outcomes than smaller hospitals.

However, the greater the difference between service settings being compared, the more difficult it is to be sure that patients were similar, or to isolate

Table 3.—Determining Whether Differences in Prognosis, Rather Than Differences in the Intervention, Explain Differences in Outcomes

| |
|---|
| Were all important prognostic factors measured? |
| Were measures of patients' prognostic factors reproducible and accurate? |
| To what extent were patients similar with respect to these factors? |
| Was multivariate analysis used to adjust for imbalances in prognostic factors? |
| Did additional analyses (particularly in low-risk subgroups) demonstrate the same results as the primary analysis? |

which aspects, if any, of the process of care relate to the outcomes observed. This is especially true when comparisons are made on a broad geographic footing between regions or countries in which populations and processes of care differ in many ways. One recent study compared outcomes of Canadian and American patients enrolled in a major trial of thrombolytic therapy for acute myocardial infarction.[10] Rates of revascularization and use of specialist services were much higher in the United States. The investigators used an appropriately broad range of outcomes measures and observed that in terms of symptoms, functional status, psychological well-being, and health-related quality of life, Canadian patients fared somewhat worse than their American counterparts—a finding of obvious concern to Canadian practitioners. However, some of the difference may be because the types of patients recruited by Canadian investigators were destined for worse outcomes irrespective of management. Canadians may also have a different cultural threshold for reporting symptoms or functional impairment.

A third source of variations in outcomes that may occur within similar health systems is the type of treatment provided. This is the sort of comparison that was done in the outcomes studies of TURP vs open prostatectomy described in this article's opening scenario. Such comparisons may avoid some of the broad health system effects and sociocultural or even genetic differences that threaten the validity of outcomes comparisons made across widely disparate populations. However, it is still possible that differences in outcomes may have been due to differences in patients receiving the alternative management strategies, for without randomization, patients will inevitably differ in ways other than the treatment being provided to them. This phenomenon is called "selection bias." When two alternative procedures are being compared in research, selection bias arises from the exercise of good clinical judgment in routine practice. For example, urologists may choose younger, healthier patients to undergo the more extensive open prostatectomy, and older, sicker patients for TURP. Patients then end up differing in obvious or subtle ways that affect their likelihood of having a good or bad outcome. Epidemiologists use the term "confounding" to describe this problem. The validity of any form of observational research is threatened by case selection biases that create noncomparable groups of patients and confound any outcomes comparisons.

Researchers must therefore somehow adjust for differences between groups of patients. The sophistication of these so-called risk adjustment methods is growing rapidly.[11] However, researchers and quality-of-care evaluators are unlikely to know all the prognostic factors that interact with treatments to affect outcomes. Randomization is important precisely because it distributes these unknown factors in an unbiased manner. The problem worsens when one considers that all known prognostic features may not have been measured, and if they have been measured, they may not have been measured or recorded accurately. Inaccurate measurement or recording is a particular concern when information comes from administrative databases. For instance, Jollis et al[12] compared information about cardiac risk factors in an administrative database in patients undergoing angiography with information collected prospectively for a clinical database by a cardiac fellow who actually saw the patients. A chance-corrected measure of agreement ($\kappa$ statistic) showed good agreement only for diabetes (83% agreement) and whether patients had an acute myocardial infarction (76%); agreement was moderate for hypertension (56%), poor for the presence of heart failure (39%), and no better than chance (9%) for unstable angina. Hannan et al[13] found similar discrepancies in comparing a cardiac surgery registry with an administrative database in New York State. These inaccuracies mattered: the ability of evaluators to predict mortality was clearly higher with the detailed clinical data as opposed to the administrative database.[13] Thus, the accuracy, reproducibility, and fairness of adjustments for differences in patients can be undermined by poor data quality.

The problem of limited or inaccurate data in insurance databases or computerized hospital discharge abstracts may be partly ameliorated by supplementing the information with chart audits.[14] This is time-consuming and expensive, but may be the only way to reduce the chances of missing or misconstruing important differences among groups of patients. A more efficient mechanism may be to establish specific registry mechanisms geared to measuring key patient characteristics, process of care elements, and relevant outcomes.

How, then, can you best assure yourself that, short of randomization, investigators have made the fairest possible outcomes comparison possible? We summarize the steps in Table 3. First, did the researchers convince you, through their review of the literature and on the basis of what you know about the determinants of prognosis, that they measured all of the important prognostic factors? This is more likely to occur if the analysis involves chart audits or, better still, a specific clinical registry, as opposed to reliance on available administrative data. Second, since these measurements are only as good as the data that go into them, you should consider whether these measures of patients' prognostic factors are reproducible and accurate. Third, did the researchers show the extent to which the groups being compared differed on the prognostic factors that they measured? Fourth, did they use some form of multivariate analysis wherein they tried to adjust simultaneously not only for the obvious prognostic factors, but also for other more subtle differences that may have confounded the comparisons?

Localio and colleagues[15] have recently reported on the consequences of not taking into account all possible prognostic factors. A large corporation's managed care program sought to determine which of the hospitals serving the corporation's employees delivered better quality of care as reflected in part by fewer in-hospital deaths. A consultant concluded that the hospitals differed, and this conclusion influenced the company's choices about hospital selection. As it turned out, an appropriate analysis conducted by a group of academic investigators concluded that the difference between even the hospital with the worst record and the rest could be easily attributable to the play of chance. Furthermore, when the investigators included an adjustment for age, a prognostic factor that had been left out of the consultant's initial analysis, the rank order of the hospitals changed.[16]

Because observational data are so susceptible to selection biases that may confound the outcome comparisons, the researchers should determine whether their results persist when they analyze the data in different ways. For example, if there is a severe imbalance in allocation of patients with a particularly important prognostic factor, it may make sense to eliminate all patients with that factor and repeat the analyses. Unfortunately, even relative balance on a prognostic factor does not guarantee comparability. One reason is that administrative data and

registries tend to use fairly simple categories, such as whether a disease is or is not present. Yet, the category "disease present" may be associated with a wide range of underlying dysfunction, and therefore equally variable prognosis. Patients with chronic lung disease or chronic heart failure, for instance, can vary from mild to severe, with very different prognostic implications. Thus, apparent balance on the proportion of patients with these diagnoses can mask a situation in which one group has many more severely affected patients than the other. This is even true for advanced age as a prognostic factor, since elderly persons may vary considerably in their overall robustness.

Because of this problem, a useful double-check in any outcomes comparison is to ensure that the findings are replicable within a relatively low-risk subgroup of the patients being examined. By eliminating patients in categories associated with widely varying physiological states, we increase the likelihood of a "level playing field" for comparisons.

How do our two studies of prostate surgery measure up in this regard? Andersen et al[1] considered patients' ages at surgery, but relied only on diagnoses coded in the computerized hospital records as indicating compromised health status. Even with these limited data, fewer open prostatectomy patients had high-risk diagnoses. They were also younger and had less heart disease and cancer. In a multivariate analysis to try to adjust for these differences, it did appear that TURP continued to confer a 30% to 40% relative increase in the risk of death over several years of follow-up. Extensive sensitivity analyses were performed, including a specific examination of low-risk patients (described as "healthiest men"). Although low-risk patients also showed an excess risk with TURP, the relative magnitude of the increased risk of death was smaller for low-risk patients than for high-risk patients. As Andersen et al[1] stated: "The extent to which this difference is attributable to the surgical intervention itself remains an open question. The two groups of patients are quite different with regard to age and preoperative health status, and available data may not be sufficient to control such differences through statistical analysis."

Concato et al[2] used chart review methods with a detailed and systematic abstraction of information related to health status based on inpatient and ambulatory care records. They carefully confirmed that two reviewers independently agreed on patients' health status assessments. Patients in the TURP group were again found to be older and sicker. However, in a multivariate analysis, the adjusted excess risk of TURP diminished as the degree of detail on comorbidity was increased. Their best estimate was that TURP actually conferred no increased risk relative to open prostatectomy. Unfortunately, owing to the small sample size, their results were very imprecise, with 95% confidence limits ranging from much increased to much reduced risk with TURP (eg, from 0.57 to 1.87). Thus, the Yale study highlights the issue of noncomparability and selection biases, but does not rule out harms of the magnitude demonstrated by the Danish investigators. Moreover, the study provides data on outcomes for only a single city; the results may not be generalizable.

## CONCLUSIONS AND RESOLUTION

Given the limitations of observational studies of large databases, can we better define the role of this sort of health services research? Observational studies do remain important in the generation of hypotheses about causal pathways from a pathophysiological standpoint. Moreover, once randomized trials have helped define what treatments are likely to work best for your patients, observational outcomes studies generate information about what happens when these practices are used in the real world as opposed to the selected populations of patients and practitioners participating in randomized trials. This information deepens our understanding of practical effectiveness as opposed to theoretical efficacy, and may add new insights since trials do not always measure all the outcomes of interest to patients and physicians.

However, this complementary or supplementary role of large-scale observational studies departs sharply from using administrative data or clinical registries to decide which specific management strategies will yield better outcomes: eg, surgery vs medical, invasive vs noninvasive, different surgical procedures, and so on. To determine the relative merits of treatments, randomized trials are usually possible and preferable given the unavoidable biases of observational studies.

Do observational studies have any role at all in choosing best practices? Randomized trials are expensive and difficult to conduct and cannot be undertaken for all the clinical questions in which practitioners are interested. Observational studies may identify situations in which one therapy appears so much better than an alternative that bias would be a very unlikely explanation for the difference. As well, the hypothesis-generating role of observational studies is illustrated by the example of open prostatectomy. (Unfortunately, the convenience of transurethral surgery, together with deeply held beliefs about its safety, probably precludes ever mounting a large-scale trial comparing transurethral and open prostatectomy.) Finally, if the outcomes of interest are very rare, such as unusual idiosyncratic side effects of a drug, researchers can only obtain adequate sample sizes through use of administrative databases.

There are other situations in which randomization is not feasible, such as looking for systematic variations in outcomes of similar procedures provided by different practitioners or institutions ("who" or "where" rather than "what"; see Table 2). It is untenable to assume that all hospitals or providers practice equally well and observational outcomes comparisons have a role in assessing quality of care. This is especially applicable for some well-defined services (eg, coronary artery bypass grafting) where there are validated risk-adjustment algorithms[17-20] and dedicated registries to measure risk factors and outcomes, so that these comparisons are probably meaningful. In general, however, potential harm to patients from poor quality care must be weighed against the harm to skilled health workers and fine institutions caused by poorly founded inferences about inferior outcomes.

Given the relatively weak inferences possible from most observational studies of outcomes, alternative strategies for ensuring the quality of medical care should always be considered. For some processes of care (though certainly not all, as we caution in the next article in this series), we can accurately document what went on and make confident judgments about its appropriateness. For example, randomized trials show that preoperative antibiotic and antithrombotic prophylaxis improves patients' outcomes after various surgical procedures. Systematically omitting these treatments puts patients at risk and indicates a need for practitioners and institutions to improve their quality of care. We suggest that in most instances it is most efficient to use randomized trials or meta-analyses of trials to establish optimal management strategies, and then assess if quality of care is maintained by monitoring the process of care to ensure that well-proven practices are consistently applied to eligible patients.

What, then, of your patient? Perhaps predictably, given what we know about the limitations of observational studies, your exploration has been inconclusive. Indeed, had you used MEDLINE on CD-ROM for the years prior to 1990, the relevant literature would not have

moved you much further. Related work[21,22] on increased mortality after TURP as opposed to open prostatectomy has incorporated extra detail on differences among patients drawn from chart reviews and failed to eliminate the excess mortality seen with TURP; however, the adjustments were arguably less detailed than those used by Concato et al.[2] One very small randomized trial has also shown a trend to excess mortality with TURP.[23] On the other hand, there has been no definitive trial comparing the two forms of surgery and TURP remains the predominant procedure for benign prostatic hyperplasia.

The retired internist returns in 4 weeks as planned. "Was I right about the risks of the keyhole method?" he asks. You admit that the abandonment of open prostatectomy may have been premature, but caution that his age and medical status make him a poor candidate for the more extensive procedure, even if you could find a urologist competent to do it. Hearing your own advice, you again appreciate that similar selection biases may be the real reasons for the apparently higher mortality after TURP. Fortunately, your patient has had an excellent response to the α-blocker and the issue of prostatectomy can be set aside for some time. As you usher him from the office, he grumbles: "By the way, did you see that the operative mortalities for all the local heart surgeons are on the front page of the newspaper? Thank heavens I retired."

**References**


1. Andersen TF, Bronnum-Hansen H, Sejr T, Roepstroff C. Elevated mortality following transurethral resection of the prostate for benign hypertrophy! but why? *Med Care.* 1990;28:870-881.
2. Concato J, Horwitz RI, Feinstein AR, Elmore JG, Schiff SF. Problems of comorbidity in mortality after prostatectomy. *JAMA.* 1992;267:1077-1082.
3. Guyatt GH, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, II: how to use an article about therapy or prevention, A: are the results valid? *JAMA.* 1993;270:2598-2601.
4. Levine M, Walter S, Lee H, Haines T, Holbrook A, Moyer V, for the Evidence-Based Medicine Working Group. Users' guide to the medical literature, IV: how to use an article about harm. *JAMA.* 1994; 271:1615-1619.
5. White K. Improved medical statistics and health services systems. *Pub Health Rep.* 1967;82:847-854.
6. Hannan EL, Kilburn H Jr, O'Donnell JF, et al. A longitudinal analysis of the relationship between in-hospital mortality in New York State and the volume of abdominal aortic aneurysm surgeries performed. *Health Serv Res.* 1992;27:517-542.
7. Jollis JG, Peterson ED, DeLong ER, et al. The relation between the volume of coronary angioplasty procedures at hospitals treating Medicare beneficiaries and short-term mortality. *N Engl J Med.* 1994;331:1625-1629.
8. Showstack JA, Rosenfeld KE, Garnick DW, Luft HS, Schaffarzick RW, Fowles J. Association of volume with outcome of coronary artery bypass graft surgery: scheduled vs nonscheduled operations. *JAMA.* 1987;257:785-789.
9. Hannan EL, Kilburn H Jr, Bernard H, O'Donnell JF, Lukacik G, Shields EP. Coronary artery bypass graft surgery: the relationship between in-hospital mortality rate and surgical volume after controlling for clinical risk factors. *Med Care.* 1991;29:1094-1107.
10. Mark DB, Naylor CD, Hlatky MA, et al. Use of medical resources and quality of life after acute myocardial infarction in Canada and the United States. *N Engl J Med.* 1994;331:1130-1135.
11. Daley J, Shwartz M. Developing risk-adjustment methods. In: Iezzoni LI, ed. *Risk Adjustment for Measuring Health Care Outcomes.* Ann Arbor, Mich: Health Administration Press; 1994:199-238.
12. Jollis JG, Ancukiewicz M, DeLong ER, Pryor DB, Muhlbaier LH, Mark DB. Discordance of databases designed for claims payment versus clinical information systems: implications for outcomes research. *Ann Intern Med.* 1993;119:844-850.
13. Hannan EL, Kilburn H Jr, Lindsey ML, Lewis R. Clinical versus administrative data bases for CABG surgery: does it matter? *Med Care.* 1992; 30:892-907.
14. Malenka DJ, McLerran D, Roos N, Fisher ES, Wennberg JE. Using administrative data to describe casemix: a comparison with the medical record. *J Clin Epidemiol.* 1994;47:1027-1032.
15. Localio AR, Hamory BH, Sharp TJ, Weaver SL, TenHave TR, Landis JR. Comparing hospital mortality in adult patients with pneumonia. *Ann Intern Med.* 1995;122:125-132.
16. Wu AW. The measure and mismeasure of hospital quality: appropriate risk-adjustment methods in comparing hospitals. *Ann Intern Med.* 1995;122: 149-150.
17. Tu JV, Jaglal SB, Naylor CD, and the Steering Committee of the Provincial Adult Cardiac Care Network. Multicenter validation of a risk index for mortality, intensive care unit stay, and overall hospital length of stay after cardiac surgery. *Circulation.* 1995;91:677-684.
18. O'Connor GT, Plume SK, Olmstead EM, et al, for the Northern New England Cardiovascular Disease Study Group. Multivariate predictors of in-hospital mortality associated with coronary artery bypass graft surgery. *Circulation.* 1992;85:2110-2118.
19. Higgins TL, Estafanous FG, Loop FD, Beck GJ, Blum JM, Paranandi L. Stratification of morbidity and mortality outcome by preoperative risk factors in coronary artery bypass patients: a clinical severity score. *JAMA.* 1992;267:2344-2348.
20. Edwards FH, Clark RE, Schwartz M. Coronary artery bypass grafting: the Society of Thoracic Surgeons National Database experience. *Ann Thorac Surg.* 1994;57:12-19.
21. Roos NP, Wennberg JE, Malenka DJ, et al. Mortality and reoperation after open and transurethral resection of the prostate for benign prostatic hyperplasia. *N Engl J Med.* 1989;320:1120-1124.
22. Malenka DJ, Roos N, Fisher ES, et al. Further study of the increased mortality following transurethral prostatectomy: a chart-based analysis. *J Urol.* 1990;144:224-228.
23. Meyhoff H-H. Transurethral versus transvesical prostatectomy: clinical, urodynamic, renographic and economic aspects: a randomized study. *Scand J Urol Nephrol.* 1987;4(suppl 102):1-26.

# Users' Guides to the Medical Literature

## XII. How to Use Articles About Health-Related Quality of Life

Gordon H. Guyatt, MD, MSc; C. David Naylor, MD, MSc, DPhil; Elizabeth Juniper, MCSP, MSc; Daren K. Heyland, MD; Roman Jaeschke, MD, MSc; Deborah J. Cook, MD, MSc; for the Evidence-Based Medicine Working Group

### CLINICAL SCENARIO

You are a physician following a 35-year-old man who has had active Crohn disease for 8 years. The symptoms were severe enough to require resectional surgery 4 years ago, and despite treatment with sulfasalazine and metronidazole, the patient has had active disease requiring oral steroids for the last 2 years. Repeated attempts to decrease the prednisone have failed, and the patient has required doses of greater than 15 mg per day to control symptoms. You are impressed by both the methods and results of a recent article[1] documenting that such patients benefit from oral methotrexate and suggest to the patient that he consider this medication. When you explain some of the risks of methotrexate, particularly potential liver toxicity, the patient is hesitant. How much better, he asks, am I likely to feel while taking this medication?

### INTRODUCTION

There are 3 reasons we offer treatment to our patients. We believe our interventions increase longevity, prevent future morbidity, or make patients feel better. The first 2 of these 3 end points are relatively easy to measure. At least in part because of difficulty in measurement, clinicians have for many years been ready to substitute physiological or laboratory tests for the direct measurement of the third. In the last 20 years, however, clinicians have recognized the importance of direct measurement of how people are feeling and how they are able to function in daily activities. Investigators have developed increasingly sophisticated methods of making these measurements.

Since, as clinicians, we are most interested in aspects of life quality directly related to health rather than issues such as finances or the quality of the environment, we frequently refer to measurements of how people are feeling as health-related quality of life (HRQL).[2] Investigators measure HRQL using questionnaires that typically include questions about how patients are feeling or what they are experiencing associated with response options such as yes or no, 7-point scales, or visual analogue scales. Investigators aggregate responses to these questions into domains or dimensions (such as physical or emotional function) that yield an overall score.

Controversy exists concerning the boundaries of HRQL and the extent to which individual patient's values must be included in its measurement.[3-5] Is it sufficient to know that patients with chronic obstructive lung disease in general value being able to climb stairs without getting short of breath, or does one need to establish that the individual patient values climbing stairs without dyspnea? Further controversy exists about how the relative values of items and domains need to be established and how these values should be determined. Is it enough to know that both dyspnea and fatigue are important to people with lung disease, or does one need to establish their relative importance? If establishing their relative importance is necessary, which of the many available approaches should one use?

In this article, we take a simple approach. We use HRQL to refer to the health aspects of their lives that people, in general, value, and we are ready to accept patients' statement of what they value without precise determination of ranking of items or domains.

Clinicians often have limited familiarity with methods of measuring how patients feel. At the same time, they are facing articles that recommend administering or withholding treatment on the basis of its impact on patients' well-being. This Users' Guide is designed for clinicians asking the question: Will this treatment make my patient feel better? As in other guides, we will use the framework of the validity of the methods, interpretation of the results, and application of the results to one's patients (Table). In addition, we begin the guide with a commentary on when one should and should not be concerned about HRQL measurement. Our guidelines borrow heavily from our previous work.[2,5] While this article focuses on using HRQL measures to help with treatment decisions, we hope that it may also improve clinical care by

From the Departments of Clinical Epidemiology and Biostatistics (Drs Guyatt and Cook and Ms Juniper) and Medicine (Drs Guyatt, Heyland, Jaeschke, and Cook), McMaster University, Hamilton, Ontario, and the Institute for Clinical Evaluative Sciences and the Clinical Epidemiology Unit, Sunnybrook Health Science Centre, University of Toronto, Toronto, Ontario (Dr Naylor).

The original list of members (with affiliations) appears in the first article of this series (*JAMA*. 1993;270:2093-2095). A list of new members appears in the 10th article of the series (*JAMA*. 1996;275:1435-1439). The following members contributed to this article: Paul Glasziou, MB, PhD; Virginia Moyer, MD, MPH; and Peter Tugwell, MD, MSc.

Reprints: Gordon H. Guyatt, MD, MSc, McMaster University HealthSciences Centre, 1200 Main St W, Room 2C12, Hamilton, Ontario, Canada L8N 3Z5.

Users' Guides to the Medical Literature section editor: Drummond Rennie, MD, Deputy Editor (West), *JAMA*.

Guidelines for How to Use Articles About Health-Related Quality of Life (HRQL)

**Are the results valid?**
- Primary guides
  - Have the investigators measured aspects of patients' lives that patients consider important?
  - Did the HRQL instruments work in the way they are supposed to?
- Secondary guides
  - Are there important aspects of HRQL that have been omitted?
  - If there were trade-offs between quality and quantity of life, or an economic evaluation, have the investigators used the right measures?

**What were the results?**
- What was the magnitude of effect on HRQL?

**Will the results help me in caring for my patients?**
- Will the information from the study help me inform my patients?
- Did the study design simulate clinical practice?

emphasizing aspects of patients' experience, including functional, emotional, and social limitations, which clinicians sometimes neglect.

## DO YOU NEED TO WORRY ABOUT HRQL?

In the early days of clinical trials, few if any treatment studies included measurements of HRQL, and no one worried much. When should you be concerned if investigators have not paid adequate attention to how patients feel?

In general, delaying mortality is sufficient reason to administer a treatment. Some years ago, investigators showed that around-the-clock oxygen therapy for patients with severe chronic airflow limitation improved mortality.[6] The fact that HRQL data weren't reported in the original article turns out not to be an important omission. Since the intervention prolongs life, our enthusiasm for continuous oxygen administration is not blunted by a subsequent report suggesting that more intensive oxygen therapy had little or no impact on HRQL.[7] Similarly, while feeling better is important to patients with heart failure, when interventions either extend[8] or shorten[9] life span, we usually do not need an HRQL assessment to inform our clinical decisions.

There are exceptions to this rule. While many of our life-prolonging treatments have a negligible impact on or actually improve HRQL, this is not always the case. If treatment leads to a deterioration in HRQL, patients may be concerned that small gains in life span come at too high a cost. Interventions that highlight this concern include chemotherapy for cancer and human immunodeficiency virus disease. In the extreme, life may be prolonged, but patients' families may wonder if, for example, their fate is a persistent vegetative state, they are not better off dead. A patient's own preferences expressed through an advance directive may support this view.

When the goal of treatment is to improve how people are feeling (rather than to prolong their lives) and physiological correlates of patients' experience are lacking, HRQL measurement is imperative. For example, we would pay little attention to studies of antidepressants that failed to measure patients' mood, or trials of antimigraine medication that failed to measure pain.

The difficult decisions occur when the relation between physiologic or laboratory measures and HRQL outcomes is uncertain. Practitioners have relied on substitute end points not because they weren't interested in making patients feel better, but because they assumed a strong link between physiologic measurements and patients' well-being. A recent trial in patients with symptomatic postmenopausal osteoporosis examined the effect of sodium fluoride on bone density and vertebral fractures.[10] The investigators believed that increased bone mass and fewer vertebral fractures would lead to decreased pain and increased function. Does their failure to measure the effect of treatment on areas of unequivocal importance to patients, including pain, physical function, and household and leisure activities,[11] affect the clinical message of the results? Similarly, investigators measuring the effects of antianginal medication have often been satisfied with increased duration of exercise on the treadmill without direct measurement of decreased symptoms or increase in activity in day-to-day life. Are we ready to prescribe medication on the basis of increased laboratory exercise capacity?

Bone density, vertebral fractures, and exercise capacity, or similar measures such as joint count, ejection fraction, or pulmonary function, are surrogate end points for what we really want to measure: the effect of treatment on our patients' lives. Whether these surrogate measures are adequate depends on how confident we are of the link with how people feel. When this issue has been investigated empirically, the relation between physiologic and clinical measures and patients' symptoms is usually modest and often highly variable.[2-17] Though these findings lead us to recommend caution in assuming that improvement in physiologic or clinical function will result in patients feeling better, each clinician (and, when appropriate, the patient) must decide on her own threshold.

Referring to the opening scenario, investigators reported the results of a randomized trial of methotrexate in 141 patients with chronically active Crohn disease despite at least 3 months of prednisone therapy.[1] Patients who received methotrexate were twice as likely to be in clinical remission following 16 weeks of treatment than those who received placebo (39.4% vs 19.1%, $P=.02$), and actively treated patients received less prednisone and showed less disease activity. Is additional information regarding HRQL necessary to interpret the results of this study? As depicted in the scenario, the decision to give methotrexate depends on weighing the benefits and risks, and the patient's question about how much better he is likely to feel with medication may well be relevant to his decision. Without information about the effect of the medication on HRQL, therefore, neither the clinician nor the patient can make a fully informed choice.

## ARE THE RESULTS VALID?

### Primary Guides

**Have the Investigators Measured Aspects of Patients' Lives That Patients Consider Important?**—We have described how investigators often substitute end points that make intuitive sense to them for those that patients value. Clinicians can recognize these situations by asking themselves the question: If the end points measured by the investigators were the only thing that changed, would patients be willing to take the treatment? In addition to change in clinical or physiologic variables, patients would require that they feel better or live longer.

How can clinicians be secure that investigators have measured aspects of life that patients value? Investigators may show that the outcomes they have measured are important to patients by asking them directly. For example, in a study examining HRQL in patients with chronic airflow limitation, we used a literature review and interviews with clinicians and patients to identify 123 items reflecting possible ways their illness might affect patients' HRQL.[18] We then asked 100 patients which of the items were problems for them and how important those items were. We found that the most important problem areas for patients were their dyspnea during day-to-day activities and their chronic fatigue. An additional area of difficulty was emotional function, including feeling frustrated and impatient.

If the authors don't present direct evidence that their outcome measures are important to patients, they may cite prior work. For example, a randomized trial of respiratory rehabilitation in patients with chronic lung disease used an HRQL measure based on the responses of patients in the study we've described above and referred to that study.[19] Ideally, the report

will include a summary of the developmental process sufficiently detailed to obviate the need to go back to the prior report.

Alternatively, investigators may describe the content of their measures in detail. An adequate description of the content of a questionnaire allows clinicians to use their experience to decide whether what is being measured is important to patients. For instance, the authors of an article describing a randomized trial of surgery vs watchful waiting for benign prostatic hyperplasia "assessed the degree to which urinary difficulties bothered the patients or interfered with their activities of daily living, sexual function, social activities, and general well-being."[20] Few would doubt the importance of these items.

In the study of methotrexate for patients with inflammatory bowel disease (IBD), the patients completed the Inflammatory Bowel Disease Questionnaire (IBDQ), which addresses patients' bowel function, emotional function, systemic symptoms, and social function. Although the authors don't mention this in their article, the 32 items in the IBDQ were chosen because patients with IBD labeled them as the most important in their daily lives.[21]

**Did the HRQL Instruments Work in the Way They Are Supposed to?**—Measuring how people are feeling is not easy. Investigators must demonstrate that their instruments allow strong inferences about the effect of treatment on HRQL. We will now review how an HRQL measure should perform (we call the way it performs its *measurement properties*) if it is going to be useful.

*Signal and Noise.*—There are 2 distinct ways in which investigators use HRQL instruments. They may wish to help clinicians distinguish between people who have a better or worse HRQL, or to measure whether people are feeling better or worse over time.[22] For instance, suppose a trial of a new drug for patients with heart failure shows that it works best in patients with the New York Heart Association (NYHA) functional classification class IV symptoms. We could use the NYHA class for 2 purposes. One would be to discriminate between patients as to their NYHA class in deciding who to treat. We might also want to determine whether the drug was effective in improving an individual patient's functional status and therefore monitor changes in patients' NYHA functional class.

While for both purposes we require a high ratio of signal to noise, when we are discriminating between people at a single point in time, the signal comes from differences between patients (if everyone gets the same score, we can't tell who is better off and who is worse off), and the noise comes from variability within subjects (if patients' scores fluctuate wildly, we're not going to be able to say much about their relative well-being).[23] The technical term usually used for the ratio of variability between patients to the total variability is *reliability*.

Instruments used to evaluate change over time must, in contrast, be able to pick up any important changes in the way patients are feeling, even if those changes are small. Thus, the signal comes from the difference in score in patients who have improved or deteriorated, and the noise from the variability in score in patients who have not changed. The term we use for the ability to detect change (the ratio of signal to noise over time) is *responsiveness*.

An unresponsive instrument can result in a false-negative trial in which the intervention improves how patients feel, and yet the instrument fails to detect the improvement. This problem may be particularly salient for questionnaires that have the advantage of covering all relevant areas of HRQL, but the disadvantage of covering each area superficially. A crude instrument such as the NYHA functional classification (with only 4 categories) may work well for stratifying patients, but may not be able to detect small but important improvement with treatment.

In studies that show no difference in change in HRQL when patients receive a treatment vs a control intervention, clinicians should look for evidence that the instruments have been able to detect small or medium-sized effects in previous investigations. In the absence of this evidence, instrument unresponsiveness becomes a plausible reason for the failure to detect differences in HRQL. For example, a randomized trial of a diabetic education program reported no changes in 2 measures of well-being and attributed the result to, among other factors, lack of integration of the program with standard therapy.[24] Given that the program improved knowledge and self-care and patients felt less dependent on physicians, another explanation is inadequate responsiveness of the 2 HRQL measures.

In the trial of methotrexate in Crohn disease, concern about responsiveness decreases because the study showed statistically significant differences between treatment and control groups. As it turns out, the IBDQ had detected small to medium-sized differences in previous investigations.[21,25,26]

*Validity.*—Validity has to do with whether the instrument is measuring what it is intended to measure. The absence of a reference or criterion standard for HRQL creates a challenge for anyone hoping to measure how patients are feeling. We can be more confident that an instrument is doing its job if it appears targeted to the right problems (the technical term for this is *face validity*). Empirical evidence that it measures the domains of interest will also help.

To provide such evidence, investigators have borrowed validation strategies from psychologists who have for many years had to decide whether questionnaires assessing intelligence, attitudes, and emotional function were really measuring what is intended. Investigators interested in attitudes may show apparent differences between individuals that really reflect variability in the tendency to provide socially acceptable answers rather than differences in underlying attitudes; investigators may demonstrate apparent effects of rehabilitation on HRQL, but may really be detecting differences in satisfaction with care. In either case, the instrument would be detecting a signal, but it would be the wrong signal.

Establishing validity therefore involves examining the logical relationships that should exist between measures. For example, we would expect that, in general, patients with lower treadmill exercise capacity will have more dyspnea in daily life than those with higher exercise capacity, and we would expect to see substantial correlations between a new measure of emotional function and existing emotional function questionnaires. When we are interested in evaluating change over time, we examine correlations of change scores: patients who deteriorate on their treadmill exercise capacity should, in general, show increases in dyspnea, while those whose exercise capacity improves should experience less dyspnea; a new emotional function measure should show improvement in patients who improve on existing measures of emotional function. The technical term for this process is testing an instrument's *construct validity*.

Clinicians should look for evidence of the validity of HRQL measures used in clinical studies. Reports of randomized trials using HRQL measures seldom review evidence for the validity of the instruments they use, but clinicians can gain some reassurance from statements (backed by citations) that the questionnaires have been previously validated. In the absence of evident face validity or empirical evidence of validity, clinicians are entitled to skepticism about the study's measurement of HRQL.

In the methotrexate in IBD study, the investigators refer to the IBDQ as "previously validated" and provide 2 relevant citations.[21,25] These articles describe extensive validation of the questionnaire, including correlations of change that document the instrument's usefulness for measuring change over time.

### Secondary Guides

**Are There Important Aspects of HRQL That Have Been Omitted?**—Investigators may have addressed HRQL issues, but have not done so comprehensively. Exhaustive measurement may be more or less important in a particular context. One can think of a hierarchy that begins with symptoms, moves on to the functional consequences of the symptoms, and ends with more complex elements such as emotional function. If, as a clinician, you believe your patient's sole interest is in whether a treatment relieves the primary symptoms and most important functional limitations, you will be satisfied with a limited range of assessment. Recent randomized trials in patients with migraine[27,28] and postherpetic neuralgia[29] restricted themselves primarily to the measurement of pain; studies of patients with rheumatoid arthritis[30,31] and back pain[32] measured pain and physical function, but not emotional or social function.

As a clinician, you can judge whether or not these omissions are important to you or, more importantly, your patients. We would encourage you, however, to bear in mind the broader impact of disease on patients' lives. Disease-specific measures that explore the full range of patients' problems and experience remind us of domains we might otherwise forget. We can trust these measures to be comprehensive if the developers have conducted a detailed survey of patients suffering from the illness or condition.

If you are interested in going beyond the specific illness and comparing the impact of treatments on HRQL across diseases or conditions, you will require a more comprehensive assessment. None of the disease-specific, system- or organ-specific, function-specific (such as instruments that examine sleep or sexual function), or problem-specific (such as pain) measures are adequate for comparisons across conditions. These comparisons require generic measures designed for administration to people with any underlying health problem (or no problem at all) that cover all relevant areas of HRQL.

One type of generic measure, health profiles, yields scores for all domains of HRQL (including, for example, mobility, self-care, and physical, emotional, and social function). There are a number of well-established health profiles, including the Sickness Impact Profile[33] and the short forms of the instruments used in the Medical Outcomes Study[34,35] that have advantages of simplicity, self-administration, and the ability to put changes in specific functions in the context of overall HRQL. Inevitably, such instruments cover each area superficially. This may limit their responsiveness—indeed, several randomized trials have found that generic instruments were less powerful in detecting treatment effects than specific instruments.[19,36-40] Ironically, generic instruments may also suffer from not being sufficiently comprehensive: they may completely omit patients' primary symptoms.

Disease-specific measures may comprehensively sample all aspects of HRQL relevant to a specific illness and also be responsive, but they are unlikely to deal with adverse effects. For instance, the IBDQ measures all important disease-specific areas of HRQL, including symptoms directly related to the primary bowel disturbance, systemic symptoms, and emotional and social function. Coincidentally, it measures some methotrexate adverse effects, including nausea and lethargy, because these are also experienced by patients with IBD not taking methotrexate, but not other adverse effects such as rash or mouth ulcers. The investigators could have administered a generic instrument to tap in to non–IBD-related aspects of HRQL, but once again would likely have failed to measure adverse effects in sufficient detail. Adverse effect–specific instruments are limited; the investigators chose a checklist approach and documented the frequency of occurrence of adverse events both severe and not severe enough to warrant discontinuation of treatment.

**If There Were Trade-offs Between Quality and Quantity of Life, or an Economic Evaluation, Have the Investigators Used the Right Measures?**—While providing information about the broad domains of HRQL and therefore allowing comparisons across conditions, health profiles are ill-suited for health policy decisions that involve integrating costs. Health policy decisions require choices about resource allocation across diseases, conditions, or medical problems, and also involve considerations of cost. These choices require standardized comparisons that allow one to relate the impact of very different treatments (such as drugs, surgery, or rehabilitation programs) on very different conditions (such as chronic lung disease, renal failure, or Parkinson disease). Inevitably, they involve putting a value on health states and may thus require sophisticated weighting for patient preferences, and necessitate relating health states to anchors of death and full health. Such measures may aid policymakers in making the right decisions about how public money is allocated.

Measures that provide a single number that summarizes all of HRQL are preference or value weighted, and have the preferences or values anchored to death and full health are called *utility measures*. Typically, utility measures use a scale from 0 (death) to 1.0 (full health) to summarize HRQL. Since they weight the duration of life according to its quality, their output is often called *quality-adjusted life years* (QALYs). Thus, utilities are holistic measures that ask patients to express, in a single value, their strengths of preferences for particular health states.

Boyle and colleagues,[41] in a classic article, used a utility measure to calculate that treating critically ill infants weighing 1000 to 1499 g at birth cost $3200 per QALY gained, while treating infants with a birth weight of 500 to 999 g cost $22 400 per QALY gained.[41] Estimates for the cost per QALY for treating patients receiving renal dialysis have ranged from approximately $30 000 to $50 000.[42,43] While different weighting schemes yield different results and may therefore be considered arbitrary, a number of increasingly simple utility measures are now available, have provided interesting results in clinical trials, and may facilitate integrating cost into policy decisions. However, the use, measurement, and interpretation of utility measures remain controversial.[44] The investigators in the methotrexate trial did not use a health profile or a utility measure, thus limiting use of the data for comparisons across disease states and preventing a formal economic analysis.

### What Were the Results?

**What Was the Magnitude of Effect on HRQL?**—Understanding the results of a trial involving HRQL involves special challenges. Patients with acute back pain who were prescribed bed rest had mean scores on the Owestry Back-Disability Index, a measure that focuses on disease-specific functional status, 3.9 points worse than control patients.[32] Patients with severe rheumatoid arthritis allocated to cyclosporine had a mean disability score 0.28 unit better than control patients.[30] Are these differences trivial, small but important, of moderate magnitude, or do they constitute large and extremely important differences between treatments?

These examples show that the interpretability of most HRQL measures is not self-evident. There are a number of

methods available for understanding the magnitude of HRQL effects. Investigators may relate changes in HRQL questionnaire score to well-known functional measures (such as the NYHA functional classification), to clinical diagnosis (such as the change in score needed to move people in or out of the diagnostic category of depression), or to the impact of major life events.[45] They may relate changes in HRQL score to patients' global ratings of the magnitude of change they have experienced,[46] or to the extent they rate themselves as better or worse than other patients.[47] Whatever the strategy, if investigators don't provide an indication of how to interpret changes in HRQL score, the findings are of limited use to clinicians.

Even if we did know that 3.9 points on the Owestry Back-Disability Index or 0.28 unit on a rheumatoid arthritis disability index signified, for instance, small but important changes, mean differences between groups may be difficult to interpret. Clinicians may find the proportion of patients who achieved small, medium, and large gains due to treatment more informative.

The investigators who conducted the trial of methotrexate for Crohn disease do not help clinicians interpret the magnitude of difference in HRQL. The mean difference in IBDQ score between treatment and control groups at 16 weeks was 0.59. Other investigations suggest that differences of approximately 0.5 may represent small but important changes, while large improvements correspond to a difference in score of greater than 1.0.[46-49] Thus, the mean difference between treated and control patients in the methotrexate study likely falls into the category of small but important change in HRQL.

### Will the Results Help Me in Caring for My Patients?

**Will the Information From the Study Help Me Inform My Patients?**—People with the same chronic disease often vary markedly in the problems they experience. Even if the problems are the same, the magnitude of the impact of those problems in their lives may differ. Assessment of HRQL will only help in the care of an individual patient if that patient's problems are similar to those of patients in the trial.

Knowing whether HRQL results of a study are relevant for your patients means understanding their experience of illness. Even the most common problems of a chronic disease don't affect all those afflicted. For instance, 92% of patients with IBD complain of frequent bowel movements, and 82% complain of abdominal cramps.[50] With respect to emotional function, 78% feel frustrated and 76% feel depressed. The patients who experienced these difficulties vary in the extent to which they felt the problems were important. Thinking back to the scenario, before answering the question about how the treatment would affect the patient's life, the clinician would have to find out the problems the patient was currently experiencing, the importance he attached to those problems, and the value he might attach to having the problems ameliorated.

Reflecting further on the process of communicating with patients, HRQL instruments that focus on specific aspects of patients' experience may be more useful than global measures. Patients with chronic lung disease may find it more informative to know that their compatriots offered a treatment became less dyspneic and fatigued in daily activity, rather than simply that they judged their HRQL as improved. HRQL measures will be most useful when the results facilitate their practical use by you and your patients.

**Did the Study Design Simulate Clinical Practice?**—Treatments affect HRQL both by reducing disease symptoms and consequences and by creating new problems. Adverse effects may make the cure worse than the disease. Clinicians conducting clinical trials are usually blind to treatment allocation and try to maintain patients on the study medication as long as possible. Patients may therefore soldier on in the face of considerable adverse effects, and this may be reflected in their HRQL.

This is not how we conduct our clinical practice. If patients experience significant adverse effects, we discontinue the medication, particularly if there is a suitable alternative. Thus, the design of the clinical trial may create an artificial situation with misleading estimates of the impact of treatment on HRQL. This issue is of particular concern for treatments such as antihypertensive drugs in which much of the impairment in HRQL may be due not to the medical condition, but to the treatment.

The trial of methotrexate in Crohn disease simulated clinical practice well. If the patient is experiencing problems similar to those of the trial patients, and if those problems are important to him, he is likely to achieve comparable benefit to patients enrolled in the trial.

## CONCLUSION

We encourage clinicians to consider the impact of their treatments on patients' HRQL, and to look for information regarding this impact in clinical trials. Responsive, valid, and interpretable instruments measuring experiences of importance to most patients should increasingly help guide our clinical decisions.

We acknowledge a useful review of the manuscript by Brian Feagan, MD, who reassured us we were on the right track with our scenario. We offer special thanks to Deborah Maddock who has provided outstanding administrative support and coordination for the activities of the Evidence-Based Medicine Working Group.

### References

1. Feagan BG, Rochon J, Fedorak RN, et al. Methotrexate for the treatment of Crohn's disease: the North American Crohn's Study Group Investigators. *N Engl J Med.* 1995;332:292-297.
2. Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life: basic sciences review. *Ann Intern Med.* 1993;70:225-230.
3. Gill TM, Feinstein AR. A critical appraisal of the quality of quality-of-life measurements. *JAMA.* 1994;272:619-626.
4. Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life: a conceptual model of patient outcomes. *JAMA.* 1995;273:59-65.
5. Guyatt GH, Cook DJ. Health status, quality of life, and the individual. *JAMA.* 1994;272:630-631.
6. Nocturnal Oxygen Therapy Trial Group. Continuous or nocturnal oxygen therapy to hypoxemic chronic obstructive lung disease. *Ann Intern Med.* 1980;93:391-398.
7. Heaton RK, Grant I, McSweeney AJ, Adams KM, Petty TL. Psychologic effects of continuous and nocturnal oxygen therapy in hypoxemic chronic obstructive pulmonary disease. *Arch Intern Med.* 1983;143:1941-1947.
8. Mulrow CD, Mulrow JP, Linn WD, Aguilar C, Ramirez G. Relative efficacy of vasodilator therapy in chronic congestive heart failure: implications of randomized trials. *JAMA.* 1988;259:3422-3426.
9. Packer M, Carver JR, Rodeheffer RJ, et al. Effect of oral milrinone on mortality in severe chronic heart failure: the PROMISE Study Research Group. *N Engl J Med.* 1991;325:1468-1475.
10. Pak CY, Sakhaee K, Piziak V, et al. Slow-release sodium fluoride in the management of postmenopausal osteoporosis. *Ann Intern Med.* 1994; 120:625-632.
11. Cook DJ, Guyatt GH, Adachi JD, et al. Quality of life issues in women with vertebral fractures due to osteoporosis. *Arthritis Rheum.* 1993;36:750-756.
12. Guyatt GH, Thompson PJ, Berman LB, et al. How should we measure function in patients with chronic heart and lung disease? *J Chronic Dis.* 1985; 38:517-524.
13. Mahler DA, Weinberg DH, Wells CK, Feinstein AR. The measurement of dyspnea: contents, interobserver agreement, and physiologic correlates of two new clinical indexes. *Chest.* 1984;85:751-758.
14. Rector TS, Kubo SH, Cohn JN. Patients' self-assessment of their congestive heart failure, II: content, reliability and validity of a new measure—the Minnesota Living With Heart Failure Questionnaire. *Heart Failure.* 1987;3:198-209.
15. Juniper EF, Guyatt GH, Ferrie PJ, Griffith LE. Measuring quality of life in asthma. *Am Rev Respir Dis.* 1993;147:468-479.
16. Wiklund I, Comerford MB, Dimenas E. The relationship between exercise tolerance and quality of life in angina pectoris. *Clin Cardiol.* 1991;14:204-208.
17. Osteoporosis Quality of Life Research Group. Measuring quality of life in women with osteoporosis. *J Bone Miner Res.* In press.
18. Guyatt GH, Berman LB, Townsend M, Pugsley SO, Chambers LW. A measure of quality of life for clinical trials in chronic lung disease. *Thorax.* 1987; 42:773-778.
19. Goldstein RS, Gort EH, Guyatt GH, Stubbing D, Avendano MA. Prospective randomized controlled trial of respiratory rehabilitation. *Lancet.* 1994;344: 1394-1397.
20. Wasson JH, Reda DJ, Bruskewitz RC, Elinson J, Keller AM, Henderson WG. A comparison of trans-

urethral surgery with watchful waiting for moderate symptoms of benign prostatic hyperplasia: the Veterans Affairs Cooperative Study Group on Transurethral Resection of the Prostate. *N Engl J Med.* 1995;332:75-79.
21. Guyatt GH, Mitchell A, Irvine EJ, et al. A new measure of health status for clinical trials in inflammatory bowel disease. *Gastroenterology.* 1989;96:804-810.
22. Kirshner B, Guyatt GH. A methodologic framework for assessing health indices. *J Chronic Dis.* 1985;38:27-36.
23. Guyatt GH, Kirshner B, Jaeschke R. Measuring health status: what are the necessary measurement properties? *J Clin Epidemiol.* 1992;45:1341-1345.
24. De Weerdt I, Visser AP, Kok GI, de Weerdt O, Van der Veen EA. Randomized controlled multicentre evaluation of an education programme for insulin-treated diabetic patients: effects on metabolic control, quality of life, and costs of therapy. *Diabet Med.* 1991;8:338-345.
25. Irvine EJ, Feagan B, Rochon J, et al. Quality of life: a valid and reliable measure of therapeutic efficacy in the treatment of inflammatory bowel disease. *Gastroenterology.* 1994;106:287-296.
26. Greenberg GR, Feagan BG, Martin F, et al. Oral budesonide for active Crohn's disease: Canadian Inflammatory Bowel Disease Study Group. *N Engl J Med.* 1994;331:836-841.
27. Salonen R, Ashford E, Dahlof C, et al. Intranasal sumatriptan for the acute treatment of migraine. *J Neurol.* 1994;241:463-469.
28. Mathew NT, Saper JR, Silberstein SD, et al. Migraine prophylaxis with divalproex. *Arch Neurol.* 1995;52:281-286.
29. Trying S, Barbarash RA, Nahlik JE, et al. Famciclovir for the treatment of acute herpes zoster: effects on acute disease and postherpetic neuralgia. *Ann Intern Med.* 1995;123:89-96.
30. Tugwell P, Pincus T, Yocum D, et al. Combination therapy with cyclosporin and methotrexate in severe rheumatoid arthritis: the Methotrexate-Cyclosporine Combination Study Group. *N Engl J Med.* 1995;333:137-141.
31. Kirwan JR. The effect of glucocorticoids on joint destruction in rheumatoid arthritis: the Arthritis and Rheumatism Council Low-Dose Glucocorticoid Study Group. *N Engl J Med.* 1995;333:142-146.
32. Malmivaara A, Hakkinen U, Aro T, et al. The treatment of acute low back pain: bed rest, exercises, or ordinary activity. *N Engl J Med.* 1995;332:351-355.
33. Bergner M, Bobbit RA, Carter WB, Gilson BS. The Sickness Impact Profile: development and final revision of a health status measure. *Med Care.* 1981; 19:787-805.
34. Tarlov AR, Ware JE Jr, Greenfield S, Nelson EC, Perrin E, Zubkoff M. The Medical Outcomes Study: an application of methods for monitoring the results of medical care. *JAMA.* 1989;262:925-930.
35. Ware JE, Kosinski M, Bayliss MS, et al. Comparison of methods for the scoring and statistical analysis of SF-36 health profile and summary measures: summary of results of the Medical Outcomes Study. *Med Care.* 1995;33:AS264-AS279.
36. Tandon PK, Stander H, Schwarz RP Jr. Analysis of quality of life data from a randomized, placebo controlled heart-failure trial. *J Clin Epidemiol.* 1989; 42:955-962.
37. Smith D, Baker G, Davies G, Dewey M, Chadwick DW. Outcomes of add-on treatment with lamotrigine in partial epilepsy. *Epilepsia.* 1993;34:312-322.
38. Chang SW, Fine R, Siegel D, Chesney M, Black D, Hulley SB. The impact of diuretic therapy on reported sexual function. *Arch Intern Med.* 1991; 151:2402-2408.
39. Tugwell P, Bombardier C, Buchanan WW, et al. Methotrexate in rheumatoid arthritis: impact on quality of life assessed by traditional standard-item and individualized patient preference health status questionnaires. *Arch Intern Med.* 1990;150:59-62.
40. Laupacis A, Wong C, Churchill D. The use of generic and specific quality-of-life measures in hemodialysis patients treated with erythropoietin. *Control Clin Trials.* 1991;12:168S-179S.
41. Boyle MH, Torrance GW, Sinclair JC, Horwood SJ. Economic evaluation of neonatal intensive care of very-low-birth-weight infants. *N Engl J Med.* 1988; 308:1330-1337.
42. Hornberger JC, Garber AM, Chernew ME. Is high-flux dialysis cost-effective? *Int J Technol Assess Health Care.* 1993;9:85-96.
43. Hornberger JC. The hemodialysis prescription and cost effectiveness. *J Am Soc Nephrol.* 1993;4: 1021-1027.
44. Naylor CD. Cost-effectiveness analysis: are the outputs worth the inputs? *ACP J Club.* 1996;124: a12-a14.
45. Testa MA, Anderson RB, Nackley JF, Hollenberg NK, for the Quality-of-Life Hypertension Study Group. Quality of life and antihypertensive therapy in men: a comparison of captopril with enalapril. *N Engl J Med.* 1993;328:907-913.
46. Juniper EF, Guyatt GH, Willan A, Griffith LE. Determining a minimal important change in a disease-specific quality of life questionnaire. *J Clin Epidemiol.* 1994;47:81-87.
47. Redelmeier DA, Goldstein RS, Guyatt GH. Assessing the minimal important difference in symptoms: a comparison of two techniques. *J Clin Epidemiol.* 1996;49:1215-1219.
48. Jaeschke R, Guyatt G, Keller J, Singer J. Measurement of health status: ascertaining the meaning of a change in quality-of-life questionnaire score. *Control Clin Trials.* 1989;10:407-415.
49. Juniper EF, Guyatt GH, Feeny DH, Ferrie PJ, Griffith LE, Townsend M. Measuring quality of life in children with asthma. *Qual Life Res.* 1996;5:35-46.
50. Mitchell A, Guyatt G, Singer J, et al. Quality of life in patients with inflammatory bowel disease. *J Clin Gastroentrol.* 1988;10:306-310.

# Users' Guides to the Medical Literature

## XIII. How to Use an Article on Economic Analysis of Clinical Practice

## A. Are the Results of the Study Valid?

Michael F. Drummond, PhD; W. Scott Richardson, MD; Bernie J. O'Brien, PhD; Mitchell Levine, MD; Daren Heyland, MD; for the Evidence-Based Medicine Working Group

### CLINICAL SCENARIO

You are a general internist on the staff of a large community hospital. Your chief of medicine knows of your interest in evidence-based medicine, and she asks you to help her solve a problem. The hospital's pharmacy and therapeutics committee has been trying to decide on formulary guidelines for the use of streptokinase or tissue-type plasminogen activator (t-PA) in the treatment of acute myocardial infarction (AMI). Members of the committee have been arguing for weeks about the Global Utlization of Streptokinase and Tissue Plasminogen Activator for Occluded Coronary Arteries (GUSTO) trial[1] and whether the added expense of t-PA is worth it. The committee has reached an impasse and has asked the chief of medicine for some outside help to reach a good decision. Knowing that the hospital faces pressure to keep costs down, the chief wants good information about this question to bring to the next committee meeting later this week. She asks you to help her find out if a formal economic analysis that compares thrombolytic agents for AMI has been done and then help her present it to the committee.

### The Search

From your office computer you enter the hospital library's CD-ROM MEDLINE system via the hospital's information network. In the current MEDLINE file, you cross the terms "myocardial infarction" (11 099 citations), "thrombolytic therapy" (3350 citations), and "cost-benefit analysis" (4232 citations). This yields a set of only 11. Reviewing these on screen, you find 3 articles directly relevant to your question. One is an economic analysis done as part of the GUSTO study,[2] and another is an economic analysis using data from the GUSTO trial in a decision model.[3] Your searching program includes a "Local Messages" field, and this field reports that both of these studies are available in your hospital's library. Your search also turns up another analysis based on modeling,[4] but the "Local Messages" note indicates that this journal is not available in your library. You request a copy via interlibrary loan, but realize it will probably arrive long after the committee's meeting later this week. You thus turn to the first 2 articles, hoping to find some evidence you can use to help the committee.

### INTRODUCTION

In the course of their work, clinicians make many decisions about the care of individual patients. Clinicians are also asked to participate in decisions for large groups of patients, whether to set clinical policy for an institution ("Should streptokinase or t-PA be recommended routinely for patients with an AMI who present to our hospital?"), or to set health policy at a more macro level ("Which thrombolytic agents should our national or local health authority choose to purchase and provide for our citizens who suffer AMI?"). When making decisions for such patient groups, clinicians need to not only weigh the benefits and risks, but should also consider whether these benefits will be worth the health care resources consumed. Resources used to provide health care are vast, but not limitless. This is particularly the case in managed care settings where, in essence, a fixed sum is available to provide care for enrollees. Thus, more and more, clinicians will have to convince colleagues and health policymakers that the benefits of their interventions justify the costs.

To inform these decisions, clinicians can use economic analyses of clinical practices. Economic analysis is a set of formal, quantitative methods used to compare alternative strategies with respect to their resource use and their expected outcomes.[5,6] Economic evaluations seek to inform resource allocation decisions, not make them. Economic analyses have been attracting more attention in recent years and could potentially inform decisions at different levels in the health care system, such as managing major institutions like hospitals and in determining regional or national policy.[7-9]

Randomized trials generate data about relative treatment efficacy, but sometimes investigators may also collect data about cost. As with other integrative studies such as decision analyses[10] and practice guidelines,[11] economic analyses may use estimates of cost and effectiveness from summaries of several studies of therapy, diagnosis, and prognosis. Either way, the main distinction between economic analyses and other studies is the explicit measurement and valuation of resource consumption or cost. The integration of cost data often involves placing values on the health outcomes so that they can be related to the costs of alternative treatment strategies.

From the Centre for Health Economics, University of York, York, England (Dr Drummond); Department of Medicine, University of Rochester School of Medicine and Dentistry, Rochester, NY (Dr Richardson); Department of Clinical Epidemiology and Biostatistics, McMaster University and Centre for Evaluation of Medicines, St Joseph's Hospital, Hamilton, Ontario (Drs O'Brien and Levine); and Royal Alexandra Hospital, Edmonton, Alberta (Dr Heyland).

The original list of members (with affiliations) appears in the first article of this series (*JAMA*. 1993;270:2093-2095). A list of new members appears in the 10th article of the series (*JAMA*. 1996;275:1435-1439). The following members contributed to this article: Gordon H. Guyatt, MD, MSc (chair); Roman Jaeschke, MD, MSc; Deborah J. Cook, MD, MSc; Hertzel Gerstein, MD, MSc; Stephen Walter, PhD; John Williams, Jr, MD, MHS; and C. David Naylor, MD, MSc, DPhil.

Reprints: Gordon H. Guyatt, MD, MSc, McMaster University Health Sciences Centre, 1200 Main St W, Room 2C12, Hamilton, Ontario, Canada L8N 3Z5.

Users' Guides to the Medical Literature section editor: Drummond Rennie, MD, Deputy Editor (West), *JAMA*.

In helping you understand economic analyses, we will introduce you to how these analyses are conducted and review some of their strengths and weaknesses. This is not, however, an article on how to perform economic analysis; should you wish to do so, you should look elsewhere.[12-14] Since you may frequently encounter economic analyses that are based on decision models, you may also find it useful to review the earlier articles in the series on clinical decision analysis[10] when reading such studies.

## THE FRAMEWORK FOR THE USERS' GUIDES

We will approach articles on economic analysis of clinical strategies with the same 3 organizing questions introduced in earlier articles in this series:

### Are the Results Valid?

This question addresses whether an economic analysis truly determines which of the clinical strategies would provide the most benefit for the available resources. Just as with other types of studies, the validity of an economic analysis is primarily determined by the strength of the methods used.

### What Were the Results?

If the answer to the first question was yes, and the economic analysis likely yields an unbiased assessment of the costs and outcomes of the clinical strategies under study, then the results are worth examining further. The guides under this second question consider the size of the expected benefits and costs from adopting the most efficient strategy and the level of uncertainty in the results.

### Will the Results Help in Caring for My Patients?

If the economic analysis yields valid and important results, you can then examine how to apply these results in your own clinical setting.

Table 1 summarizes the specific questions you can ask in addressing these 3 areas. We will explore the guides by applying them to the articles we found in our search. This article will deal with the validity guides, while the next in the series will address the results and applicability.

## ARE THE RESULTS VALID?

### Did the Analysis Provide a Full Economic Comparison of Health Care Strategies?

Economic analyses compare 2 or more treatments, programs, or strategies. If 2 strategies are analyzed but only costs are compared, this comparison would inform only the resource-use half of the decision and is termed a *cost analysis*. Comparing 2 or more strategies only by their efficacy (such as in a randomized trial) informs only the outcomes portion of the decision. A full economic comparison requires that both the costs and outcomes be analyzed for each of the strategies being compared. To help you understand the structure of the comparison further, some additional questions will be useful.

**Was a Broad Enough Viewpoint Adopted?**—Costs and outcomes can be evaluated from a number of viewpoints: the patient, the hospital, the third-party payer (eg, health maintenance organization), or society at large. Each viewpoint may be relevant depending on the question being asked, but broader viewpoints are most relevant to those concerned about the overall allocation of health care resources.[9] That is, an evaluation adopting, for example, the viewpoint of the hospital will be useful in estimating the budgetary impact of alternative therapies for that institution. However, economic evaluation is usually directed at informing policy from a broader societal perspective.

For example, in an evaluation of an early discharge program, it is not sufficient to report only hospital costs, since patients discharged early may consume substantial resources in the community. These costs may not be borne by the hospital, but are likely to impact on a third-party payer or the patient in some way or another. This was a limitation of the study by Topol et al,[15] which assessed the feasibility and cost savings of hospital discharge 3 days after AMI, considering only hospital and professional charges. We have no knowledge of other community services consumed and whether these differed between early discharge and conventional discharge patients.

One of the main reasons for considering narrower viewpoints in conducting an economic analysis is to assess the impact of change on the main budget holders, since budgets or payments may need to be adjusted before a new therapy can be adopted. This is particularly true in countries like the United States, where resource-allocating decisions are made in a decentralized way by a range of actors rather than a health ministry. Weisbrod et al[16] pointed out that while a community-oriented mental illness program was worthwhile from the perspective of society as a whole, it would be more costly to the organization responsible for providing the care. Even within the same institution, narrow budgetary viewpoints can prevail. In our example comparing streptokinase with t-PA, it would be wrong just to focus on the relative costs of the drugs, which fall on the pharmacy budget, if there are also impacts on the use of other hospital resources.

The patient's perspective may also merit specific consideration if costs (eg, in travel) reduce access to care. Also, some patients may not be able to participate in community care programs if these impose major costs in terms of informal nursing support in the home. In some countries, most notably the United States, patients may also be responsible for a sizable proportion of their health care bills. Many economic analysts do not track all of these costs, owing to the time and effort required. However, the patient's perspective is partially integrated into the analysis by measuring the outcomes of therapy, such as impact on quality of life.

The way in which the articles by Mark et al[2] and Kalish et al[3] handle these and other key methodological issues is presented in Table 2. Mark et al[2] point out the importance of considering a broad, societal viewpoint, whereas Kalish et al[3] do not discuss the issue. In practice, both analyses concentrate on the identification and quantification of direct medical care costs, both inside and outside the hospital. The reasons for exclusion of other cost items, such as patients' costs, are not explicitly discussed, but may relate to the practical problems of data collection.

The breadth of outcomes considered varies according to the type of economic analysis. In cost-effectiveness analyses the health outcomes are not valued, but reported in physical units such as *life years gained* or *cases successfully treated*. In a variant of cost-effectiveness analysis, sometimes called *cost-utility analysis*, outcomes of different types are weighted to produce a composite

Table 1.—Users' Guides for Economic Analysis of Clinical Practice

| |
|---|
| **Are the results valid?** |
| Did the analysis provide a full economic comparison of health care strategies? |
| Were the costs and outcomes properly measured and valued? |
| Was appropriate allowance made for uncertainties in the analysis? |
| Are estimates of costs and outcomes related to the baseline risk in the treatment population? |
| **What were the results?** |
| What were the incremental costs and outcomes of each strategy? |
| Do incremental costs and outcomes differ between subgroups? |
| How much does allowance for uncertainty change the results? |
| **Will the results help in caring for my patients?** |
| Are the treatment benefits worth the harms and costs? |
| Could my patients expect similar health outcomes? |
| Could I expect similar costs? |

Table 2.—Key Methodological Features of the 2 Studies

| Feature | Mark et al[2] | Kalish et al[3] |
|---|---|---|
| Overall study design | Cost-effectiveness and cost-utility analysis concurrent with clinical trial | Cost-utility analysis using a decision-analytic model |
| Viewpoint for analysis | Societal | Not stated |
| Alternatives compared | t-PA or streptokinase for patients with acute myocardial infarction | t-PA or streptokinase for patients with acute myocardial infarction |
| Benefit measure(s) | Life-years saved and quality-adjusted life-years saved | Quality-adjusted life-years saved |
| Source(s) of effectiveness data | GUSTO trial (1-y survival) and Duke Cardiovascular Disease Database (long-term survival) | GUSTO trial (1-y survival) and Worcester Heart Attack Study (long-term survival) |
| Source(s) of quality of life (utility) weights | Sample of 2600 US patients enrolled in the GUSTO trial | GISSI-2 trial |
| Estimates of resource use | 23 105 US patients enrolled in the GUSTO trial (for initial hospitalization); sample of 2600 US patients (for resource use up to 1 y) | Brigham and Women's Hospital and the literature |
| Source(s) of cost data | Duke cost accounting system and Medicare DRG rates | Brigham and Women's Hospital and the literature |
| Discounting | 5% per year | 5% per year |
| Sensitivity analysis | Varied estimates of survival and cost; also varied discount rate and considered importance of disabling strokes | Varied estimates of survival cost and stroke rate; also varied discount rate |

*t-PA indicates tissue-type plasminogen activator; GUSTO, Global Utilization of Streptokinase and Tissue Plasminogen Activator for Occluded Coronary Arteries; GISSI-2, Gruppo Italiano per lo Studio della Sopravvivenza nell'Infarto Miocardico; and DRG, diagnosis related group.

index, such as the quality-adjusted life year (QALY)[12] or healthy years equivalent.[17] Quality adjustment involves placing a lower value on time spent with impaired physical and emotional function than time spent in full health. On a scale where 0 represents death and 1 represents full health, the greater the impairment, the lower the value of a particular health state. These approaches are particularly useful when alternative treatments produce outcomes of different types, or when increased survival is bought at the expense of reduced quality of life.

Finally, in cost-benefit analyses, the health consequences are valued by asking health care consumers what they would be willing to pay for health services that achieve combinations of outcomes of particular types. This has an advantage in that it would be possible to assess directly whether the intervention is worthwhile to society, as all costs and outcomes would be valued in the same units (usually dollars). However, this approach may introduce a bias toward interventions for the rich, if their willingness to pay were higher than that of the poor. Nevertheless, it is worth remembering that most of the methods of economic evaluation ultimately lead toward some type of social valuation, such as how much we are willing to pay to gain an extra year of life or an extra QALY. Also, the QALY approach introduces another kind of bias in favor of those individuals with potentially more years to live in a good health state.

In the study by Mark et al,[2] the primary analysis was cost-effectiveness analysis, using the outcome *years of life saved*. The outcome in QALYs was considered in a secondary analysis. In the study by Kalish et al[3] the primary analysis used QALYs. In both cases the value of states of health were obtained by the time trade-off approach; that is, by asking patients how many years in their current state of health they would be willing to give up to live their remaining years in excellent health. Mark et al[2] obtained these values from patients in the GUSTO trial 1 year after treatment. Kalish et al[3] obtained them from a subset of patients in the Gruppo Italiano per lo Studio della Sopravvivenza nell'Infarto Miocardico (GISSI-2) trial.

Another type of consequence is the impact that therapy may have on the patient's ability to work and hence her or his contribution to the nation's production. These impacts are known as *indirect costs and benefits* in much of the health economics literature, but this terminology is falling from favor as it is at odds with the accounting use of the term *indirect costs*, to mean overhead. The issue of inclusion or exclusion of productivity changes is a frequent topic of debate. On one hand, these represent resource-use changes just like those occurring in the health care system. On the other hand, production may not actually be lost if a worker is absent for a short period. Also, for longer periods of absence, a previously unemployed worker may be employed. Furthermore, inclusion of productivity changes biases evaluations in favor of programs for those individuals who are employed full-time. Therefore, you should be skeptical about any economic analysis that includes productivity changes without clearly presenting the implications.

Neither of the thrombolytic studies discussed here considered productivity changes. The inclusion would be unlikely to substantially influence the comparison between streptokinase and t-PA, and may not be appropriate. However, the exclusion of lost productivity could constitute another argument for thrombolysis over a treatment strategy of no thrombolysis.

**Were All the Relevant Clinical Strategies Compared?**—The second assessment of the breadth of an economic evaluation relates to the range of alternative strategies examined. A frequently omitted strategy is that of maintaining the status quo. Another mistake is to view alternatives as being all or nothing. In medicine it is not often a question of whether one should adopt a particular test or apply a particular therapy, but how much of it should be applied. Thus, the interesting and more clinically relevant questions often relate to whether a given procedure should be applied selectively or routinely, whether a treatment should be given to low-risk patients as well as to high-risk patients, or whether the dose of a drug should be intensified.

One difficulty faced by economic analysts is that the comparisons they would like to make are to some extent limited by the availability of clinical data. A particular concern is the fact that clinical trials of many new medicines make a comparison with placebo rather than another active therapy. This means that, often, economic analyses cannot be based on either a particular clinical trial or an overview of several trials. Rather, they become integrative studies that, of necessity, employ a number of assumptions. Therefore, users of economic analyses need to check on the methods of the studies generating the clinical data for the economic analysis and whether such studies are really comparable. They may be concerned if the clinical data used in an economic evaluation came from studies that enrolled patients of different baseline risk, or measured clinical outcomes in a slightly different way.

Both the articles by Mark et al[2] and Kalish et al[3] examine only the strategies compared in the GUSTO trial. This is reasonable because previous randomized trials had shown that thrombolysis was both effective and cost-effective when compared with no treatment, so the issue of a do-nothing strategy does not

arise. However, the question of which patients should be treated with a particular therapy is likely to be important (we return to this point later).

### Were the Costs and Outcomes Properly Measured and Valued?

**Was Clinical Effectiveness Established?**—To be valid, economic evaluations require evidence on the effectiveness of the alternatives being compared. The standards for assessment of effectiveness correspond to those discussed in earlier guides in the series. Although evidence based on experiments, such as that obtained from randomized trials, is considered the best evidence for answering questions of therapy, economic evaluations are more valid if effectiveness data reflect normal clinical practice as closely as possible. Some economic evaluations are now being undertaken concurrently with randomized trials. Others are being based on systematic overviews of a number of trials. For example, Mugford et al[18] used data from a systematic overview of 58 controlled trials to estimate the cost-effectiveness of giving prophylactic antibiotics routinely to reduce the incidence of wound infection after cesarean delivery.

The decision about whether to base an economic evaluation on results of a single trial, an overview of a number of trials, or a broader synthesis (in a modeling study) of trial and other evidence is not straightforward. In principle, all 3 approaches can be used. The considerations that guide the choice of approach in a given situation are as follows.

An evaluation based on prospective economic data collection alongside a single methodologically rigorous trial has high internal validity. However, the results may not be widely generalizable (that is, they may have low external validity) if the setting for the trial was atypical, the protocol highly prescriptive, or compliance higher than one would expect in routine clinical practice. An evaluation based on an overview of a number of trials is likely to be more precise, as the pooled estimate of effectiveness will have a narrower confidence interval (CI), and is likely to be more widely generalizable because of a wider range of patients, practice settings, and ways of administering the intervention in several trials.

Sometimes data from trials require adjustment when used in an economic analysis. In their economic evaluation of misoprostol, a drug for prophylaxis against gastric ulcer in patients receiving long-term nonsteroidal anti-inflammatory drugs (NSAIDs), Hillman and Bloom[19] used clinical data from a trial undertaken by Graham et al.[20] This evaluation compared misoprostol (400 μg and 800 μg daily) with placebo in a double-blind randomized controlled trial of 3 months' duration. An important issue for economic analysis was that ulcers prevented by misoprostol may generate savings in health care expenditure, which could balance the cost of adding the drug. However, it was not possible to use the rates of ulcer observed in the trial for the economic analysis without adjustment. First, lesions were discovered by endoscopy, which was performed monthly. Many of these ulcers would not have come to the notice of the patient or her physician in regular practice. Second, the compliance rate observed in the trial was higher than that typically observed in patients taking NSAIDs. Therefore, Hillman and Bloom adjusted the observed ulcer rates to reflect the fact that 40% of endoscopically determined lesions remain silent. They also adjusted for lower compliance by using the ulcer rates in the evaluable cohort and assuming that only 60% of this efficacy would be achieved in practice.

Sometimes the length of follow-up in the clinical trial may be too short for the purposes of economic evaluation, as this tends to use long-term end points such as survival. The problem of length of follow-up is equally relevant for both costs and benefits. In some cases an increase in length of follow-up in a clinical trial by a number of months may make a lot of sense. For example, although it is common in trials of thrombolytic therapy to record 30-day mortality, most major trials, such as the GUSTO study, incorporate 1-year follow-up.

In other fields, such as lowering cholesterol levels, data on final outcomes such as all-cause mortality may take years to obtain. Here modeling studies have been undertaken, making projections of long-term outcomes from short-term trial data relating to intermediate end points, such as percentage reduction in cholesterol. Therefore, the problem of short-term follow-up is compounded by the use of an intermediate end point. The wisdom of this approach depends on the validity of the hypothesis linking intermediate and final outcomes. In at least 1 case, projections based on short-term evidence turned out to be wrong. Schulman et al[21] concluded that early use of zidovudine therapy in asymptomatic individuals with human immunodeficiency virus infection was cost-effective based on projections of disease progression from a clinical trial with 1-year follow-up. However, a subsequent study with 3-year follow-up showed that the advantages of therapy in the first year were eroded in subsequent years.[22] The authors also called into question the uncritical use of CD4 cell counts as a surrogate end point for assessment of benefit from long-term antiviral therapy.

Where long-term evidence is lacking, economists are in a quandary, particularly where the treatment concerned is already in use. Do they say nothing at all, or undertake a modeling study that may help the decision maker understand the likely range of cost-effectiveness outcomes? The same problem confronts the user of economic evaluation results. Should a decision be postponed until definitive data are available, or should an interim policy be formulated, pending further results?

Of the 2 thrombolysis studies discussed here, the one by Mark et al[2] was undertaken concurrently with the clinical trial, whereas that by Kalish et al[3] is a modeling study using the GUSTO trial results as its main source of clinical evidence. Therefore, the cost-effectiveness results are likely to be more similar than in a situation, for example, where the modeling study draws on clinical data from a number of different sources.

The main methodological difference between the 2 studies is that the resource consumption (eg, days in hospital, number of outpatient visits) in the study by Mark et al[2] are those actually observed during the trial. By contrast, the estimates in the study by Kalish et al[3] are drawn from other sources, although the probabilities of resource-consuming events (eg, coronary artery bypass surgery) are taken from the GUSTO trial.

Finally, it should be noted that by using observational databases, both articles extrapolated survival data beyond the 1 year observed in the trial. This reaffirms the point that, even when good quality clinical data are available, modeling is often necessary to conduct an economic evaluation.

**Were Costs Measured Accurately?**—While the viewpoint determines the relevant range of costs and outcomes to be included in an economic evaluation, there are many issues relating to their measurement and evaluation. First, it is useful to report the physical quantities of resources consumed or released by the treatments separately from their prices or unit costs. Not only does this allow us to scrutinize the method of assigning monetary values to resources, it also helps us to interpret the results of a study from one setting to another, as prices are known to vary by location.

Second, there are different approaches to valuing costs or cost savings. One approach is to use published charges. However, charges may differ from real costs, depending on the sophistication of accounting systems and the relative

bargaining power of health care institutions and third-party payers.[23] Where there is a systematic deviation between costs and charges, the analyst may adjust the latter by a *cost-to-charge ratio.* However, very little is currently known about how charges differ from costs, so simple adjustments may not suffice. From the third-party payer's perspective, charges will bear some relation to the amounts actually paid, although in some settings payments vary by payer. From a societal perspective we would like the real costs, since these reflect what society is forgoing, in benefits elsewhere, to provide a given treatment.

For example, Cohen et al[24] compared costs and charges for conventional angioplasty, directional coronary atherectomy, intracoronary stenting, and bypass surgery. Previous studies had suggested that total hospital charges for directional coronary atherectomy or intracoronary stenting are significantly higher than those for conventional angioplasty. However, when costs were examined, by adjusting itemized patient accounts by department-specific cost-to-charge ratios, it was found that the in-hospital costs of angioplasty and directional coronary atherectomy were similar. Also, although the cost of coronary stenting was approximately $2500 higher than that of conventional angioplasty, the magnitude of this difference was smaller than the $6300 increment previously suggested on the basis of analysis of hospital charges. The implication is that we may be deterred from using coronary atherectomy or stenting because of the high cost, whereas this may be an artifact of hospital accounting systems or bargaining power, rather than a reflection of the real value to society of the resources consumed by those procedures.

Mark et al[2] use costs from the Duke Transition One cost-accounting system, Medicare diagnosis related group (DRG) reimbursement rates, and Medicare physicians' fees in their estimations. Since the costs of the thrombolytic agents are an important component of the analysis, drug costs are calculated in 2 ways: from the *Drug Topics Red Book* average of 1993 wholesale prices,[25] and from the average costs of the drugs in 16 randomly selected GUSTO hospitals. The impact on cost-effectiveness of the different estimation methods is examined. Kalish et al[3] used medication costs and Medicare DRG reimbursement rates for 1 hospital. They took costs of treating serious hemorrhage and the costs of managing coronary artery disease and stroke from the literature.

**Were Data on Costs and Outcomes Appropriately Integrated?**—When making comparisons between alternatives in terms of cost per life year gained or cost per QALY gained, it is important to compute the incremental cost-effectiveness ratio of one therapy over another. This is because the most relevant information for the decision maker relates to the extra benefit that would be gained compared with any extra cost. Of course, if one therapy is dominated by another, having both higher benefits and lower costs, then the incremental comparison is not needed. In this case both articles calculate the incremental cost per life year or QALY gained from the use of t-PA, compared with streptokinase.

One important point to note about incremental analysis is that the incremental cost-effectiveness ratio of a given intervention is critically dependent on the comparison made. The most relevant comparison is current care, which could include doing nothing where this is ethically defensible. In the example discussed here, most would argue that streptokinase is the appropriate comparison and that doing nothing is not really an option. Where there are multiple interventions, each of which could be delivered at different scales or intensities, the ranking of options becomes quite complex.[26]

A final issue in the measurement and valuation of costs and consequences relates to the adjustment for differences in their timing. It is normally assumed that we prefer benefits sooner and prefer to postpone costs because of uncertainty about the future and because resources, if invested, usually yield a positive return. The accepted way of allowing for this in economic evaluations is to discount costs and benefits occurring in the future to present values.[12] The effect of this is to assign a lower weight in the analysis to costs and benefits occurring in the future. An annual discount rate of 5% is common in the published literature, although this choice is not necessarily theoretically or empirically justified. There are also debates about whether health outcomes should be discounted at the same rate as costs.[27,28]

In both studies considered here, the authors discount costs and benefits occurring in the future at a rate of 5% per year. Mark et al[2] also report results for discount rates of 0% and 10%, whereas Kalish et al[3] report results for rates of 1% and 10%.

### Was Appropriate Allowance Made for Uncertainties in the Analysis?

Uncertainty in economic evaluation can arise either from lack of precision in estimation or from methodological controversy. The conventional way of allowing for uncertainty in economic analyses is to undertake a sensitivity analysis (discussed in an earlier guide[10]) where the estimates for key variables are altered to assess what impact they have on study results.

In addition, conducting economic evaluations concurrently with clinical trials provides the opportunity to apply conventional tests of statistical significance to the resource quantities or costs.[20] Also, where measurements from a clinical trial inform us of the distribution of cost variables, it is possible to set the range of estimates for sensitivity analysis in relation to the statistical properties of the distribution (eg, 2 SDs from the mean). This raises a number of important issues, such as the size of the "economically important difference" when comparing the cost or cost-effectiveness of 2 alternatives, and the appropriateness of, and methods for, statistical tests on cost-effectiveness ratios.

Both articles report extensive sensitivity analyses, many of which relate to different methodological choices (eg, source of cost estimates) rather than to observed variability in the data. Mark et al[2] use the 95% CI for the increase in 1-year survival to explore the possible range in cost per life year saved. They also perform statistical tests for differences in cost but not for differences in cost-effectiveness ratios.

Because economic evaluation methods are in their infancy compared with those for randomized trials, investigators still debate many issues.[30] We've already mentioned one major issue: the appropriateness of alternative methods for valuing outcomes. Other issues relate to the appropriateness of considering some types of outcome (such as the costs of lost production if individuals are away from work because of illness) or the choice of discount rate. Some methodological uncertainties can be taken into account by sensitivity analysis (eg, if the choice of discount rate does not affect the choice of strategy in a given situation, then this particular controversy, though important, may not be critical to the decision).

The other way in which methodological uncertainties can be accommodated is in the reporting and discussion of results. Economists are often criticized for failing to reach a firm conclusion, but if the result is truly equivocal, that information will be important for the decision maker. It is important to remember that economic evaluation is no more than an aid to decision making, since there are often many difficult value judgments in reaching a decision.

### Are Estimates of Costs and Outcomes Related to the Baseline Risk in the Treatment Population?

Finally, we must recognize that in clinical practice the costs and outcomes of treatment are likely to be related to the baseline risk in the treatment population. For example, the cost-effectiveness of drug therapy for elevated cholesterol level, compared with no treatment, will depend on age, sex, pretreatment cholesterol level, and other risk factors; the greater the patients' risk, the lower the cost per unit of benefit.[31]

Division of patients into risk categories is common in clinical practice. In a study of the cost-effectiveness of β-blockers after AMI, Goldman et al[32] found that the cost per life year gained was $2400 for those patients at high risk, compared with $13 000 for those at low risk. The differences in the cost-effectiveness ratios were driven primarily by the patient's ability to benefit from therapy, rather than treatment cost.

Both articles investigate the impact of patient age on cost-effectiveness, as older patients have a higher mortality risk and fewer years of life left to live. In addition, Mark et al[2] investigate the impact of infarction location on the cost-effectiveness estimates.

In this article we have outlined some of the threats to validity in economic evaluations. In the next article on economic analysis, we will show you how to determine the results and how to use them in your practice.

#### References


1. The GUSTO Investigators. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. *N Engl J Med.* 1993;329:673-682.
2. Mark DB, Hlatky MA, Califf RM, et al. Cost-effectiveness of thrombolytic therapy with tissue plasminogen activator as compared with streptokinase for acute myocardial infarction. *N Engl J Med.* 1995;332:1418-1424.
3. Kalish SC, Gurwitz JH, Krumholz HM, Avorn J. Cost-effectiveness of model of thrombolytic therapy for acute myocardial infarction. *J Gen Intern Med.* 1995;10:321-330.
4. Goel V, Naylor CD. Potential cost-effectiveness of intravenous tissue plasminogen activator versus streptokinase for acute myocardial infarction. *Can J Cardiol.* 1992;8:31-38.
5. Eisenberg JM. Clinical economics: a guide to the economic analysis of clinical practices. *JAMA.* 1989; 262:2879-2886.
6. Detsky AS, Naglie IG. A clinician's guide to cost-effectiveness analysis. *Ann Intern Med.* 1990; 113:147-154.
7. Elixhauser A, Luce BR, Taylor WR, Reblando J. Health care CBA/CEA: an update on the growth and composition of the literature. *Med Care.* 1993; 31(suppl):JS1-JS11, JS18-JS149.
8. Backhouse ME, Backhouse RJ, Edey SA. Economic evaluation bibliography. *Health Econ.* 1992; 1(suppl):1-236.
9. Russell LB, Gold MR, Seigel JE, Daniels N, Weinstein MC, for the Panel on Cost-Effectiveness in Health and Medicine. The role of cost-effectiveness analysis in health and medicine. *JAMA.* 1996; 276:1172-1177.
10. Richardson WS, Detsky AS, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VII: how to use a clinical decision analysis, A: are the results of the study valid? *JAMA.* 1995;273:1292-1295.
11. Wilson MC, Hayward RSA, Tunis SR, Bass EB, Guyatt G, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VIII: how to use clinical practice guidelines, B: what are the recommendations and will they help you in caring for your patient? *JAMA.* 1995;274:1630-1632.
12. Drummond MF, Stoddart GL, Torrance GW. *Methods for the Economic Evaluation of Health Care Programmes.* Oxford, England: Oxford University Press; 1987.
13. Luce BR, Elixhauser A. *Standards for Socioeconomic Evaluation of Health Care Products and Services.* Berlin, Germany: Springer Verlag; 1990.
14. Kamlet MS. *A Framework for Cost-Utility Analysis of Government Health Care Programs: Report to the Office of Disease Prevention and Health Promotion.* Washington, DC: Public Health Service, US Dept of Health and Human Services in cooperation with the Foundation of Health Services Research; 1990.
15. Topol EJ, Burek K, O'Neill WW, et al. A randomized controlled trial of hospital discharge three days after myocardial infarction in the era of reperfusion. *N Engl J Med.* 1988;318:1083-1088.
16. Weisbrod BA, Test MA, Stein LI. Alternative to mental hospital treatment, II: economic benefit-cost analysis. *Arch Gen Psychiatry.* 1980;37:400-405.
17. Mehrez A, Gafni A. Quality-adjusted life years, utility theory, and healthy-years equivalents. *Med Decis Making.* 1989;9:142-149.
18. Mugford M, Kingford J, Chalmers I. Reducing the incidence of infection after caesarian section: implications of prophylaxis with antibiotics for hospital resources. *BMJ.* 1989;299:1003-1006.
19. Hillman AL, Bloom BS. Economic effects of prophylactic use of misoprostol to prevent gastric ulcer in patients taking nonsteroidal anti-inflammatory drugs. *Arch Intern Med.* 1989;149:2061-2065.
20. Graham DY, Agrawal NM, Roth SH. Prevention of NSAID-induced gastric ulcer with the synthetic prostaglandin misoprostol: a multicentre, double-blind, placebo-controlled trial. *Lancet.* 1988; 2:1277-1280.
21. Schulman KA, Lynn LA, Glick HA, Eisenberg JM. Cost-effectiveness of low dose zidovudine therapy for asymptomatic patients with human immunodeficiency virus (HIV) infection. *Ann Intern Med.* 1991;114:798-802.
22. Concorde Coordinating Committee. Concorde: MRC/ANRS randomised double-blind controlled trial of immediate and deferred zidovudine in symptom-free HIV infection. *Lancet.* 1994;343:871-881.
23. Finkler SA. The distinction between cost and charges. *Ann Intern Med.* 1982;96:102-109.
24. Cohen DJ, Breall JA, Kalon KLH, et al. Economics of elective coronary revascularization: comparison of costs and charges for conventional angioplasty, directional atherectomy, stenting and bypass surgery. *J Am Coll Cardiol.* 1993;22:1052-1059.
25. Sifton DW. *1993 Red Book: Pharmacy's Fundamental Reference.* Montvale, NJ: Medical Economics Data; 1993.
26. Karlsson G, Johannesson M. The decision rules of cost-effectiveness analysis. *Pharmaco Econ.* 1996; 9:113-120.
27. Parsonage M, Neuberger H. Discounting and health benefits. *Health Econ.* 1992;1:71-76.
28. Cairns J. Discounting and health benefits: another perspective. *Health Econ.* 1992;1:76-79.
29. O'Brien BJ, Drummond MF, Labelle RJ, Wilian A. In search of power and significance: issues in the design and analysis of stochastic economic appraisals. *Med Care.* 1994;32:150-163.
30. Udvarhelyi IS, Colditz GA, Rai A, Epstein MA. Cost-effectiveness and cost-benefit analysis in the medical literature: are the methods being used correctly? *Ann Intern Med.* 1992;116:238-244.
31. Epstein MA. Cost-effectiveness of antihyperlipidemic therapy in the prevention of coronary heart disease: the case of cholestyramine. *JAMA.* 1987;258:2381-2387.
32. Goldman L, Sia STB, Cook EF, Rutherford JD, Weinstein MC. Costs and effectiveness of routine therapy with long-term beta-adrenergic antagonists after acute myocardial infarction. *N Engl J Med.* 1988;319:152-157.

# Users' Guides to the Medical Literature

## XIII. How to Use an Article on Economic Analysis of Clinical Practice

## B. What Are the Results and Will They Help Me in Caring for My Patients?

Bernie J. O'Brien, PhD; Daren Heyland, MD; W. Scott Richardson, MD; Mitchell Levine, MD; Michael F. Drummond, PhD; for the Evidence-Based Medicine Working Group

### CLINICAL SCENARIO

You recall from the first of our 2 articles[1] concerning economic analysis of clinical practice that your chief of medicine has asked you to review relevant economic evidence from the literature and report to the hospital's pharmacy and therapeutics committee, which is trying to decide on formulary guidelines for the use of streptokinase and tissue-type plasminogen activator (t-PA) in the treatment of acute myocardial infarction (AMI). Your literature search identified 2 recent key cost-effectiveness studies: an analysis of economic data collected prospectively as part of the Global Utilization of Streptokinase and Tissue Plasminogen Activator for Occluded Coronary Arteries (GUSTO) trial[2] of streptokinase vs t-PA by Mark et al,[3] and a decision-analytic model by Kalish et al.[4] In the first article of this 2-part series we showed you how to evaluate the validity of the different economic appraisal study methods. In this article, we will show you how to interpret the results of an economic evaluation and how to examine the applicability of such data to your local practice setting and patients. We will do so by applying the Users' Guides to economic analysis of clinical practice in Table 1 to both studies.

### WHAT ARE THE RESULTS?

#### What Were the Incremental Costs and Outcomes of Each Strategy?

Let us start with the incremental costs. Look in the text and tables for the listings of all the costs considered for each treatment option and remember that costs are the product of the quantity of a resource used and its unit price. These should include the costs incurred to produce the treatment such as the physician's time, nurse's time, materials, and the like—what we might term the *up-front costs*, as well as the *downstream costs*, which refer to resources consumed in the future and are associated with clinical events that are attributable to the therapy. The study by Mark et al[3] quantifies resources used by treatment group in 3 periods of time over 1 year: initial hospitalization, discharge to 6 months, and 6 months to 1 year. Both treatment groups were very similar in their use of hospital resources over the year; both experienced a mean length of stay of 8 days, of which 3.5 days were in the intensive care unit. Both groups had the same rate of coronary artery bypass graft (CABG) surgery (13%) and percutaneous transluminal coronary angioplasty (PTCA) (31%) on initial hospitalization. As summarized in Table 2, the 1-year health care costs, excluding the thrombolytic agent, were $24 990 per patient treated with t-PA, and $24 575 per patient treated with streptokinase. As is clear from Table 2, the main cost difference between the 2 groups is the cost of the thrombolytic drugs themselves: $270 for t-PA and $320 for streptokinase. The overall difference in cost between patients treated with t-PA and patients treated with streptokinase is therefore our incremental cost at $2845 over the first year. This is discounted at 5% per year for a final figure of $2760. The authors argue that there is no cost difference between the 2 groups after 1 year. These data for incremental costs for t-PA are very similar to those estimated by Kalish et al,[4] who found a difference of $2535 in the use of t-PA to treat AMI in preference to streptokinase.

The measure of effectiveness chosen in the study by Mark et al[3] is the gain in life expectancy associated with t-PA. The available follow-up experience was to 1 year, with 89.9% surviving in the streptokinase group vs 91.1% in the t-PA group ($P<.001$). To translate these observations into life expectancy gains, the authors project survival curves for another 30 years or more using first a 14-year AMI survivorship database from

From the Department of Clinical Epidemiology and Biostatistics, McMaster University, and Centre for Evaluation of Medicines, St Joseph's Hospital, Hamilton, Ontario (Drs O'Brien and Levine); Royal Alexandra Hospital, Edmonton, Alberta (Dr Heyland); Department of Medicine, University of Rochester, School of Medicine and Dentistry, Rochester, NY (Dr Richardson); and the Centre for Health Economics, University of York, York, United Kingdom (Dr Drummond).

The original list of members (with affiliations) appears in the first article of this series (*JAMA*. 1993;270:2093-2095). A list of new members appears in the 10th article of the series (*JAMA*. 1996;275:1435-1439). The following members contributed to this article: Gordon H. Guyatt, MD, MSc (chair); Roman Jaeschke, MD, MSc; Deborah J. Cook, MD, MSc, Herlzel Gerstein, MD, MSc; Stephen Walter, PhD; John Williams, Jr, MD, MHS; and C. David Naylor, MD, MSc, DPhil.

Reprints: Gordon H. Guyatt, MD, MSc, McMaster University Health Sciences Centre, 1200 Main St W, Room 2C12, Hamilton, Ontario, Canada L8N 3Z5.

Users' Guides to the Medical Literature section editor: Drummond Rennie, MD, Deputy Editor (West), *JAMA*.

Table 1.—Users' Guides for Economic Analysis of Clinical Practice

**Are the results valid?**
- Did the analysis provide a full economic comparison of health care strategies?
- Were the costs and outcomes properly measured and valued?
- Was appropriate allowance made for uncertainties in the analysis?
- Are estimates of costs and outcomes related to the baseline risk in the treatment population?

**What were the results?**
- What were the incremental costs and outcomes of each strategy?
- Do incremental costs and outcomes differ between subgroups?
- How much does allowance for uncertainty change the results?

**Will the results help in caring for my patients?**
- Are the treatment benefits worth the harms and costs?
- Could my patients expect similar health outcomes?
- Could I expect similar costs?

Table 2.—Costs, Effects, and Cost-effectiveness Summary for Tissue-type Plasminogen Activator (t-PA) vs Streptokinase From Mark et al[3]

| | Treatment Group | | Difference (t-PA−Streptokinase) | Difference Discounted at 5% per Year |
|---|---|---|---|---|
| | t-PA | Streptokinase | | |
| Costs, in US$ | | | | |
| Health care costs for 1 y (excluding thrombolytic)* | 24 990 | 24 575 | 415 | . . . |
| Thrombolytic drug cost | 2750 | 320 | 2430 | . . . |
| Total 1-year cost | 27 740 | 24 895 | 2845 | 2709.6 (=ΔC)† |
| Effects | | | | |
| Life expectancy, y | 15.41 | 15.27 | 0.14 | 0.029 (=ΔE)† |
| Incremental cost-effectiveness of t-PA | . . . | . . . | . . . | ΔC/ΔE=$32 678 per life year gained |

*Treatment groups assumed to have no cost differences beyond 1 year.
†These discounted differences were not reported in the article, but have been imputed. ΔC indicates incremental cost, and ΔE, incremental effect. Ratio differs due to rounding error.

Duke University and then an assumption that remaining survivorship will follow a statistical distribution known as Gompertz. Having projected 2 survival curves, the authors calculate the area under each curve, which represents the expected value of survival time or life expectancy. For patients receiving t-PA, life expectancy was 15.41 years and 15.27 years for patients receiving streptokinase. As summarized in Table 2, the difference in life expectancy is 0.14 year per patient; or phrased another way, for every 100 patients treated with t-PA in preference to streptokinase, we would expect to gain 14 years of life.

In other situations, quantifying incremental effectiveness may be more difficult. Not all treatments change survival, and those that do not may affect different dimensions of health in many ways. For example, drug treatment of asymptomatic hypertension may result in short-term health reductions from drug adverse effects, in exchange for long-term expected health improvements, such as reduced risk of strokes. Note that in our t-PA example the outcome is not unambiguously restricted to survival benefit because there is a small but statistically significant increased risk of nonfatal hemorrhagic stroke associated with t-PA.[2] The existence of trade-offs between different aspects of health, for example between length of life vs quality of life, means that to arrive at a summary measure of net effectiveness, we must implicitly or explicitly weight the "desirability" of different outcomes relative to each other.

There is a large and growing literature on quantitative approaches for combining multiple health outcomes into a single metric using patient preferences.[5] Foremost among current practice is the construction of quality-adjusted life-years (QALYs) as a measure that captures the impact of therapies in the 2 broad domains of survival and quality of life. (QALYs were described in more detail earlier in this series.[6,7]) For economic appraisal, the added attraction of the QALY is that it provides decision makers with outcomes data that can be compared across diseases and treatments (eg, thrombolytic therapy for AMI vs nonsteroidal anti-inflammatory drugs [NSAIDs] for arthritis) as well as within a given therapy area. However, the QALY approach is not without criticism and some authors have proposed an alternative preference-weighted outcome measure known as *healthy years equivalents*.[8]

Both cost-effectiveness studies attempt to apply utility weights to estimate QALYs; the study by Mark et al[3] calculates QALYs as a secondary analysis using preference weights measured in the trial, and the study by Kalish et al[4] calculates QALYs as the primary outcome using values from the literature. Both studies conclude that, under plausible preference weights for nonfatal outcomes, the overall cost-effectiveness estimates are robust.

In summary, both studies use the efficacy data from the GUSTO trial as their starting point to conclude that t-PA treatment is more costly than streptokinase treatment, but that it provides an increase in survival (quality-adjusted or otherwise). The next calculation in both studies is to determine the incremental cost-effectiveness ratio for t-PA. This is illustrated using the data from the study by Mark et al[3] in Table 2. After discounting future costs and effects at 5% per year to reflect time preference (for the rationale, see our first article[1]), the difference (t-PA minus streptokinase) in cost per patient over the year (and by extension into the future because they assume no cost differences beyond 1 year) is $2709.60, which is divided by the difference in life expectancy per patient (0.029) to yield a ratio of $32 678 per year of life gained. A simple interpretation of this ratio is that it is the "price" at which we are buying additional years of life by using t-PA in preference to streptokinase; the lower this price, the more attractive is the use of t-PA. The study by Kalish et al[4] reaches a similar incremental cost-effectiveness ratio (with their adjusted denominator of QALYs and using the 30-day risk reduction GUSTO data) of $30 300 per QALY. These are the main results of the studies; we will discuss their interpretation later in this article.

## Do Incremental Costs and Outcomes Differ Between Subgroups?

In an editorial accompanying the GUSTO economic analysis, Lee[9] stresses that "cost-effectiveness should focus on strategies, not drugs. The cost-effectiveness of t-PA depends on how the drug is administered and to whom it is given." The first point relates mainly to the fact that the GUSTO trial had a protocol for accelerated administration of t-PA; slower regimens of administration of the same drug had previously shown no clinical advantage.[10] The second point is that because some patients (eg, the elderly) have a greater prior risk of mortality, the t-PA treatment effect will likely yield a higher absolute risk reduction in mortality.[2]

This second point has important implications for cost-effectiveness as can be seen in Table 3, which presents cost per life-year estimates among 8 subgroups on the basis of infarction site and patient age. Because the baseline risk of mortality in AMI varies by age and infarct site, the mortality benefit from treatment with t-PA also varies, and it is clear from Table 3 that t-PA is more cost-effective in older patients with anterior infarcts. To take the extreme cases, the cost per life-year gained in a person aged 40 years or younger with an inferior infarct is $203 071, compared with a person aged 75 years or older with an anterior infarct at only $13 410 per life-year gained.

In reviewing these studies you decide that the variation in yield per dollar ex-

Table 3.—Incremental Cost-effectiveness of Tissue-type Plasminogen Activator vs Streptokinase in Patient Subgroups From the Global Utilization of Streptokinase and Tissue Plasminogen Activator for Occluded Coronary Arteries (GUSTO)*

| | Cost (in $) Per Life-Year Gained by Age Subgroup, y | | | |
|---|---|---|---|---|
| | ≤40 | 41-60 | 61-75 | >75 |
| Inferior myocardial infarction | 203 071 | 74 816 | 27 873 | 16 246 |
| Anterior myocardial infarction | 123 609 | 49 877 | 20 601 | 13 410 |

*Data from the GUSTO Investigators.[2] Table adapted from Mark et al.[3]

pended may have some important implications for your pharmacy and therapeutics committee decision, because they wish to use t-PA only in selected patients.

### How Much Does Allowance for Uncertainty Change the Results?

Both t-PA cost-effectiveness studies explore uncertainty using sensitivity analysis, examining the impact on incremental cost-effectiveness of alternative values for uncertain variables. (One-way and multi-way sensitivity analysis was described in detail in the Users' Guides on decision analysis.[6,7])

A useful starting point for a sensitivity analysis is to examine the impact of variation in the effectiveness measure on the cost-effectiveness estimates. Where effectiveness is based on clinical trial data, the analyst does not have to make an additional judgment about the plausible range over which to vary the data, but can use a conventional measure of precision around a treatment effect such as the 95% confidence interval (CI). Using data from the study by Mark et al,[3] we know the t-PA treatment effect was a 1.1% increase in 1-year survivorship with a 95% CI of 0.46% to 1.74%. Applying this variation to the denominator of the incremental cost-effectiveness ratio, Mark et al[3] report a range of $71 039 per life-year gained to $18 781 around their baseline estimate of $32 678, with smaller benefit yielding a higher ratio. Both studies conclude that their estimates of cost-effectiveness are most sensitive to uncertainty in the magnitude of mortality benefit. It should be noted, however, that this form of analysis only partially captures the uncertainty in the cost-effectiveness ratio because it assumes the numerator (cost) does not vary. Investigators are currently developing more formal procedures for estimating CIs for cost-effectiveness ratios that permit the numerator and denominator to vary.[11]

## WILL THE RESULTS HELP IN CARING FOR MY PATIENTS?

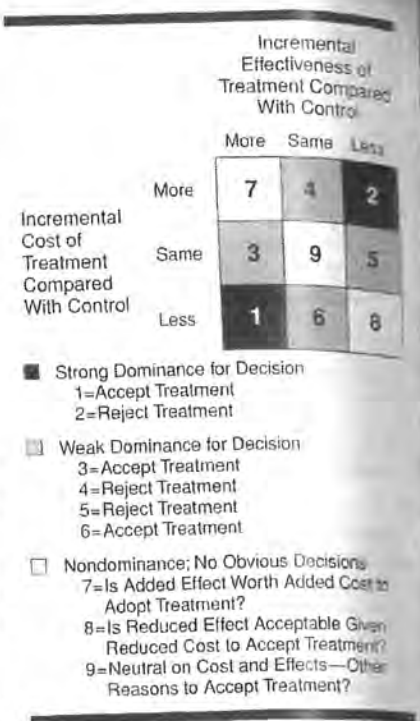
Having established the results of the 2 economic studies and the precision of the estimates, we now turn to 2 important issues of interpretation. The first issue is how incremental cost-effectiveness ratios can be interpreted to help in decision making, and the second issue is the extent to which the cost and/or effects from the study can be applied to your practice setting.

### Are the Treatment Benefits Worth the Harms and Costs?

In the Figure we present a simple framework for categorizing economic study results when data on incremental costs and effects have been determined. This 3×3 matrix has 9 cells to categorize studies depending on whether the new treatment is more, the same, or less costly than the control and whether it has more, the same, or less effectiveness.

In category 1, the new treatment is both less costly and more effective than the control, so the new treatment is said to be strongly *dominant*. For example, treatment to eradicate *Helicobacter pylori* for duodenal ulcer is strongly dominant over acid suppression with an $H_2$-receptor antagonist because it is less costly and results in fewer recurrences of ulcer over a 1-year period.[12] Category 2 represents strong dominance to reject a new therapy where the costs are higher and the effectiveness is worse than the control. Then follow 4 cases of so-called weak dominance where one of either costs or effectiveness is equivalent between the 2 therapies: category 3 indicating weak dominance to accept the treatment (equivalent cost but better effectiveness) and category 4 indicating weak dominance to reject the treatment (greater cost with equivalent effectiveness). By analogy, categories 5 and 6 indicate weak dominance to reject and accept, respectively.

All the shaded cells in the Figure indicate comparative cost and effectiveness combinations that provide evidence of strong or weak dominance. To inform decision making, no further analysis, such as calculation of cost-effectiveness ratios, is required for these shaded cells. However, further analysis is needed if results fall into the nondominance unshaded cells of 7, 8, or 9. First, it may arise that the treatment is associated with no statistically significant or clinically important difference in either effectiveness or costs, although it should be noted that the process of implementation and change of programs will generate costs not captured in the analysis.




Nine possible outcomes arising in the comparison of treatment control in terms of incremental cost and incremental effectiveness.

The most common nondominance circumstance is category 7, where the new therapy offers additional effectiveness but at an increased cost (or its mirror image in category 8). Both t-PA studies in our example fall into category 7. In this circumstance, as undertaken by both our t-PA studies, it is useful to calculate the incremental cost-effectiveness ratios of the new therapy as we discussed above and illustrated in Table 2.

Having estimated the incremental cost-effectiveness of t-PA over streptokinase and assuming for the moment that these data apply to your practice setting, how do you decide whether approximately $33 000 is an acceptable price to pay for saving 1 additional year of life? The first important point to note is that this question involves a value judgment and cannot be resolved by the analyst using only the study data. As noted in the conclusion of the GUSTO economic analysis, the study data can inform the decision but cannot make the choice. Some appeal must be made to external criteria to ascertain whether a jurisdiction or society is willing to pay this price for this improvement in outcome.

There are a number of approaches to the interpretation of incremental cost-effectiveness ratios. In an ideal world of complete information we would have data indicating the health outcomes that would be forgoing from other interventions and programs, within and

health care, not funded as a consequence of using t-PA. This is what economists refer to as *opportunity cost*. However, data to accomplish this task are very limited and investigators have promulgated a variety of second-best interpretive strategies. One approach assumes that previous decisions to adopt new medical therapies of known cost-effectiveness reveal an underlying set of values with which to judge the acceptability of the current treatment candidate. Our 2 t-PA cost-effectiveness studies both use this interpretive strategy to assess their $30 000 per life-year estimates: both cite the cost-effectiveness of 2 to 3 other interventions, some nonCardiac, that are currently funded and both conclude that an acceptable cost-effectiveness threshold would be $50 000 per QALY gained (for Kalish et al[4]) and per life-year gained (for Mark et al[8]).

Investigators have debated the validity of such interpretive strategies for incremental cost-effectiveness ratios at both theoretical[13,14] and practical levels.[15] For example, Johannesson and Weinstein[13] maintain that prioritizing resource allocations among health care programs based on rank orderings of interventions by incremental cost-effectiveness does lead to an efficient allocation of resources, in the sense that we are getting the greatest health yield for the resources expended. However, Birch and Gafni[14] contend that this is only the case where 2 assumptions hold true; programs exhibit *constant returns to scale* and are perfectly *divisible*. What do these 2 terms mean? *Constant returns to scale* implies a linear relationship between costs and outcomes at different levels of production; in many cases this may not hold true because we observe economies of scale, an example being the regionalization of cardiac surgery in 1 center where high volume can produce lower cost per case and often better clinical outcomes. *Divisibility of programs* implies that we can reallocate $1 or $1000 to t-PA and purchase benefits at the same rate implied by the cost-effectiveness ratio; this divisibility does not hold because to treat 1 additional patient with t-PA would require a block of resources equal, at least, to the cost of t-PA. While this methodologic debate continues, Drummond et al[15] caution readers about the practical problems of comparisons between cost-effectiveness studies that may have used very different methods, data, and assumptions.

In summary, you should exercise caution when drawing conclusions from incremental cost-effectiveness ratios. The ultimate criterion is one of local opportunity cost: what are the health benefits you will no longer realize if resources are expended on t-PA? The practical difficulty of applying this criterion is that many existing programs or services currently provided may not have been evaluated and so the opportunity cost of reducing or removing them is unknown or speculative.

### Could My Patients Expect Similar Health Outcomes?

After understanding the results, you should now turn to whether they will apply to your own practice setting. There are 2 levels of applicability for economic appraisal to the local setting. The first is the extent to which the evidence from the clinical trial(s) that forms the basis for the estimated treatment effect can be applied to routine clinical practice in any jurisdiction. A distinction is sometimes made between the efficacy of a treatment—as observed in a highly selected and compliant clinical trial population—and its effectiveness in the real world. For economic evidence to be relevant to policy decisions we would prefer evidence to be more related to effectiveness than efficacy. The second aspect is the extent to which the observed effect and cost data are transferable between jurisdictions. Threats to the transferability of cost-effectiveness data include variation in clinical practice patterns and variation in the prices of health care resources.

The applicability of clinical data to populations other than those studied was previously discussed in our Users' Guide on therapy or prevention.[16] To assess whether patients in your setting can expect the same health outcomes, you must examine 2 factors: (1) Are the patients in the study similar to my patients? (2) Is the clinical management of the study patients similar to my local practice? If your patients meet the inclusion and exclusion criteria of the primary article(s) for effectiveness used in the economic evaluation, then there is little difficulty in passing judgment that the patients are indeed similar. In many circumstances your patients may not be a perfect replicate of the study population, and then you should proceed by considering whether there are reasons to suppose your patients will respond differently to treatment than those included in the study. If the analysis is based on patients different from yours, check the subgroup and sensitivity analyses to see if relevant clinical variables were examined to permit extrapolation to your patients. Note that both of our economic studies used effectiveness data from the GUSTO trial,[2] which was a large, simple trial where the inclusion and exclusion criteria were sufficiently broad and likely to reflect the mix of patients presenting with AMI in many local settings.

Next, determine if the intervention is, or would be, used in the same way in your community. Local deviation from the observed patient management in the trial can have implications for generalizing both costs and outcomes from the study to the local setting. With respect to outcomes the key question is whether practice differs with respect to factors that will influence the magnitude of the treatment effect. First, let us consider whether these data apply to nonstudy hospitals in the United States. Kalish et al[4] doubt whether the efficacy data from the GUSTO trial are good predictors of effectiveness in routine practice:

> It has been questioned whether the results achieved in the GUSTO trial are possible in actual practice, largely due to the small time delay between symptom onset and treatment in this trial.[11,17] The benefit of tPA in the GUSTO trial was seen primarily among patients treated within four hours of symptom onset,[2] and the majority of patients who have AMI in the United States are not treated within four hours.[18]

Another issue is whether the GUSTO efficacy data are applicable to centers outside the United States. The GUSTO trial enrolled patients from 15 different countries; the majority of these patients (56%) were recruited from the United States. Patients from the United States were managed differently from non-US patients in a number of ways, including greater use of invasive revascularization such as PTCA and CABG, and greater use of nonprotocol medications such as antiarrhythmics and calcium antagonists.[19] Statistical analysis by logistic regression reveals that although mortality reduction with accelerated t-PA vs streptokinase was greater in the United States (1.2% absolute decrease vs 0.7% elsewhere), the test for treatment-by-country interaction against streptokinase was not significant ($P$=.30). In other words, if the truth were that there was no difference between the United States and other countries, differences equal to or greater than 1.2% vs 0.7% would be found in 30% of similar trials. Thus, while the results do not exclude a difference in effect between countries, neither do they provide substantial support for this hypothesis.

### Could I Expect Similar Costs?

In considering the transferability of cost (and cost-effectiveness) estimates between jurisdictions, it is useful to remember that the cost of a treatment is the summation of the product of physical resources consumed (eg, drugs, tests) and their unit prices. Cost data may not transfer well between jurisdictions for 2 reasons: (1) clinical practice patterns vary in such a way that resource consumption

associated with the treatment differs from that reported in the study and (2) local prices for resources differ from those used in the study. To address these points a good economic evaluation should report resource use and prices separately so that a reader can ascertain whether practice patterns and prices apply to their jurisdiction. The economic analysis by Mark et al[3] gives detailed reporting of resources and prices so the reader can judge whether, for example, the 73% rate of cardiac catheterization, 31% rate of PTCA, and 13% rate of CABG are applicable to their institution.

As previously noted, the GUSTO economic analysis is undertaken only on a sample of the US patients from the multinational trial, and the intensity of resource use was lower in other countries. Such resource use differences reflect a number of factors including availability of resources and financial incentives to health care providers. For example, the length of hospital stay was significantly lower in US hospitals than non-US hospitals (8 vs 10 days; $P<.001$) despite a greater incidence of complications among US patients. This difference likely reflects downward pressure exerted on length of stay in the United States by the prospective payment system to hospitals based on diagnosis related groups.

Variation in the prices of health care resources can threaten the validity of cross-jurisdictional inferences about cost-effectiveness. The problem is not due to variation in overall price levels between countries, but variation in the price of one health care input relative to another (ie, relative prices). For example, in a cost-effectiveness study of misoprostol as prophylaxis against gastrointestinal events in persons taking NSAIDs for arthritis, Drummond et al[20] found that among 4 countries compared, the price of misoprostol was highest in the United States but, surprisingly, the cost-effectiveness analysis was most favorable in the United States, indicating that prophylaxis actually reduced costs. This result is explained largely by different prices for health care resources because the use of misoprostol reduced the risk of surgery, the relative price of which was highest in the United States. The results of the GUSTO economic analysis[3] are clearly dependent on the relative prices of t-PA and streptokinase. Furthermore, we know that these relative drug prices vary between countries. For example, if the drug costs were those typical in Europe (approximately $1000 for 100 mg of t-PA and $200 for 1.5 million units of streptokinase), the cost-effectiveness ratio would be $13 943 per year of life saved.

Finally, it should be recognized that countries may differ with respect to the value they place on health benefits vs other commodities. There is no reason why $50 000 per life-year as an acceptable cost-effectiveness threshold for the United States is applicable to, for example, a less-industrialized country where the opportunity cost of such resources will be much higher. Countries vary in their willingness to pay for health and health care as evidenced by the varying proportions of gross national product they devote to the latter.

## RESOLUTION OF THE SCENARIO

Returning to our scenario and referring to the framework in the Figure, both t-PA cost-effectiveness studies indicate that t-PA is not dominant over streptokinase but falls into category c, implying that a trade-off between increased effectiveness at increased cost needs to be resolved. Since the effectiveness, resource use, and price data are applicable to your hospital, you inform the committee that the analyses you have reviewed can help inform their decision, but they must make the choice and decide what cost-effectiveness threshold is acceptable. You help frame this choice as one of local opportunity cost; by diverting resources to t-PA, what health benefits will be forgone from other treatments or programs no longer being funded? The committee decides that universal use of t-PA in all AMI cases will be very costly and divert resources from other health-producing programs in the hospital (although the benefits of these programs have not been as clearly documented as the new program!). They decide that t-PA should be used selectively based on the cost-effectiveness evidence in Table 3 and adopting the cutpoint of $50 000 per life-year suggested by Mark et al.[3] The committee decides that the preferred clinical strategy in their hospital is streptokinase in patients younger than 60 years with an inferior infarct and patients 60 years or younger with an anterior infarct; all other patients would receive t-PA.

### References


1. Drummond MF, Richardson WS, O'Brien BJ, et al, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, XIII: how to use an article on economic analysis of clinical practice, A: are the results of the study valid? *JAMA*. 1997;277:1552-1557.
2. The GUSTO Investigators. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. *N Engl J Med*. 1993;329:673-682.
3. Mark DB, Hlatky MA, Califf RM, et al. Cost-effectiveness of thrombolytic therapy with tissue plasminogen activator as compared with streptokinase for acute myocardial infarction. *N Engl J Med*. 1995;332:1418-1424.
4. Kalish SC, Gurwitz JH, Krumholz HM, Avorn J. Cost-effectiveness of model of thrombolytic therapy for acute myocardial infarction. *J Gen Intern Med*. 1995;10:321-330.
5. Torrance GW, Feeny D. Utilities and quality-adjusted life years. *Int J Technol Assess Health Care*. 1989;5:559-575.
6. Richardson WS, Detsky AS, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VII: how to use a clinical decision analysis, A: are the results of the study valid? *JAMA*. 1995;273:1292-1295.
7. Richardson WS, Detsky AS, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VII: how to use a clinical decision analysis, B: what are the results and will they help me in caring for my patients? *JAMA*. 1995;273:1610-1613.
8. Mehrez A, Gafni A. The healthy-years equivalents: how to measure them using the standard gamble approach. *Med Decis Making*. 1991;11:140-146.
9. Lee TH. Cost-effectiveness of tissue plasminogen activator. *N Engl J Med*. 1995;332:1443-1444.
10. Ridker PM, O'Donnell C, Marder VT, Hennekens CH. Large-scale trials of thrombolytic therapy for acute myocardial infarction: GISS1-2, ISIS-3, and GUSTO-1. *Ann Intern Med*. 1993;119:530-532.
11. O'Brien BJ, Drummond MF, Labelle RJ, Willan A. In search of power and significance: issues in design and analysis of stochastic cost-effectiveness studies in health care. *Med Care*. 1994;32:150-163.
12. O'Brien BJ, Goeree R, Mohamed AH, Hunt R. Cost-effectiveness of *Helicobacter pylori* eradication for the long-term management of duodenal ulcer in Canada. *Arch Intern Med*. 1995;155:1958-1964.
13. Johannesson M, Weinstein MC. On the decision rules of cost-effectiveness analysis. *J Health Econ*. 1993;12:459-467.
14. Birch S, Gafni A. Changing the problem to fit the solution: Johannesson and Weinstein's (1993) application of economics to real world problems. *J Health Econ*. 1993;12:469-476.
15. Drummond MF, Torrance GW, Mason J. Cost-effectiveness league tables: more harm than good? *Soc Sci Med*. 1993;37:33-40.
16. Guyatt GH, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group. Users' Guides to the Medical Literature, II: how to use an article about therapy or prevention, B: what were the results and will they help me in caring for my patient? *JAMA*. 1994;271:59-63.
17. Ridker PM, O'Donnell C, Marder V, Hennekens CH. A response to 'Holding GUSTO up to the light.' *Ann Intern Med*. 1994;120:882-885.
18. Ridker PM, Manson J, Goldhaber SZ, Hennekens CH, Buring JE. Comparison of delay time to hospital presentation for physicians and non-physicians with acute myocardial infarction. *Am J Cardiol*. 1992;45:505-512.
19. Van de Werf F, Topol EJ, Lee KL, et al. Variations in patient management and outcomes for acute myocardial infarction in the United States and other countries. *JAMA*. 1995;273:1586-1591.
20. Drummond MF, Bloom BS, Carrion G, et al. Issues in the cross national assessment of health technology. *Int J Technol Assess Health Care*. 1992;8:671-682.

# Users' Guides to the Medical Literature

## XIV. How to Decide on the Applicability of Clinical Trial Results to Your Patient

Antonio L. Dans, MD; Leonila F. Dans, MD; Gordon H. Guyatt, MD, MSc; Scott Richardson, MD; for the Evidence-Based Medicine Working Group

### CLINICAL SCENARIO

You are the attending physician on duty when a poor, 45-year-old man presents to the emergency department of a general hospital in the Philippines. He has severe chest pain for 2 hours, associated with clammy perspiration. Physical examination reveals a blood pressure of 110/70 mm Hg, a pulse rate of 92 beats per minute, a normal first heart sound, and clear lungs. An electrocardiogram discloses 3-mm ST-segment elevation in the inferior leads. As intravenous lines are placed, and the patient is prepared for admission to the coronary care department, you consider whether you should offer this patient a thrombolytic agent. Though your response is that the impecunious patient cannot afford the treatment, you ponder the right course of action in a richer patient. As your duty ends that night, you resolve to prepare for the next patient admitted for an acute myocardial infarction (MI) by retrieving the best evidence on the use of thrombolytics.

### THE SEARCH

Streptokinase is the only thrombolytic agent that your patients might afford. You, therefore, confine your search to this drug, trying to locate the best trial or, if possible, a meta-analysis. Using Grateful Med software (National Library of Medicine, Bethesda, Md), you select *myocardial infarction* from the list of medical subject headings used to index articles. On the second subject line, you use the term *streptokinase*. You limit your search to English-language articles, and to find quantitative reviews or original studies, you use the term *meta-analysis* or *randomized controlled trial* as the publication type.

You retrieve a systematic meta-analysis of randomized trials that deal only with effectiveness[1] and not toxicity. You, therefore, also review a single trial from ISIS-2 Collaborative Group[2] that you choose on the basis of its size (17 000 patients), strong design (including double-blinding), and the wide variety of settings in which the study was undertaken. You refer to earlier Users' Guides to evaluate the validity of the studies,[3,4] as well as the magnitude and precision of the treatment effects and toxicity.[5] The articles pass the validity criteria, and the treatment reduced the event rate from 17.4% to 12.8%.[1] This outweighs the potential harm of "bleeds requiring transfusion," which occurred in 0.5% of patients treated with streptokinase compared with 0.2% in the placebo group.[2]

An answer does not come easily to the last question: "How can you apply the results to your patients?" Asians constituted a small minority of the patients in the trials, and you are uncertain about your hospital staff's ability to cope with technical requirements for administering the drug or dealing with any complications.

As clinicians look more often to randomized controlled trials (RCTs) to guide their clinical care, they must decide how to apply RCT results to individual patients in their practice setting. This Users' Guide addresses the issue of applicability, which involves the implications of the trial results for patient care. Applicability is closely related to concepts of generalizability and external validity, but is broader in its scope, including issues related to the overall impact of treatment in individual patients. In considering applicability, clinicians first must decide whether the biology of the treatment effect will be similar in patients they are facing; second, their patients' risk of a target event, which the treatment is designed to prevent; third, the adverse effects that may accompany treatment; and fourth, their own ability to deliver the intervention in a safe and effective manner.[6] Clinicians managing patients who differ economically, racially, and culturally from those recruited in typical clinical trials face particular challenges in addressing applicability. Such patients include those from the inner cities of North America, the Native American reservations, or less industrialized countries. Clinicians seeing these patients cannot afford to repeat every trial simply because of doubts regarding applicability. The end result is that applicability becomes a fait accompli—an issue that may often be ignored rather than confronted.

Earlier in this series, we addressed the applicability problem in the Users' Guide for articles about therapy or prevention: "A better approach than rigidly applying the study's inclusion and exclusion criteria is to ask whether there are compelling reasons why the results should not be applied to the patient. A compelling reason usually won't be found, and most often you can generalize the results to your patient with confidence."[3]

From the Departments of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario (Dr Guyatt); the Departments of Internal Medicine (Dr A. L. Dans) and Pediatrics (Dr L. F. Dans), University of the Philippines College of Medicine, Manila; and the Department of Medicine, University of Texas Health Sciences Center, San Antonio (Dr Richardson).

The original list of members (with affiliations) appears in the first article of this series (*JAMA*. 1993;270:2093-2095). A list of new members appears in the 10th article of the series (*JAMA*. 1996;275:1435-1439). The following members contributed to this article: Deborah J. Cook, MD, MSc; Hertzel Gerstein, MD, MSc; Ann Holbrook, MD, PharmD, MSc; Les Irwig MBBCh, PhD; Virginia Moyer, MD, MPH; and Thomas B. Newman, MD.

Reprints: Gordon H. Guyatt, MD, MSc, McMaster University Health Sciences Centre, 1200 Main St W, Room 2C12, Hamilton, Ontario, Canada L8N 3Z5.

Users' Guides to the Medical Literature section editor: Drummond Rennie, MD, Deputy Editor (West), *JAMA*.

Table 1.—The Guides

| Issues |
|---|
| Biologic |
| (1) Are there pathophysiologic differences in the illness under study that may lead to a diminished treatment response? |
| (2) Are there patient differences that may diminish the treatment response? |
| Social and economic |
| (3) Are there important differences in patient compliance that may diminish the treatment response? |
| (4) Are there important differences in provider compliance that may diminish the treatment response? |
| Epidemiologic |
| (5) Do my patients have comorbid conditions that significantly alter the potential benefits and risks of the treatment? |
| (6) Are there important differences in untreated patients' risk of adverse outcomes that might alter the efficiency of treatment? |

Physicians may encounter problems following this advice. We didn't give a good definition of a "compelling reason" or provide guidelines on how to systematically address the question. In this article, we correct these deficiencies by presenting a set of guidelines for evaluating the applicability of the results of RCTs to populations other than the participants. We present the guides as questions that probe for situations when clinicians may be forced to reject applicability. We phrase the questions so that a "yes" answer will lead clinicians to suspect a problem of applicability. Table 1 summarizes the guides, categorizing them into biologic issues (which help us decide if the treatment can work), socioeconomic issues (which help us decide if the treatment will work), and epidemiologic issues (which help us decide how efficient the treatment will be). As we discuss each issue, we will offer sources of information that will help physicians answer their questions.

## THE GUIDES—BIOLOGIC ISSUES

### Are There Pathophysiologic Differences in the Illness Under Study That May Lead to a Diminished Treatment Response?

Diseases with a single name may represent conditions with important pathophysiologic differences. These differences can sometimes lead to diminished treatment responses due to divergence in pathogenetic mechanisms or biological differences in the causative agent. Hypertension in blacks, which has been observed to be relatively responsive to diuretics and unresponsive to β-blockers,[7] provides an example of the former. This selective response reflects a state of relative volume excess that investigators now theorize may have served protective functions in their hot and arid ancestral environments.[8]

Malaria provides an example of a condition that may vary because of biological differences in the causative agent. Malaria treatment protocols vary depending on drug resistance patterns.[9] In these examples, clinicians should anticipate variation in response to treatment and should temper hasty conclusions regarding the applicability of trial results.

#### Sources of Evidence

Sources of information regarding disease pathophysiology in populations include basic and laboratory studies, animal studies, genetic studies, and observational studies documenting pathologic changes in affected individuals and evaluating the biology of causative agents (eg, surveys on drug resistance patterns of infectious diseases).[10] In some cases, variation in response to treatment may be the first clue to a difference in pathophysiology. This was the case in the example of hypertension in blacks.

To address our scenario of applicability of streptokinase to the treatment of MI in the Philippines, we reviewed a case series of autopsies performed on Filipino patients who had MI.[11] Pathologic changes in the coronary arteries and myocardium were similar to those noted among North Americans,[12] while nonatherosclerotic causes of coronary disease were rare. Clinical surveys have demonstrated that Filipinos share the same risk factors for coronary disease[13] as North Americans.[14] Thus, we can be confident that disease pathogenesis is similar.

### Are There Patient Differences That May Diminish the Treatment Response?

Between-population differences in response to treatment may arise from differences in drug metabolism, immune response, or environmental factors that affect drug toxicity. Differences in drug metabolism may directly influence the efficacy of a treatment regimen. If they are not identified, slow metabolizers of a drug could face the risk of greater toxic effects, while a significant decrease in efficacy might occur in rapid metabolizers. Such differences are usually based on genetic polymorphism in the activity of metabolizing enzymes. A well-known example is hepatic *N*-acetyltransferase, an enzyme with increased activity among Asians.[15] For this reason, clinicians offer higher drug dosages for agents such as isoniazid, hydralazine, and procainamide hydrochloride. Other examples of genetic polymorphism include pseudocholinesterase activity in the metabolism of suxamethonium and glucose-6-phosphate dehydrogenase activity in the metabolism of sulfonamides and other drugs.[16]

Differences in patients' immune response may also modulate treatment effect. *Haemophilus influenzae* vaccine, for example, has a lower efficacy in Alaskan natives than in nonnative populations.[17] Finally, environmental factors may affect response to therapy. For instance, the incidence of thyroid dysfunction from amiodarone differs in low vs high iodine environments.[18]

#### Sources of Evidence

Pharmacokinetic and bioavailability studies are important sources of evidence regarding differences in treatment response. Such studies generally require small sample sizes and commonly available equipment. Unfortunately, for a wide variety of drugs, technology for assays remains unavailable. Reasonable alternatives include dose-ranging and descriptive studies of patients receiving treatment, which can also provide information on immune response to vaccines and environmental factors that may increase or decrease the toxic effects of drugs. Postmarketing surveillance studies and large RCTs require large sample sizes and long-term follow-up, but (as in the example of the decreased effect of *H influenzae* vaccine in Alaskan natives) may provide definitive information about differential response to therapy.

Although we found no studies evaluating the pharmacokinetic profile of streptokinase when given to Filipinos, postmarketing studies show that Filipinos experience the same reperfusion arrhythmias and bleeding complications when given streptokinase at the same dose as North Americans.[19] These studies provide some assurance of similarities in the response to adverse effects of treatment.

## SOCIAL AND ECONOMIC ISSUES

When satisfied that biologic differences do not compromise treatment applicability, clinicians must examine constraints related to the social environment that may diminish treatment effectiveness.

### Are There Important Differences in Patient Compliance That May Diminish the Treatment Response?

To the extent that groups of people exhibit different compliance with treatment, clinicians may expect variation in treatment effectiveness. Variability in compliance between populations may stem from resource limitations in a particular setting or less obvious attitudinal or behavioral idiosyncrasies. Both types of problems may, for example, affect the safety of outpatient administration of anticoagulants. Neither indigent pa-

tients nor their society may be able to afford repeated clinic visits and tests for treatment monitoring. Alcoholic patients, whatever their financial situation, may be less likely to comply with monitoring. Inadequate monitoring, whatever the reason, increases bleeding risk from overanticoagulation, shifting the balance between benefit and harm (even to the point where harm outweighs benefit).

### Sources of Evidence

While clinicians perform poorly at untutored guessing of patient compliance, a systematic examination of compliance in individual patients, or groups of patients, is likely to aid in identifying varying compliance patterns. Clinicians may also refer to more general sources of evidence, such as sociologic descriptions of attitudes of specific groups of people. In the Philippines, an attitude called *bahala na* connotes a lack of capacity or will to control one's fate.[20] A near equivalent would go something like "let's just wait and see, there's really nothing much we can do about the situation." This external locus of control[21] may have an adverse effect on patient compliance. In our scenario, we don't expect patient compliance to be a problem since we give streptokinase intravenously as a single dose.

### Are There Important Differences in Provider Compliance That Might Diminish the Safety and Efficacy of the Treatment?

In this guide, provider compliance refers to a host of diagnostic tests, monitoring equipment, intervention requirements, and other technical specifications that clinicians must satisfy to safely and effectively administer a treatment. Financial conditions in a health care center, access to equipment, technologic expertise, and availability and skill of health personnel may influence treatment effectiveness. For instance, while carotid endarterectomy may benefit low-risk patients when surgery-associated stroke is low, the net effect for such patients in centers with higher surgery-associated stroke rates may be an increase in adverse outcomes.[22]

In less industrialized countries, many hospitals and clinics do not have easy access to sophisticated equipment, so problems of provider compliance are common. For example, while rheumatic atrial fibrillation remains a common problem in Asian countries, few laboratories in rural areas perform the tests necessary for titration of warfarin dose. This limitation is likely to reverse the critical balance between effectiveness and safety of treatment.

### Sources of Evidence

Because of experience regarding availability of equipment, laboratory tests, and health personnel resources, practitioners themselves are a good source of information regarding feasibility interventions. Clinicians' assessments can be supplemented by formal quality-of-care assessments and postmarketing surveillance of adverse effects. Whatever the source of information, a thorough understanding of the technical requirements for safe and effective administration should guide decisions regarding the ability to comply.

Administration of streptokinase carries potential hazards, foremost of which is catastrophic bleeding. Facilities for emergency administration of cryoprecipitate, fresh frozen plasma, or whole blood must be available.[23] In hospitals without efficient blood banking systems, it may be difficult to cope with bleeding emergencies. This increases the potential hazards of treatment and may tip the balance between benefit and harm.

## EPIDEMIOLOGIC ISSUES

When satisfied that biologic, social, or economic differences do not compromise applicability, the clinician must examine the patient's characteristics that can influence either the magnitude of the benefit or the risks of treatment (and thus, the trade-off between the 2).[24] The last 2 guides address these issues.

### Do My Patients Have Comorbid Conditions That Significantly Alter the Potential Benefits and Risks of the Treatment?

The presence of other conditions in a particular locality may affect treatment efficiency in 2 possible ways: competing diagnostic possibilities or competing causes of outcome. The management of pneumonia in developing countries provides an example of a competing diagnostic possibility.

The acute respiratory tract infection management protocol includes a symptom-driven algorithm for differentiating pneumonia from nonpneumonia. This protocol identifies children who need antibiotics and has proven effective in reducing mortality from pneumonia among children younger than 5 years.[25] Unfortunately, similarities exist in the clinical presentation of pneumonia and malaria. In malaria endemic areas, clinicians may expect an increase in false-positive "pneumonias." These patients with false-positive pneumonialike presentations will not respond to antibiotics for pneumonia, and a delay in instituting antimalarial treatment may result. If the drop in accuracy is large enough, the balance between harm and benefit will change. To resolve this issue, investigators have initiated a study to determine if the acute respiratory tract infection protocol can maintain its effectiveness in malaria endemic areas (S. P. Lupisan, unpublished data, 1998).

Competing causes of target events may also affect the magnitude of benefit. An example comes from the management of acute MI in some Filipino hospitals. A recent study disclosed 30 in-hospital deaths in a cohort of 149 patients admitted to a charity hospital (ISIP Study Group, unpublished data, 1996). On the basis of results from the meta-analysis, clinicians might expect streptokinase to reduce this 20% death rate by 25%.[1] However, a closer look at the local data shows a contrast with the original studies in which virtually all deaths were a direct result of cardiac ischemia. In the Philippine study, noncardiac causes (mostly pneumonia with sepsis) were responsible for 11 of the 30 deaths. Streptokinase will not reduce mortality in such patients. Adequate antibiotic coverage may result in a greater (and more economical) reduction in mortality for patients who develop pneumonia.

In addition to reducing benefit, other morbidity may affect the magnitude of risk. Surgical mortality may increase in malnourished patients, shifting the balance between benefit and risk. On occasion, other morbidity can also work in the opposite direction—increasing efficiency. For example, a patient with a large infarct, in whom the clinician is considering warfarin, may also have atrial fibrillation. Since anticoagulation reduces stroke risk in such patients, the presence of atrial fibrillation strengthens the indication or treatment.

### Sources of Evidence

Cohort studies provide the most reliable information on comorbid conditions. In the MI scenario, we used data from the local study of 149 charity patients to evaluate the impact of other morbid conditions.[21] As we noted, we can expect streptokinase to prevent around 5 of 19 cardiac deaths (but none of those from other causes), and the absolute reduction in all-cause mortality is a decline from 30 (20.1%) of 149 to 25 (16.8%) of 149.

### Are There Important Differences in Untreated Patients' Risk of Adverse Outcomes That Might Alter the Efficiency of Treatment?

In our Users' Guide on therapy, we addressed the relationship between a patient's risk of an adverse event and the magnitude of the treatment impact. Because the issue is so important in assessing applicability of trial results, we will review it in detail.

Table 2.—Baseline Mortality Rate Without Treatment and Estimated Number Needed to Treat or to Save 1 Life Using Streptokinase in Filipinos With Acute Myocardial Infarction, Tabulated According to Age and Wall Involvement

| Characteristics | Age <60 y | | Age ≥60 y | |
|---|---|---|---|---|
| | Mortality Rate | No. Needed to Treat | Mortality Rate | No. Needed to Treat |
| Wall involvement | | | | |
| Infarction | 0.02 | 179 | 0.13 | 27 |
| Non–Q-wave myocardial infarct | 0.04 | 89 | 0.18 | 23 |
| Anterolateral wall infarct | 0.05 | 71 | 0.19 | 20 |
| Massive anterior wall infarct | 0.14 | 26 | 0.23 | 16 |

In the therapy Users' Guide, we introduced the notion of number needed to treat (NNT). Thinking about NNT requires an understanding of the concepts of relative risk, relative risk reduction, and absolute risk reduction. Readers desiring a full discussion of these concepts can refer to the earlier article.[3] Because it estimates the number of patients who need to receive treatment (with implications about the associated toxic effects and cost) to prevent an adverse event, clinicians can use the NNT to consider a treatment's efficiency.

The NNT is the inverse of the absolute risk reduction resulting from a particular treatment in a particular group of patients. If a patient's risk without treatment is 20%, then we expect 20 of 100 patients without treatment to experience an adverse event. When we administer a treatment with a relative risk reduction of 10%, only 18 treated patients will experience adverse events. Thus, for every 100 patients treated, we prevent 2 events, and the NNT is 50. If the expected event rate in untreated patients is cut by half to 10%, and the relative risk reduction remains the same, in treating 100 patients, we will prevent only 1 adverse event, and the NNT will double to 100.

This reasoning, and much of what follows, assumes that relative risk reduction remains constant across subgroups. While testing this assumption can be difficult,[26] there are situations in which the assumption will fail, and clinicians should be alert to this possibility.[27,28] Fortunately, however, in most instances the assumption will not introduce important inaccuracies in the NNT.[29,30]

One source of difference in expected event rates is country of origin and residence. Keys[31] compared the 20-year incidence of coronary deaths in the United States, 5 European countries, and Japan.[31] He found an extremely low incidence of coronary death in the Japanese cohort, despite correction for baseline differences in recognized risk factors. Similar results have been observed in preliminary reports of the ongoing Multinational Monitoring of Cardiovascular Disease and Their Determinants project.[32] In this study, involving 39 centers from 26 countries, east Asians showed a much lower incidence of coronary death than their western counterparts. Age-standardized mortality rates for coronary heart disease were lowest among Japanese (40 of 100 000), and highest in North Ireland (414 of 100 000).

Thinking of the NNT, this 10-fold difference in incidence among the Japanese would translate to a 10-fold increase in the NNT for a drug preventing coronary deaths. This decrease in efficiency may warrant a reconsideration of applying the results of a trial to low-risk patients. We consider the issue of balancing costs and effects in our Users' Guides for determining a level of recommendation.[33]

### Sources of Evidence

Cohort studies on the course of disease in untreated patients can provide excellent risk data, and such studies are even more useful when they define subsets of patients at varying risk. Of 424 Filipinos with MI who were eligible for streptokinase (but in whom the drug was not administered) and who participated in a cohort study conducted in 9 centers in metropolitan Manila, 37 (11.1%) suffered cardiac death.[24] This provides a good estimate of the expected event rate. If streptokinase had been given, it would have prevented 25% of the deaths, reducing the absolute mortality rate to 8.3%. Thus, 2.8% of the those otherwise destined to die would have been spared (the absolute risk reduction), and the NNT is 100 divided by 2.8 or approximately 36 patients.

The expected event rates varied in the patient subpopulation.[24] Young patients with small infarcts had a much lower expected mortality (and thus much larger NNTs) than old patients with large infarcts. Using prognostic information from these various subgroups, we constructed Table 2, which shows the expected mortality according to age and left ventricular wall involvement and the corresponding NNT to save 1 life in each group. As the table shows, NNT can range from 16 (when treatment is applied to patients with a poor prognosis) to as much as 179 (when treatment is applied to patients with a good prognosis).

Varying patient risk will affect benefit of treatment no matter what the environment in which you practice. Even if you work in the Western tertiary care environment in which investigators conducted their original studies, you will still face high- and low-risk patients. The critical trade-off between risk and benefit may vary in these patient groups, mandating a different treatment decision.[30]

## COMMENT

These guides address the task of applying the results of clinical trials done on restricted, specially selected populations to other groups. Although inspired by the predicament in less industrialized countries, the guides are relevant to all situations where clinicians must make decisions regarding applicability. By breaking down the problem into specific questions, we have provided guides for busy clinicians who make daily attempts to strike a balance between making "unjustifiably broad generalizations and being too conservative in one's conclusions."[4]

When clinicians suspect limited applicability (ie, when a response of "yes" is encountered for any of the questions), what can they do? This will depend on whether the anticipated differences are important, and if important, whether they are remediable. For example, differences in disease pathophysiology (guide 1) do not always mean that applicability is limited. Management of a cataract, for instance, will probably be the same regardless of the cause. Differences in treatment response (guide 2) can sometimes be accommodated by altering administration of a treatment (such as adjusting the dose of a drug). Education, training, provision of necessary equipment, and other attempts at optimizing compliance may address problems in patient and provider compliance (guides 3 and 4).

For differences in comorbid conditions or expected target event rates (guides 5 and 6) the clinician's response will depend on the differences observed. If an increase in efficiency is anticipated (as when disease prognosis is worse or the incidence of an adverse outcome is greater), a recommendation to treat can be more easily accepted. A decrease in efficiency, on the other hand, should lead clinicians to be more cautious in accepting a treatment recommendation.

When the answer to 1 or more of the guide questions is "yes," and the differences noted are important and not easily remediable, clinicians should not assume that the trial results can be readily applied. In these instances, an additional RCT may be warranted.

## RESOLUTION OF THE SCENARIO

What should we recommend regarding thrombolytic use for the Filipino patient admitted for acute MI? There is no reason to believe that Filipinos have a different disease pathogenesis or a different response to treatment with thrombolytics (guides 1 and 2). Patient compliance will not be an important issue since the drug is given intravenously as a single dose (guide 3). The technical requirements for administration are often, but not always, available, and when they are not, the risks may outweigh the benefits of thrombolytic administration (guide 4).

Two issues remain to be resolved, both dealing with the magnitude of treatment impact. Pneumonia is an important comorbid condition, accounting for one third of deaths, at least in some charity hospitals (guide 5). However, rates of cardiac death are still sufficiently high (11.1%) that the relative risk reduction we can achieve with streptokinase (28%) will result in an NNT of 32 for the overall population (guide 6). For subgroups of patients, however, the NNT will range from 16 to 179, depending on the age and the size of the infarct (Table 2).

Should we recommend the routine use of streptokinase among Filipinos presenting with acute MI? The guides have brought us closer to an answer. We have confirmed applicability of the thrombolytic data on the effectiveness of streptokinase, but only in centers with adequate blood banking facilities. We have also refined estimates of treatment impact, based on knowledge of the course of disease among Filipinos. However, the cost of the drug is approximately $250 per treatment and in the Philippines the average annual per capita income is only $1300 (National Statistics Office, unpublished data, 1994).[31] These figures highlight the difficult economic trade-off associated with administering streptokinase.

The judgment about whether to give streptokinase will depend on who pays for the treatment (in the Philippines, usually the patients themselves), patient and family values, what resources are available (usually limited in our charity hospital setting), and the competing needs (for example, the need for antibiotics because of a high incidence of pneumonia, in turn a result of overcrowding in the hospital wards). For equally applicable treatments, our final decision may differ for a much less costly, but equally effective and applicable treatment, such as aspirin for our MI patient.

### References


1. Midgette AS, O'Connor GT, Baron JA, Bell J. Effect of intravenous steptokinase on early morality in patients with suspected acute myocardial infarction: a meta-analysis by anatomic location of infarction. *Ann Intern Med.* 1990;113:961-968.
2. ISIS-2 Collaborative Group. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17 187 cases of suspected acute myocardial infarction: ISIS-2. *Lancet.* 1988;2:349-360.
3. Oxman AD, Cook DJ, Guyatt GH, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VI: how to use an overview. *JAMA.* 1994;272:1367-1371.
4. Guyatt GH, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, II: how to use an article about therapy or prevention, A: are the results of the study valid? *JAMA.* 1993;270:2598-2601.
5. Guyatt GH, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, II: how to use an article about therapy or prevention, B: what were the results and will they help me in caring for my patients? *JAMA.* 1994;271:59-63.
6. Friedman LM, Furberg CD, DeMets DL. *Fundamentals of Clinical Trials.* 2nd ed. Little, Mass: PSG Publishing Co Inc; 1985.
7. Falkner B, Kushner H. Effect of chronic sodium loading on cardiovascular response in young blacks and whites. *Hypertension.* 1990;15:36.
8. Wilson TW. History of salt supplies in West Africa and blood pressure today. *Lancet.* 1986;1:784.
9. World Health Organization. World malaria situation in 1992, part 1. *Wkly Epidemiol Rec.* 1994;69:309-314.
10. Davis CE. Generalizing from clinical trials. *Control Clin Trials.* 1994;15:11-14.
11. Canlas MM, Dominguez AE, Abarquez RF. Ten-year review of the clinicopathologic findings of coronary artery disease at the University of the Philippines, Philippine General Hospital (1969-1978). *Phil J Int Med.* 1980;18:65-74.
12. Roberts WC, Potkin BN, Solus DE, et al. Mode of death, frequency of healed and acute myocardial infarction, number of major epicardial coronary arteries severely narrowed by atherosclerotic plague, and heart weight in fatal atherosclerotic coronary artery disease: analysis of 889 patients studied at necropsy. *J Am Coll Cardiol.* 1990;15:196-202.
13. Balgos AA, Lopez MB, delos Santos E, et al. The significance of risk factors in myocardial infarction—a 2-year retrospective study at the University of the Philippines, Philippine General Hospital. *Philippine J Cardiol.* 1984;12:104-108.
14. Farmer JA, Gotto AM. Dyslipidemia and other risk factors for coronary artery disease. In: Braunwald E, ed. *Heart Disease—A Textbook of Cardiovascular Medicine.* 5th ed. Philadelphia, Pa: WB Saunders Co; 1997:1126-1160.
15. Horai Y, Ishizaki T. Pharmacogenetics and its clinical implication: N-acetylation polymorphism. *Ration Drug Ther.* 1987;21:1-7.
16. Goodman GA, Rall TW, Nies AS, Taylor P. Principles of therapeutics. In: *The Pharmacologic Basis of Therapeutics.* 8th ed. New York, NY: Pergamon Press Inc; 1991:71-73.
17. Ward J, Brenneman G, Letson GW, Heyward WL. Limited efficacy of a *Haemophilus* type b conjugate vaccine in Alaska Native infants: the Alaska *H influenzae* Vaccine Study Group. *N Engl J Med.* 1990;323:1415-1416.
18. Martino E, Safran M, Aghini-Lombardi F, et al. Environmental iodine intake and thyroid dysfunction during chronic amiodarone therapy. *Ann Intern Med.* 1984;101:28-34.
19. Dela Paz AG, Pineda NE, Justiniani RP, et al. Thrombolysis in acute myocardial infarction. *Philippine J Cardiol.* 1988;17:185-188.
20. Bulatao J. *Split-Level Christianity.* Manila, Philippines: University of Sto Tomas Press; 1966.
21. Raja SN, Williams S, McGee R. Multidimensional health locus of control beliefs and psychological health for a sample of mothers. *Soc Sci Med.* 1994;39:213-220.
22. Barnett HJM, Eliasziw M, Meldrum HE, Taylor DW. Do the facts and figures warrant a 10-fold increase in the performance of carotid endarterectomy on asymptomatic patients? *Neurology.* 1996;46:603-608.
23. Gersh BJ, Opie LH. Antithrombotic agents: platelet inhibitors, anticoagulants and fibrinolytics. In: Opie LH, ed. *Drugs for the Heart.* 3rd ed. Philadelphia, Pa: WB Saunders Co; 1991.
24. Glasziou PP, Irwig LM. An evidence based approach to individualising treatment. *BMJ.* 1995;311:1356-1358.
25. Sazawal S, Black RE. Meta-analysis of intervention trials on case management of pneumonia in community settings. *Lancet.* 1992;340:528-533.
26. Sharp SJ, Thompson SG, Altman DG. The relation between treatment benefit and underlying risk in meta-analysis. *BMJ.* 1996;313:735-738.
27. Rothwell PM. Can overall results of clinical trials be applied to all patients? *Lancet.* 1995;345:1616-1619.
28. Bailey KR. Generalizing the results of randomized clinical trials. *Control Clin Trials.* 1994;15:15-23.
29. Oxman AD, Guyatt GH. A consumer's guide to subgroup analysis. *Ann Intern Med.* 1992;116:78-84.
30. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA.* 1991;266:93-98.
31. Keys A. *Seven Countries: A Multivariate Analysis of Death and Coronary Heart Disease.* Cambridge, Mass: Harvard University Press; 1980.
32. World Health Organization Monitoring of Cardiovascular Disease and Their Determiants (MONICA). WHO MONICA Project: assessing CHD mortality and morbidity. *Int J Epidemiol.* 1989;18(suppl 3, pt 1):S38-S45.
33. Guyatt GH, Sackett DL, Sinclair JC, Hayward R, et al, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, IX: a method for grading health care recommendations. *JAMA.* 1995;274:1800-1804.

# Users' Guides to the Medical Literature: XIV. How to Decide on the Applicability of Clinical Trial Results to Your Patient

Antonio L. Dans; Leonila F. Dans; Gordon H. Guyatt; et al.

*JAMA*. 1998;279(7):545-549 (doi:10.1001/jama.279.7.545)

http://jama.ama-assn.org/cgi/content/full/279/7/545

# Users' Guides to the Medical Literature

## XV. How to Use an Article About Disease Probability for Differential Diagnosis

W. Scott Richardson, MD
Mark C. Wilson, MD, MPH
Gordon H. Guyatt, MD, MSc
Deborah J. Cook, MD, MSc
James Nishikawa, MD
for the Evidence-Based Medicine Working Group

### CLINICAL SCENARIO

You are an experienced clinician working at a hospital emergency department. One morning, a 33-year-old man presents with palpitations. He describes the new onset of episodes of fast, regular chest pounding, which come on gradually, last 1 to 2 minutes, and occur several times a day. He reports no relation of symptoms to activities and no change in exercise tolerance. He is very anxious and tells you he fears heart disease. He has no other symptoms, no personal or family history of heart disease, and takes no medications. His heart rate is 90 bpm and regular, and physical examination of his eyes, thyroid gland, lungs, and heart is normal. His 12-lead electrocardiogram is normal, without arrhythmia or signs of pre-excitation.

You suspect his anxiety explains his palpitations, that they are mediated by hyperventilation, and are possibly part of a panic attack. While he has no findings of cardiac arrhythmia or hyperthyroidism, you wonder if these disorders are common enough in the emergency department setting to consider seriously. You reject pheochromocytoma as too unlikely. Thus, you can list causes of palpitations, but want more information about the frequency of these causes to choose a diagnostic work-up. You ask the question: "In patients presenting with palpitations, what is the frequency of underlying disorders?"

### THE SEARCH

Your emergency department computer networks with the medical library, where MEDLINE is on CD-ROM. In the MEDLINE file for current years, you enter 3 text words: *palpitations* (89 citations), *differential diagnosis* (7039 citations), and *cause or causes* (71 848 citations). You combine these sets, yielding 17 citations, including an article by Weber and Kapoor[1] that promises to have what you want.

---

Sick persons seldom present with the diagnosis already made; instead, they present with 1 or more symptoms. These symptoms prompt the clinician to gather information through history and physical examination, identifying clinical findings that suggest explanations for the symptom(s). For example, in an older woman presenting with generalized pruritis, the clinician could identify recent anorexia and weight loss, along with jaundice and the absence of a rash. For most symptoms, the clinician must consider multiple causes for the patient's findings.

*Differential diagnosis* is the method by which the clinician considers the possible causes of a patient's clinical findings before making a final diagnosis.[2,3] Experienced clinicians often group the findings into meaningful clusters, summarized in brief phrases about the symptom, body location, or organ system involved, such as "generalized pruritis," "painless jaundice," and "constitutional symptoms" for the older woman mentioned earlier. We call these clusters *clinical problems*[3,4] and include problems of biological, psychological, or sociological origin.[3] It is for these clinical problems, rather than for the final diagnosis, that the clinician selects a patient's differential diagnosis.

When considering a patient's differential diagnosis, how is the clinician to decide which disorders to pursue? If the clinician were to consider all known causes equally likely and test for them all simultaneously (the *possibilistic* ap-

**Author Affiliations:** Departments of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario (Drs Guyatt, Cook, and Nishikawa); the Department of Ambulatory Care, Audie L. Murphy Memorial Veterans Hospital and the Department of Medicine, University of Texas Health Sciences Center at San Antonio (Dr Richardson); and the Department of Medicine, Bowman Gray School of Medicine and Wake Forest University Baptist Medical Center, Winston-Salem, NC (Dr Wilson).

The original list of members (with affiliations) appears in the first article of this series (*JAMA*. 1993;270:2093-2095). A list of new members appears in the 10th article of the series (*JAMA*. 1996;275:1435-1439). The following members contributed to this article: Les Irwig, MBBCh, PhD; Virginia Moyer, MD, MPH; Thomas B. Newman, MD, MPH; David L. Sackett, MD, MSc; Jack Sinclair, MD; and John W. Williams, Jr, MD, MHS.

**Corresponding Author and Reprints:** Gordon H. Guyatt, MD, MSc, McMaster University Health Sciences Centre, 1200 Main St W, Room 2C12, Hamilton, Ontario, Canada, L8N 3Z5.

**Users' Guides to the Medical Literature Section Editor:** Drummond Rennie, MD, Deputy Editor (West), *JAMA*.

proach), then the patient would undergo unnecessary testing. Instead, the experienced clinician is selective, considering first those disorders that are more likely (a *probabilistic* approach), more serious if left undiagnosed and untreated (a *prognostic* approach), or more responsive to treatment if offered (a *pragmatic* approach). Prior articles in this series showed how to use evidence about prognosis[6] and therapy,[7-9] so this article will focus on using evidence about disease probability.

Wisely selecting a patient's differential diagnosis involves all 3 considerations (probabilistic, prognostic, and pragmatic), as depicted in TABLE 1. The clinician's single best explanation for the patient's clinical problem(s) can be termed the *leading hypothesis* or *working diagnosis* (group 1 in Table 1). A few (usually 1-5) other diagnoses, termed *active alternatives* (group 2 in Table 1), may be worth considering further at the time of the initial work-up because they are likely, or serious, or treatable. Additional causes of the clinical problem(s), termed *other hypotheses* (group 3 in Table 1), may be too unlikely to consider at the time of the initial diagnostic work-up, but remain possible and could be considered further if the working diagnosis and active alternatives are later disproved.

Using this framework for the patient with palpitations in the scenario, you consider anxiety the working diagnosis, and you wonder whether cardiac arrhythmias, hyperthyroidism, or pheochromocytoma are active alternatives (group 2 in Table 1) or other hypotheses (group 3 in Table 1).

Selecting a patient-specific differential diagnosis should strongly influence diagnostic testing. For the leading hypothesis, the clinician may choose to confirm the diagnosis, using a highly specific test with a high likelihood ratio for a positive result.[10,11] For the active alternatives, the clinician would choose to exclude these diagnoses, using highly sensitive tests with low likelihood ratios for negative results. Usually, the clinician would not order tests initially for the other hypotheses.

How can information about disease probability help clinicians select patients' differential diagnoses? We will illustrate with some brief cases. First, consider a patient who presents with a painful eruption of grouped vesicles in the distribution of a single dermatome. An experienced clinician would make a diagnosis of herpes zoster and turn to thinking about whether to offer the patient therapy. Using Table 1, the working diagnosis is zoster and there are no active alternatives. In other words, the probability of zoster is so high (near 1.0 or 100%) that it is above a threshold where no further testing is required.

Next, consider a previously healthy athlete who presents with lateral rib cage pain after being accidentally struck by an errant baseball pitch. Again, the experienced clinician might rapidly recognize the clinical problem (posttraumatic lateral chest pain), quickly list a leading hypothesis (rib contusion) and an active alternative (rib fracture), and plan a test (radiograph) to exclude the latter. If asked, the clinician could also list disorders that are too unlikely to consider further (such as myocardial infarction). In other words, while not as likely as rib contusion, the probability of a rib fracture is above the threshold for testing, while the probability of myocardial infarction is below the threshold for testing.

These cases illustrate how clinicians can estimate the probability of disease from the patient's clinical findings, risk factors, exposures, and the like, and then compare disease probabilities to 2 thresholds. The probability above which the diagnosis is sufficiently likely to warrant therapy defines the upper threshold. This threshold is termed the *test-treatment* or simply the *treatment threshold*.[12] In the case of shingles above, the clinician judged the diagnosis of zoster to be above this treatment threshold. The probability below which the clinician believes a diagnosis warrants no further consideration defines the lower threshold. This threshold is termed the *no test-test* or simply the *test threshold*. In the case of posttraumatic torso pain above, the diagnosis of rib fracture fell above, and the diagnosis of myocardial infarction fell below, the test threshold.

Clinicians begin with pretest estimates of disease probability and then adjust the probability as new diagnostic information arrives. Test results are useful when they move our pretest probabilities across 1 of these 2 thresholds. For a disorder with a pretest probability above the treatment threshold, a confirmatory test that raises the probability further would not aid diagnostically. On the other end of the scale, for a disorder with

**Table 1.** Selecting a Patient-Specific Differential Diagnosis

| Diagnostic Hypotheses | Description of Hypotheses | Implications for Choosing Diagnostic Tests* | Implications for Choosing Initial Therapy |
|---|---|---|---|
| (1) Leading hypothesis or "working diagnosis" | Single best overall explanation of this patient's problem(s) | Choose test(s) to confirm this disorder, emphasizing high specificity and LR+ much larger than 1 | Start initial therapy for this disorder, *unless* special circumstances exist |
| (2) Active alternatives or "rule outs" | Not as good as No. 1, but likely, serious, or treatable enough to be actively sought in this patient | Choose test(s) to exclude these disorders, emphasizing high sensitivity and LR− much smaller than 1 | Consider starting initial therapy for 1 or more of these, *if* special circumstances exist |
| (3) Other hypotheses | Not likely, serious, or treatable enough to be pursued at this point, but not yet excluded | Hold off on tests for these disorders at this point | Hold off on initial therapy of these disorders at this point |
| (4) Excluded hypotheses | Causes of the problem that have been disproved | No further tests necessary | No treatment necessary |

*LR+ indicates likelihood ratio for a positive result; LR−, likelihood ratio for a negative result.

**Table 2.** Users' Guides for Articles About Disease Probability for Differential Diagnosis

| |
|---|
| **Are the results valid?** |
| Primary guides |
| Did the study patients represent the full spectrum of those who present with this clinical problem? |
| Were the criteria for each final diagnosis explicit and credible? |
| Secondary guides |
| Was the diagnostic work-up comprehensive and consistently applied? |
| For initially undiagnosed patients, was follow-up sufficiently long and complete? |
| **What were the results?** |
| What were the diagnoses and their probabilities? |
| How precise are these estimates of disease probability? |
| **Will the results help in caring for my patients?** |
| Are the study patients similar to my own? |
| Is it unlikely that the disease possibilities or probabilities have changed since this evidence was gathered? |

a pretest probability below the test threshold, an exclusionary test that lowers the probability further would not aid diagnostically. When the clinician believes the pretest probability is high enough to test for and not high enough to begin treatment (ie, between the 2 thresholds), a test could be diagnostically useful if it moves the probability across either threshold.

Where can clinicians get pretest estimates of disease probability? We can use our memories of prior cases with the same clinical problem(s), so that disorders we have diagnosed frequently would have higher probability in the current patient's illness than diagnoses we have made less frequently. Remembered cases are always available and are calibrated to our local practices. Yet our memories are imperfect, and the probabilities we estimate based on them are subject to various biases and errors.[13-15]

Original research constitutes another source of information about disease probability. For example, in a study of diagnostic tests for anemia in the aged, investigators compared blood tests with bone marrow results in 259 elderly persons, finding iron deficiency in 94 patients (36%).[16] Thus, while this study focused on evaluating tests for iron deficiency, it also provides information about disease frequency. Some research studies focus more directly on the frequency of diseases that cause symptoms,[17] such as the article by Weber and Kapoor[1] on palpitations. This Users' Guide will help you understand direct studies of disease probability, judge their validity, and decide whether to use them for estimating pretest probability for your own patients (TABLE 2).

## THE GUIDES

### Are the Results Valid?

**Did the Study Patients Represent the Full Spectrum of Those Who Present With This Clinical Problem?** This question asks about 2 related issues. First, how do the investigators define the clinical problem? The definition of the clinical problem for study determines the population from which the study patients should be selected. Thus, investigators studying "hematuria" might include patients with microscopic and gross hematuria, with or without symptoms. On the other hand, investigators studying "asymptomatic, microscopic hematuria" would exclude those with symptoms or with gross hematuria.

Such differing definitions of the clinical problem can yield different frequencies of underlying diseases. Including patients with gross hematuria or urinary symptoms might raise the frequency of acute infection as the underlying cause relative to those without symptoms. So assessing the validity of an article about differential diagnosis begins with a search for a clear definition of the clinical problem.

Having defined the target population by clinical problem, investigators next assemble a patient sample. Ideally, the study sample mirrors the target population, so that the frequency of underlying diseases in the sample approximates that of the target population. Such a patient sample is termed *representative*, and the more representative the sample is, the more accurate the resulting disease probabilities. Investigators seldom are able to use the strongest method of ensuring representativeness, obtaining a random sample of the entire population. The next strongest methods are either to include all patients with the clinical problem from a defined geographic area, or to include a consecutive series of all patients with the clinical problem who receive care at the investigators' institution(s). To the extent that a nonconsecutive case series opens the study to the differential inclusion of patients with different underlying disorders, it compromises study validity.

You can judge the representativeness of the sample by examining the setting from which patients come. Patients with ostensibly the same clinical problem can present to different clinical settings, resulting in different services seeing different types of patients. Typically, patients in secondary or tertiary care settings have higher proportions of more serious diseases or more uncommon diseases than those patients seen in primary care.[18] For instance, in a study of coronary artery disease in 1074 patients with chest pain, a higher proportion of referral practice patients had coronary artery disease than the primary care practice patients, even in patients with similar clinical histories.[19]

To evaluate representativeness, you can also note the methods by which patients were recruited. By considering how investigators identified their patients, how they avoided missing patients, and who was included and who was excluded, you can judge whether important subgroups appear to be missing. The wider the spectrum of patients in the sample, the more representative the sample should be of the whole population, and therefore the more valid the results. For example, in a study of *Clostridium difficile* colitis in 609 patients with diarrhea, the patient sample consisted of adult inpatients whose diarrheal stools were tested for cytotoxin, thereby excluding any patients whose clinicians chose not to test.[20] The inclusion of only those tested is likely to raise the probability of *C difficile* in relation to the entire population of patients with diarrhea.

Weber and Kapoor[1] defined palpitations broadly, as any one of several patient complaints (eg, fast heartbeats, skipped heartbeats, and the like) and included patients with new and recurrent palpitations. They obtained patients from

3 clinical settings (emergency department, inpatient floors, and a medical clinic) in 1 university medical center in a middle-sized North American city. Of the 229 adult patients presenting consecutively for care of palpitations at their center during the study period, 39 refused participation, and the investigators included the remaining 190 patients, including 62 from the emergency department setting. No important subgroups appear to have been excluded, so these 190 patients probably represent the full spectrum of patients presenting with palpitations.

**Were the Criteria for Each Final Diagnosis Explicit and Credible?** Clinicians often disagree about a patient's diagnosis. Such disagreement about final diagnosis could threaten the validity of a study's conclusions about disease frequency. To minimize this threat, investigators can use a set of explicit criteria when assigning each of the final diagnoses. Ideally, these criteria should include not only the findings needed to confirm each diagnosis, but also those findings useful for rejecting each diagnosis. For example, published diagnostic criteria for infective endocarditis include criteria for both verifying the infection and for rejecting it.[21,22] Investigators can then classify patients into diagnostic groups that are mutually exclusive, with the exception of patients whose symptoms are from more than 1 cause. This allows clinicians to understand which diagnoses remain possible for any undiagnosed patients after the investigators have ruled out alternatives.

Ideally, studies of disease probability would also measure the agreement beyond chance for the clinicians assigning diagnoses, as was done in a study of the causes of dizziness.[23] The greater the agreement, the more reproducible and credible are the diagnostic assignments.

While reviewing the diagnostic criteria, keep in mind that "lesion finding" is not necessarily the same thing as "illness explaining." In other words, by using explicit and credible criteria, a patient may be found to have 2 or more disorders that might explain the clinical problem, causing some doubt as to which disorder is the culprit. Better studies of disease probability will include some assurance that the disorders found actually do explain the patients' illnesses. For example, in a sequence of studies of syncope, investigators required that the symptoms occur simultaneously with an arrhythmia before that arrhythmia was judged to be the cause.[24] Alternatively, in a study of chronic cough, investigators gave cause-specific therapy and required positive treatment responses to confirm the final diagnoses.[25]

Weber and Kapoor[1] developed a priori explicit and credible criteria for confirming each possible disorder causing palpitations and listed their criteria in an appendix along with supporting citations. They evaluated study patients prospectively and assigned final diagnoses using the explicit criteria. Wherever relevant, such as for cardiac arrhythmias, they required that the palpitations occur at the same time as the arrhythmias for that cause to be diagnosed. They do not report on agreement for these assignments.

**Was the Diagnostic Work-up Comprehensive and Consistently Applied?** This criterion addresses 2 closely related issues. First, have the investigators evaluated their patients thoroughly enough to detect any of the important causes of this clinical problem? Within reason and ethics, the more comprehensive the work-up, the lower the chance that invalid conclusions about disease frequency will be reached. For example, in a retrospective study of stroke in 127 patients with mental status changes, the diagnostic work-up was reported to include neurological examination and neuroimaging; a comprehensive search for other causes of delirium was not described, and 118 cases remained unexplained.[26] As the investigators do not describe a complete and systematic search for the causes of delirium, the disease probabilities appear less credible.

The second issue is how consistently the diagnostic work-up was applied. This does not mean that every patient must undergo every test. Instead, for many clinical problems, the clinician performs a detailed, yet focused, history; a problem-oriented examination of the involved organ systems; and a few initial tests. Then, depending on the clues discovered, further inquiry proceeds down one of multiple-branching pathways. Ideally, investigators would evaluate all patients with the same initial work-up, and then "follow the clues" using prespecified testing sequences. Once a definitive test result confirms a final diagnosis, then further confirmatory testing is unnecessary and unethical.

You may find it easy deciding whether the patients' illnesses have been well investigated if they were evaluated prospectively using a predetermined diagnostic approach. It becomes harder to judge when patients are studied only after their investigation is complete or when investigation is not standardized. For example, in a study of precipitating factors in 101 patients with decompensated heart failure, while all patients underwent history and physical examination, the lack of standardization of subsequent testing makes it difficult to judge the accuracy of the disease probabilities.[27]

Weber and Kapoor[1] evaluated their patients' palpitations prospectively using 2 principal means, a structured interview completed by one of the investigators, and the combined diagnostic evaluation (ie, history, examination, and testing) chosen by the physician seeing the patient at the index visit. In addition, all patients completed self-administered questionnaires designed to assist in detecting various psychiatric disorders. A majority of patients (166/190) had electrocardiograms, and large numbers had other testing for cardiac disease as well. Thus, the diagnostic work-up was reasonably comprehensive for common disease categories, although not exhaustive. Since the subsequent testing ordered by the physicians was not fully standardized, some inconsistency may have been introduced, although it does not appear likely to have distorted the probabilities of common disease categories such as psychiatric or cardiac causes.

**For Initially Undiagnosed Patients, Was Follow-up Sufficiently Long and Complete?** Even when investigators use explicit diagnostic criteria after a comprehensive evaluation that is consistently applied, some patients' clinical problems may remain unexplained. The higher the number of undiagnosed patients, the greater the chance of error in the estimates of disease probability. For example, in a retrospective study of various causes of dizziness in 1194 patients at an otolaryngology clinic, about 27% remained undiagnosed.[28] With more than a quarter of patients' illnesses unexplained, the disease probabilities for the overall sample might be inaccurate.

If the study evaluation leaves any patients undiagnosed, investigators can follow these patients over time, searching for additional clues leading to eventual diagnoses and observing the prognosis. The longer and more complete the follow-up, the greater our confidence in the benign nature of the condition in patients who remain undiagnosed yet unharmed at the study's end. How long is long enough? No answer would correctly fit all clinical problems, but we would suggest 1 to 6 months for symptoms that are acute and self-limited and 1 to 5 years for chronically recurring or progressive symptoms.

Weber and Kapoor[1] identified a diagnosable cause of palpitations in all but 31 (16.3%) of their 190 patients. The investigators followed nearly all of the study patients (96%) for at least a year, during which 1 additional diagnosis was made in those initially undiagnosed (symptomatic correlation with ventricular premature beats). None of the 31 undiagnosed patients had a stroke or died.

## What Were the Results?

**What Were the Diagnoses and Their Probabilities?** The authors of many studies of disease probability display the main results in a table listing the diagnoses made, and the numbers and percentages of patients with those diagnoses. For some symptoms, patients may have more than 1 disease coexisting and contributing to the clinical problem. In these situations, authors often identify the major diagnosis for such patients and separately tabulate contributing causes. Alternatively, authors sometimes identify a separate, "multiple-cause" group.

Weber and Kapoor[1] present a table that tells us that 58 patients (31%) were diagnosed as having psychiatric causes, 82 patients (43%) had cardiac disorders, while 5 patients (2.6%) were found to have thyrotoxicosis and none had pheochromocytoma. This distribution differed across clinical settings: for instance, cardiac disorders were more than twice as likely in patients presenting to the emergency department compared with patients presenting to the outpatient clinic.

**How Precise Are These Estimates of Disease Probability?** Even when valid, these disease probabilities are only estimates of the true frequencies. You can examine the precision of these estimates using their confidence intervals (CIs). If the authors do not provide the CIs for you, you can calculate them yourself using the following formula (for 95% CIs):

$$95\%\text{CI} = p \pm 1.96\sqrt{(p[1-p])/n}$$

where $p$ is the proportion of patients with the cause of interest, and $n$ is the number of patients in the sample.[29] This formula becomes inaccurate when the number of cases is 5 or fewer, and approximations are available for this situation.[30,31]

For instance, consider the category of psychiatric causes of palpitations in the study by Weber and Kapoor.[1] Using the above formula, we would start with $p = 0.31$, $(1 - p) = 0.69$, and $n = 190$. Working through the arithmetic, we find the CI to be $0.31 \pm 0.066$. Thus, while the most likely true proportion is 31%, it may range between 24.4% and 37.6%.

Whether you will consider the CIs sufficiently precise depends on where the estimated proportion and CI fall in relation to your test or treatment thresholds. If both the estimate and the entire 95% CI are on the same side of your threshold, then the result is precise enough to allow firm conclusions about disease probability for use in planning tests or treatments. Conversely, if the confidence limit around the estimate crosses your threshold, the result may not be precise enough for definitive conclusions about disease probability. You might still use a valid but imprecise probability result, while keeping in mind the uncertainty and what it might mean for testing or treatment.

Weber and Kapoor[1] do not provide the 95% CIs for the probabilities they found. However, as we just illustrated, if you were concerned about how near the probabilities were to your thresholds, you could calculate the 95% CIs yourself.

## Will the Results Help in Caring for My Patients?

**Are the Study Patients Similar to My Own?** This question concerns whether the clinical setting and patient characteristics are similar enough to yours to allow you to extrapolate the results to your practice. The closer the match, the more confident you can be in applying the results. We suggest you ask yourself whether the setting or patients are so different from yours that you should discard the results.[32] For instance, consider whether your patients come from areas where 1 or more of the underlying disorders are endemic, which could make these disorders much more likely in your patients than was found in the study. Also, consider whether your patients have different cultural patterns of illness behavior or health practices that might cause important differences in the disease probabilities when compared with the patients in the study.

Weber and Kapoor[1] recruited the 190 palpitation patients from those presenting to the outpatient clinics, the inpatient medical and surgical services, and the emergency department (62 of the 190 patients) in 1 university medical center in a middle-sized North American city. Thus, these patients are likely to be similar to the patients seen in your hospital emergency department, and you can use the study results to help inform the pretest probabilities for the patient in the scenario.

**Is It Unlikely That the Disease Possibilities or Probabilities Have Changed Since This Evidence Was Gathered?** As time passes, evidence about disease frequency can become obsolete. Old dis-

eases can be controlled or eliminated. New diseases can arise. Such events can so alter the spectrum of possible diseases or their likelihood that previously valid studies may no longer be applicable to current practice. For example, consider how much the arrival of human immunodeficiency virus disease has transformed the list of possibilities for such clinical problems as generalized lymphadenopathy, chronic diarrhea, and unexplained weight loss.

Similar changes can occur as the result of progress in medical science or public health. For instance, in studies of fever of unknown origin, new diagnostic technologies have substantially altered the proportions of patients with malignancy or unexplained fevers.[33-35] Treatment advances that improve survival, such as chemotherapy for childhood leukemia, can bring about shifts in disease likelihood because the treatment might cause complications, such as secondary malignancy years after cure of leukemia. Public health measures that control some diseases, such as cholera, can alter the likelihood of the remaining causes of the clinical problems that the prevented disease would have caused, in this example, acute diarrhea.

The palpitations study by Weber and Kapoor[1] was published in 1996, and the study period was in 1991. You know of no new developments likely to cause a change in the spectrum or probabilities of disease in patients with palpitations.

## RESOLUTION OF THE SCENARIO

Using the structure outlined in Table 1, your "leading hypothesis" is that acute anxiety is causing your patient's palpitations. You offer the patient a more in-depth discussion of his psychosocial situation as the next test to explore this diagnosis (ie, the pretest probability is above your test threshold). At the same time, you do not feel that anxiety is so certain that you can stop considering other disorders (ie, the pretest probability is below your threshold for treatment without testing). After reviewing the palpitations study by Weber and Kapoor,[1] you decide to include in your "active alternatives" some cardiac arrhythmias (as common, serious, and treatable) and hyperthyroidism (less common but serious and treatable), so you arrange testing to exclude these disorders (ie, these are above your test threshold). Finally, given that none of the 190 study patients had pheochromocytoma, and since your patient has none of the other clinical features of this disorder, you place it into your "other hypotheses" category (ie, below your test threshold) and decide to hold off on testing for this condition.

We recommend applying these Users' Guides to identify good evidence on which to base initial estimates of disease probability for use in differential diagnosis. As you apply this evidence, keep in mind that selecting a patient's differential diagnosis wisely includes not only considering how likely various disorders are, but also how serious are the various diseases if left undiagnosed and untreated, and how much other clinical actions, like treatment or public health measures to reduce disease spread, could help the patient or the community.

### REFERENCES


1. Weber BE, Kapoor WN. Evaluation and outcomes of patients with palpitations. *Am J Med.* 1996;100:138-148.
2. Sox HC, Blatt MA, Higgins MC, Marton KI. *Medical Decision Making.* Boston, Mass: Butterworth; 1988.
3. Barondess JA, Carpenter CCJ, eds. *Differential Diagnosis.* Philadelphia, Pa: Lea & Febiger; 1994.
4. Glass RD. *Diagnosis: A Brief Introduction.* Melbourne, Australia: Oxford University Press; 1996.
5. Engel GL. The need for a new medical model: a challenge for biomedicine. *Science.* 1977;196:129-136.
6. Laupacis A, Wells G, Richardson WS, Tugwell P, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, V: how to use an article about prognosis. *JAMA.* 1994;272:234-237.
7. Guyatt GH, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, II: how to use an article about therapy or prevention, A: are the results of the study valid? *JAMA.* 1993;270:2598-2601.
8. Guyatt GH, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, II: how to use an article about therapy or prevention, B. *JAMA.* 1994;271:59-63.
9. Dans AL, Dans LF, Guyatt GH, Richardson WS, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, XIV: how to decide on the applicability of clinical trial results to your patient. *JAMA.* 1998;279:545-549.
10. Jaeschke R, Guyatt GH, Sackett DL, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, III: how to use an article about a diagnostic test, A. *JAMA.* 1994;271:389-391.
11. Jaeschke R, Guyatt GH, Sackett DL, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, III: how to use an article about a diagnostic test, B. *JAMA.* 1994;271:703-707.
12. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. *N Engl J Med.* 1980;302:1109-1117.
13. Tversky A, Kahneman D. Judgment under uncertainty. *Science.* 1974;185:1124-1131.
14. Dawson NV, Arkes HR. Systematic errors in medical decision making: judgment limitations. *J Gen Intern Med.* 1987;2:183-187.
15. Kassirer JP, Kopelman RI. Cognitive errors in diagnosis. *Am J Med.* 1989;86:433-441.
16. Guyatt GH, Patterson C, Ali M, et al. Diagnosis of iron deficiency anemia in the elderly. *Am J Med.* 1990;88:205-209.
17. Kroenke K. Symptoms and science: the frontiers of primary care research [editorial]. *J Gen Intern Med.* 1997;12:509-510.
18. Fletcher RH, Fletcher SW, Wagner EH. *Clinical Epidemiology: The Essentials.* 3rd ed. Baltimore, Md: Williams & Wilkins; 1996:61-64.
19. Sox HC, Hickam DH, Marton KI, et al. Using the patient's history to estimate the probability of coronary artery disease. *Am J Med.* 1990;89:7-14.
20. Katz DA, Bates DW, Rittenberg E, et al. Predicting *Clostridium difficile* stool cytotoxin results in hospitalized patients with diarrhea. *J Gen Intern Med.* 1997;12:57-62.
21. von Reyn CF, Levy BS, Arbeit RD, et al. Infective endocarditis: an analysis based on strict case definitions. *Ann Intern Med.* 1981;94:505-517.
22. Durack DT, Lukes AS, Bright DK, and the Duke Endocarditis Service. New criteria for diagnosis of infective endocarditis. *Am J Med.* 1994;96:200-209.
23. Kroenke K, Lucas CA, Rosenberg ML, et al. Causes of persistent dizziness: a prospective study of 100 patients in ambulatory care. *Ann Intern Med.* 1992;117:898-904.
24. Kapoor WN. Evaluation and outcome of patients with syncope. *Medicine.* 1990;69:160-175.
25. Pratter MR, Bartter T, Akers S, et al. An algorithmic approach to chronic cough. *Ann Intern Med.* 1993;119:977-983.
26. Benbadis SR, Sila CA, Cristea RL. Mental status changes and stroke. *J Gen Intern Med.* 1994;9:485-487.
27. Ghali JK, Kadakia S, Cooper R, Ferlinz J. Precipitating factors leading to decompensation of heart failure. *Arch Intern Med.* 1988;148:2013-2016.
28. Katsarkas A. Dizziness in aging—a retrospective study of 1194 cases. *Otolaryngol Head Neck Surg.* 1994;110:296-301.
29. Altman DG. Confidence intervals [Appendix]. In: Sackett DL, Richardson WS, Rosenberg WMC, Haynes RB, eds. *Evidence-Based Medicine: How To Practice and Teach EBM.* New York, NY: Churchill Livingstone; 1997.
30. Hanley JA, Lippman-Hand A. If nothing goes wrong, is everything all right? Interpreting zero numerators. *JAMA.* 1983;249:1743-1745.
31. Newman TB. If almost nothing goes wrong, is almost everything all right? Interpreting small numerators. *JAMA.* 1995;274:1013.
32. Glasziou P, Guyatt GH, Dans AL, Dans LF, Straus SE, Sackett DL. Applying the results of trials and systematic reviews to individual patients [editorial]. *ACP Journal Club.* November-December 1998;129:A15-A16.
33. Petersdorf RG, Beeson PB. Fever of unexplained origin: report on 100 cases. *Medicine.* 1961;40:1-30.
34. Larson EB, Featherstone HJ, Petersdorf RG. Fever of undetermined origin: diagnosis and follow-up of 105 cases, 1970-1980. *Medicine.* 1982;61:269-292.
35. Knockaert DC, Vanneste LJ, Vanneste SB, Bobbaers HJ. Fever of unknown origin in the 1980s. *Arch Intern Med.* 1992;152:51-55.

# Users' Guides to the Medical Literature: XV. How to Use an Article About Disease Probability for Differential Diagnosis

W. Scott Richardson; Mark C. Wilson; Gordon H. Guyatt; et al.

*JAMA*. 1999;281(13):1214-1219 (doi:10.1001/jama.281.13.1214)

http://jama.ama-assn.org/cgi/content/full/281/13/1214

# Users' Guides to the Medical Literature
## XVI. How to Use a Treatment Recommendation

Gordon H. Guyatt, MD, MSc
Jack Sinclair, MD
Deborah J. Cook, MD, MSc
Paul Glasziou, MB, BS, PhD
for the Evidence-Based Medicine Working Group and the Cochrane Applicability Methods Working Group

**Clinicians can often find treatment recommendations in traditional narrative reviews and the discussion sections of original articles and meta-analyses. Making a treatment recommendation involves framing a question, identifying management options and outcomes, collecting and summarizing evidence, and applying value judgments or preferences to arrive at an optimal course of action. Each step in this process can be conducted systematically (thus protecting against bias) or unsystematically (leaving the process open to bias). Clinicians faced with a plethora of recommendations may wish to attend to those that are less likely to be biased. Therefore, we propose a hierarchy of rigor of recommendations to guide clinicians when judging the usefulness of particular recommendations. Recommendations with the highest rigor consider all relevant options and outcomes, include a comprehensive collection of the methodologically highest quality data with an explicit strategy for summarizing the data (that is, a systematic review), and make an explicit statement of the values or preferences involved in moving from evidence to action. High rigor recommendations come from systematically developed, evidence-based practice guidelines or rigorously conducted decision analyses. Systematic reviews, which typically do not consider all relevant options and outcomes or make the preferences underlying recommendations explicit, offer intermediate rigor recommendations. Traditional approaches in which the collection and assessment of evidence remains unsystematic, all relevant options and outcomes may not be considered, and values remain implicit, provide recommendations of weak rigor. In an era in which clinicians are barraged by recommendations as to how to manage their patients, this hierarchy provides a potentially useful set of guides.**

*JAMA. 1999;281:1836-1843* www.jama.com

### CLINICAL SCENARIO

You are a primary care practitioner considering the possibility of anticoagulant therapy with warfarin in a new patient, a 76-year-old woman with chronic congestive heart failure and atrial fibrillation. The patient has no hypertension, valvular disease, or other comorbidity. Aspirin is the only antithrombotic agent that the patient has received over the 10 years during which she has been in atrial fibrillation. Her other medications include captopril, furosemide, and metoprolol. The duration of the patient's atrial fibrillation and her dilated left atrium on echocardiogram dissuade you from prescribing antiarrhythmic therapy. Discussing the issue with the patient, you find she places a high value in avoiding a stroke, a somewhat lower value in avoiding a major hemorrhage, and would accept the inconvenience associated with monitoring anticoagulant therapy.

You have little inclination to review the voluminous original literature relating to the benefits of anticoagulant therapy in reducing stroke or its risk of bleeding, but hope to find an evidence-based recommendation to guide your advice to the patient. In your office file relating to this problem you find a report of a primary study,[1] a decision analysis,[2] and a recent practice guideline[3] that you hope will help.

### INTRODUCTION

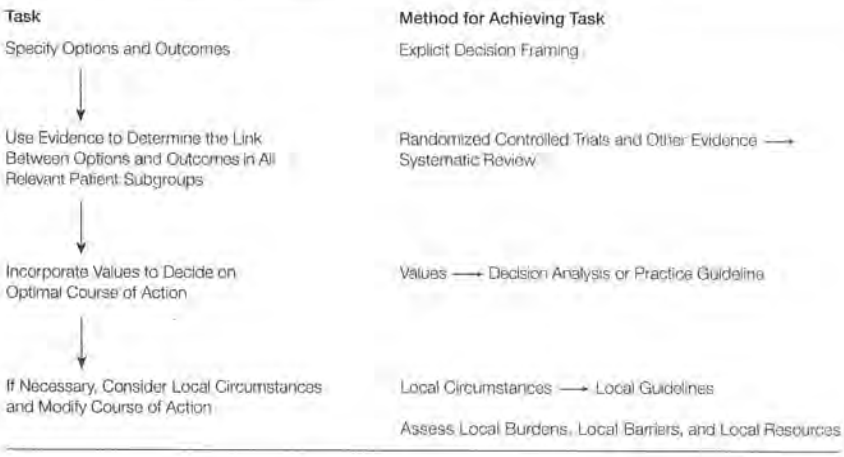
Each day, clinicians make dozens of patient management decisions. Some are relatively inconsequential, some are important. Each one involves weighing benefits and risks, gains and losses, and recommending or instituting a course of action judged to be in the patient's best interest. These decisions involve an implicit consideration of the relevant evidence, an intuitive integration of the evidence, and a weighing of the likely benefits and harms. In making choices, clinicians may benefit from structured summaries of the options and outcomes, systematic reviews of the evidence regarding the relation between options and outcomes, and recommen-

**Author Affiliations:** Departments of Clinical Epidemiology and Biostatistics (Drs Guyatt, Sinclair, and Cook), Medicine (Drs Guyatt and Cook), and Pediatrics (Dr Sinclair), McMaster University, Hamilton, Ontario; and Department of Social and Preventive Medicine, University of Queensland Medical School, Herston QLD, Australia (Dr Glasziou).
**The members of the Evidence-Based Medicine Working Group and the Cochrane Applicability Methods Working Group** are listed at the end of this article.
**Corresponding Author and Reprints:** Gordon H. Guyatt, MD, MSc, McMaster University Health Sciences Centre, 1200 Main St W, Room 2C12, Hamilton, Ontario, Canada L8N 3Z5.
**Users' Guides to the Medical Literature Section Editor:** Drummond Rennie, MD, Deputy Editor (West), *JAMA*.

dations regarding the best choices. This Users' Guide explores the process of developing recommendations, suggests how the process may be conducted systematically, and introduces a taxonomy for differentiating recommendations that are more rigorous (and thus more likely to be trustworthy) from those that are less rigorous (and thus at greater risk of being misleading).

While recommendations may be directed at health policymakers, our focus is advice for practicing clinicians. We will begin by considering the implicit steps that are involved in making a recommendation.

**Figure.** A Schematic View of the Process of Developing a Treatment Recommendation




## THE PROCESS OF DEVELOPING A RECOMMENDATION

The FIGURE presents the steps involved in developing a recommendation and the formal strategies that are available. The first step in clinical decision making is to define the decision. This involves specifying the alternative courses of action and the alternative outcomes. Often, treatments are designed to delay or prevent an adverse outcome such as stroke, death, or myocardial infarction. In our discussion, we will refer to the outcomes that treatment is designed to prevent as *target outcomes*. Treatments are associated with their own adverse outcomes—adverse or toxic effects. Ideally, the definition of the decision will be comprehensive—all reasonable alternatives will be considered and all possible beneficial and adverse outcomes will be identified. In patients like the woman with nonvalvular atrial fibrillation in the scenario, options include not treating the patient, giving her aspirin, or anticoagulant therapy with warfarin. Outcomes include minor and major embolic stroke, intracranial hemorrhage, gastrointestinal hemorrhage, minor bleeding, and the inconvenience associated with taking and monitoring medication.

Having identified the options and outcomes, decision makers must evaluate the links between the two—what will the alternative management strategies yield in terms of benefit and harm?[4] They must also consider how this impact is likely to vary in different groups of patients.[5] Having made estimates of the consequences of alternative strategies, value judgments about the relative desirability or undesirability of possible outcomes becomes necessary to allow treatment recommendations. We will use the term *preferences* synonymously with *values* or *value judgments* in referring to the process of trading off positive and negative consequences of alternative management strategies.

Recently, investigators have applied scientific principles to the collection, selection, and summarization of evidence, and the valuing of outcomes. We will briefly describe these systematic approaches.

### Linking Management Options and Outcomes—Systematic Reviews

Unsystematic identification and collection of evidence risks biased ascertainment—treatment effects may be underestimated or, more commonly, overestimated, and adverse effects may be exaggerated or ignored. Unsystematic summaries of data run similar risks of bias. One result of these unsystematic approaches may be recommendations advocating harmful treatments and failing to encourage effective therapy. For example, experts advocated routine use of lidocaine for patients with acute myocardial infarction when available data suggested the intervention was ineffective and possibly even harmful and failed to recommend thrombolytic agents when data showed patient benefit.[6]

Systematic reviews deal with this problem by explicitly stating inclusion and exclusion criteria for evidence to be considered, conducting a comprehensive search for the evidence, and summarizing the results according to explicit rules that include examining how effects may vary in different patient subgroups.[7,8] When a systematic review pools data across studies to provide a quantitative estimate of overall treatment effect, we call it a *meta-analysis*. Systematic reviews provide strong evidence when the quality of the primary studies is high and sample sizes are large and less strong evidence when designs are weaker and sample sizes small. Because judgment is involved in many steps in a systematic review (including specifying inclusion and exclusion criteria, applying these criteria to potentially eligible studies, evaluating the methodological quality of the primary studies, and selecting an approach to data analysis), systematic reviews are not immune from bias. Nevertheless, in their rigorous approach to collecting and summarizing data, systematic reviews reduce the likelihood of bias in estimating the causal links between management options and patient outcomes.

### Decision Analysis

Rigorous decision analysis provides a formal structure for integrating the evidence about the beneficial and harmful effects of treatment options with the values or preferences associated with those beneficial and harmful effects. When done well, a decision analysis will use systematic reviews of the best evidence to estimate the probabilities of the outcomes and use appropriate sources of preferences (those of society or of relevant patient groups) to generate treatment recommendations.[9,10] When a decision analysis includes costs among the outcomes, it becomes an economic analysis and summarizes trade-offs between gains (typically valued in quality-adjusted life years [QALYs]) and resource expenditure (valued in dollars).[11,12] A decision analysis will be open to bias if it fails criteria for a systematic overview in accumulating and summarizing evidence or uses preferences that are arbitrary or come from small or unrepresentative populations (such as a small group of health care providers).

**Table 1.** Methodologic Requirements for Systematic, Rigorous Recommendations

1. Comprehensive statement of management options and possible outcomes.
2. Systematic review and summary of evidence linking options to outcomes. Examination of the magnitude of impact, in terms of both benefits and risks, in relative and absolute terms.
3. Consideration of different populations, and the characteristics of these populations, that may affect impact of intervention.
4. Examination of strength of evidence linking options to outcomes. Where evidence is weak, examine the implications of plausible differences in effects.
5. Explicit, appropriate specification of values or preferences associated with outcomes.

### Practice Guidelines

Practice guidelines provide an alternative structure for integrating evidence and applying values to reach treatment recommendations. Practice guideline methodology places less emphasis on precise quantitation than does decision analysis. Instead, it relies on the consensus of a group of decision makers, ideally including experts, frontline clinicians, and patients, who carefully consider the evidence and decide on its implications. Rigorous practice guidelines will also use systematic reviews to summarize evidence and sensible strategies to attribute values to alternative outcomes.[13,14] Guidelines developers may focus on local circumstances. For example, clinicians practicing in rural parts of less industrialized countries without resource to monitor its intensity may reject anticoagulant therapy as a management approach for patients with atrial fibrillation. Practice guidelines may fail methodological standards in the same ways as decision analyses.

We will now contrast these systematic approaches to developing recommendations with historical practice.

### Current Sources of Treatment Recommendations

Traditionally, authors of original or primary research into therapeutic interventions include recommendations about the use of these interventions in clinical practice in the discussion section of their articles. Authors of systematic reviews and meta-analyses also tend to provide their impressions of the management implications of their studies. Typically, however, individual trials or overviews do not consider all possible management options, but focus on a comparison of 2 or 3 alternatives. They may also fail to identify subpopulations in which the impact of treatment may vary considerably. Finally, when the authors of overviews provide recommendations, they are not typically grounded in an explicit presentation of societal or patient preferences.

Failure to consider these issues may lead to variability in recommendations given the same data. For example, a number of meta-analyses of selective decontamination of the gut using antibiotic prophylaxis for pneumonia in critically ill patients with similar results regarding the impact of treatment on target outcomes resulted in recommendations varying from suggesting implementation, to equivocation, to rejecting implementation.[15-18] Varying recommendations reflect the fact that both investigators reporting primary studies or doing meta-analyses often make their recommendations without benefit of an explicit standardized process or set of rules.

When benefits or risks are dramatic and are essentially homogeneous across an entire population, intuition may provide an adequate guide to making treatment recommendations. Such situations are unusual. In most instances, because of their susceptibility to both bias and random error, intuitive recommendations risk misleading the clinician.

These considerations suggest that when clinicians examine treatment recommendations, they should critically evaluate the methodological quality of the recommendations. The greater the extent to which recommendations adhere to the methodological standards we have mentioned, the greater faith clinicians may place in those recommendations (TABLE 1). TABLE 2 presents a scheme for classifying the methodological quality of treatment recommendations, emphasizing the 3 key components: consideration of all relevant

**Table 2.** A Hierarchy of Rigor in Making Treatment Recommendations

| Level of Rigor | Systematic Summary of Evidence | Considers All Relevant Options and Outcomes | Explicit Statement of Values | Example Methodologies |
|---|---|---|---|---|
| High | Yes | Yes | Yes | Practice guidelines or decision analysis* |
| Intermediate | Yes | Yes or No | No | Systematic review* |
| Low | No | Yes or No | No | Traditional reviews; original articles |

*Example methodologies may not reflect the level of rigor shown. Exceptions may occur in either direction. For example, if a practice guideline or decision analysis neither systematically collects and summarizes information, nor explicitly considers societal or patients' values, it will produce recommendations that are of low rigor. If a systematic review does consider all relevant options and at least qualitatively considers values, it can produce recommendations approaching high rigor.

options and outcomes, a systematic summary of the evidence, and explicit and/or quantitative consideration of societal or patient preferences. In the next section of the article, we will describe the rating system summarized in Table 2.

## MAKING RECOMMENDATIONS: A HIERARCHY OF RIGOR

### Systematic Summary of Evidence for All Relevant Interventions Using Appropriate Values

**Quantitative Summary of Evidence and Values.** The most rigorous approach to making recommendations (which we will call a *systematic synthesis*) involves precisely quantifying all benefits and risks; determining the values of either a group of patients or the general population; where uncertainty exists, making a systematic and quantitative exploration of the range of possible true values; and using quantitative methods to synthesize the data. One approach to meeting these criteria involves conducting a formal decision analysis. Many decision analyses fail to carry out each step in the process in an optimally rigorous fashion; to do so usually requires a major research project.[9,10]

Challenges for investigators doing decision analysis include conducting the systematic reviews required to generate the best estimates of benefits and risks associated with treatment options and measuring how the general public or patients value the relevant outcomes. Typically, a decision analysis values each treatment arm in terms of QALYs. When costs are considered, the decision analysis becomes an economic analysis, and we think in terms of additional dollars spent to gain an additional QALY. The optimal therapy or the cost-effectiveness of alternatives may differ depending on untreated patients' risk of the target outcome.

What a decision analysis or economic analysis usually does not do is to value the benefits, risks, and costs and provide an explicit threshold for decision making. For example, a new treatment might cost $50 000 per QALY gained. Is this a bargain or too great a cost to warrant treatment? Often, investigators doing decision analysis will refer to the cost-effectiveness or cost-utility ratios of currently used treatments to help with this decision. For instance, the decision analysis from the scenario in this article concluded that while the cost of warfarin for patients with at least 1 factor increasing their risk of embolism was $8000 per QALY saved, the cost was $375 000 per QALY saved for a 65-year-old person with no risk factors.[2] The authors compared these figures to the $50 000 to $100 000 cost per QALY gained when screening adults for hypertension.

**Quantitative Summary of Evidence and Values: Explicit Decision Thresholds.** Investigators can use the principles of decision analysis to arrive at explicit decision thresholds and present these thresholds in ways that facilitate clinicians' understanding. One such approach involves the number of patients to whom one must administer an intervention to prevent a single target event, the number needed to treat (NNT).[19] Typically, the NNT falls as patients' risk of an adverse outcome rises and may become extremely large when patients are at very low risk. In a previous Users' Guide, we have described the threshold NNT,[20] the dividing line between when treatment is warranted (the NNT is low enough that the benefits outweigh the costs and risks) and when it is not (the NNT is too great to warrant treatment). Deriving the threshold NNT involves specifying the relative value associated with preventing the target outcome vs incurring the adverse effects and inconvenience associated with treatment.[21]

Investigators using this approach may also consider costs. If so, they face the additional requirement of specifying the number of dollars one would be willing to pay to prevent a single target event. With or without considering costs, investigators can plug the values they adduce into an equation that generates the threshold NNT.[20] They can then look at the risk of the target outcome in untreated subpopulations to whom clinicians might consider administering the intervention. Combining this information with the relative risk reduction associated with the treatment, they can determine on which side of the threshold the treatment falls.

Returning to our example, warfarin decreases the risk of stroke in patients with nonvalvular atrial fibrillation. Since anticoagulation increases bleeding risk, it is not self-evident that we should be recommending the treatment for our patients and must find a way of trading off decreased stroke and increased bleeding. We can calculate the threshold NNT by specifying the major adverse outcome of treatment, bleeding, and the frequency with which it occurs due to treatment. We then specify the impact of these deleterious effects relative to the target event the treatment prevents, a stroke. A variety of studies of relevant patient populations[22-25] suggest that, on average, patients consider 1 severe stroke equivalent to 5 episodes of serious gastrointestinal bleeding. We use these figures to calculate our threshold NNT, which proves to be approximately 152 (TABLE 3). This implies that if we need to provide anticoagulant therapy to fewer than 152 patients to prevent a stroke, we will do so; if we must provide anticoagulant therapy to more than 152 patients, then our recommendation will be to not treat.

The threshold NNT then facilitates recommendations for specific patient groups. TABLE 4 summarizes the calculation of the NNT and the associated comparison with the threshold for 2 groups of patients. A meta-analysis of randomized trials tells us that anticoagulant therapy reduces the risk of stroke by 68% (95% confidence interval, 50%-79%) and that this risk reduction is consistent across clinical trials.[26] The meta-analysis also provides risk estimates for different groups of patients with strokes. Patients older than 75 years with any previous cerebrovascular events, diabetes, hypertension, or heart disease have a stroke risk of approximately 8.1% per year. Anticoagulation reduces this risk to 2.6% with an

**Table 3.** Calculating the Threshold Number Needed to Treat (T-NNT) for Warfarin Treatment of Patients With Nonvalvular Atrial Fibrillation

We consider anticoagulant therapy for patients with nonvalvular atrial fibrillation to prevent strokes that may be fatal or, in survivors, severe or mild. The relative frequencies of the different types of stroke provide the weights used to calculate the utility, cost, and value of an "average" stroke. The adverse event caused by treatment is hemorrhage, which may be serious (adverse event 1 [AE1]) or minor (adverse event 2 [AE2]). The relative frequencies of the different types of serious hemorrhage (fatal, severe central nervous system [CNS], mild CNS, or gastrointestinal [GI]) provide the weights used to calculate the average utility, cost, and value of the "average" serious hemorrhage.

| Target Event: Stroke | | | | |
|---|---|---|---|---|
| Type of Stroke | Relative Frequency | Utility | Cost, $ | Value |
| Fatal | 0.25 | 0 | 0 | 100 000 |
| Severe | 0.25 | 0.4 | 34 200 | 60 000 |
| Mild | 0.50 | 0.8 | 7800 | 20 000 |
| Any stroke | | 0.5 | 12 450 | 50 000 |

| Adverse Event: Hemorrhage | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | Value | |
| Type of Hemorrhage | Attributable Risk | Relative Frequency | Utility | Cost, $ | Absolute | Relative to Target Event |
| Serious hemorrhage | | | | | | |
| Fatal | 0.0012 | 0.20 | 0 | 0 | 100 000 | 2.0 |
| Severe CNS | 0.00018 | 0.03 | 0.4 | 34 200 | 60 000 | 1.2 |
| Mild CNS | 0.00048 | 0.08 | 0.8 | 7800 | 20 000 | 0.4 |
| GI | 0.00414 | 0.69 | 0.8 | 3920 | 20 000 | 0.4 |
| AE1 | 0.006 | | 0.628 | 4355 | 37 200 | 0.744 |
| AE2 | 0.15 | | 0.993 | 100 | 700 | 0.014 |

Computation of relative value from utility:

$$\text{Relative value} = (1 - \text{utility of adverse event})/(1 - \text{utility of target event})$$

Thus, when utility of target event = 0.5:

| Adverse Event | Utility | Relative Value |
|---|---|---|
| Serious hemorrhage | | |
| Fatal | 0 | 1 − 0/1 − 0.5 = 2.0 |
| Severe CNS | 0.4 | 1 − 0.4/1 − 0.5 = 1.2 |
| Mild CNS | 0.8 | 1 − 0.8/1 − 0.5 = 0.4 |
| GI | 0.8 | 1 − 0.8/1 − 0.5 = 0.4 |
| AE1 | 0.628 | 1 − 0.628/1 − 0.5 = 0.744 |
| AE2 | 0.993 | 1 − 0.993/1 − 0.5 = 0.014 |

Cost of treatment = $800 per patient treated

T-NNT (not considering costs):

$$\text{T-NNT} = \frac{1}{(\text{Value}_{AE1} \cdot \text{Rate}_{AE1}) + (\text{Value}_{AE2} \cdot \text{Rate}_{AE2})}$$

(where $\text{Value}_{AE}$ = value of AE relative to that of target event)

$$= \frac{1}{(0.744 \cdot 0.006) + (0.014 \cdot 0.15)}$$

$$= 152$$

T-NNT (full model, including costs):

$$\text{T-NNT} = \frac{\text{Cost}_{target} + \text{Value}_{target}}{[\text{Cost}_{treatment} + (\text{Cost}_{AE1} \cdot \text{Rate}_{AE1}) + (\text{Cost}_{AE2} \cdot \text{Rate}_{AE2})] + [(\text{Value}_{AE1} \cdot \text{Rate}_{AE1}) + (\text{Value}_{AE2} \cdot \text{Rate}_{AE2})]}$$

$$= \frac{12\,450 + 50\,000}{[800 + (4355 \cdot 0.006) + (100 \cdot 0.15)] + [(37\,200 \cdot 0.006) + (700 \cdot 0.15)]}$$

$$= 53$$

NNT of 1 divided by 0.055, or approximately 18 per year. The NNT for this group is appreciably lower than the threshold NNT, suggesting that such patients should be treated.

Patients younger than 65 years with no risk factors have a 1-year stroke risk of 1%, which anticoagulant therapy reduces to 0.32%. The associated NNT of 146 approximates the threshold NNT of 152 and suggests the decision about whether or not to treat is a toss-up.

Clinicians or health care decision makers interested in considering costs in their decisions can look for help from the model. Costs can be included by specifying the dollar value associated with preventing adverse outcomes (for example, Laupacis and colleagues[27] have suggested the most that society might be willing to pay to gain a QALY is $100 000). When we consider costs as calculated in the decision analysis from the patient scenario,[2] we arrive at a threshold NNT of 53, suggesting a more conservative approach to anticoagulant administration (Table 3).

Investigators can use units other than NNT to develop clinically useful decision thresholds. For example, for 81 patients previously treated with cisplatin-based chemotherapy, the average minimum gain in survival that was felt to make the chemotherapy worthwhile was 4.5 months for mild toxicity and 9 months for severe toxicity.[28] Such a threshold could be integrated with information about the actual gain in life associated with the treatment to help form the basis for a recommendation about use of cisplatin therapy.

Like other quantitative approaches, considering NNT and the threshold NNT, or alternative thresholds, is intended to supplement clinical judgment, not replace it. Investigators exploring different treatment choices have found the method useful.[29] However clinicians use it, the approach highlights the necessity for both valuing the benefits and risks of treatment and understanding the magnitude of those benefits and risks in making a treatment decision.

**Quantitative Summary of Evidence, Qualitative Summary of Preferences.** Practice guidelines, if they are to minimize bias, should not substitute expert opinion for a systematic review of the literature, and should have an explicit and sensible process for valuing outcomes, an explicit consideration of the impact of uncertainty associated with the evidence and values used in the guidelines, and an explicit statement of the strength of evidence supporting the guideline. When a practice guideline meets these methodological standards, and thereby minimizes bias, we refer to the guideline as *evidence-based* (Table 1).

Once they have the evidence, investigators and clinicians are often uncomfortable with explicitly specifying preferences in moving from evidence to action. Their reluctance is understandable. Specifying a trade-off between a stroke and a gastrointestinal hemorrhage is not an exercise with which we are familiar. People may feel that identifying a specific value—a stroke is equivalent to 2.5 gastrointestinal hemorrhages, for instance—implies more precision than is realistic. Discomfort may increase further when we specify a dollar value associated with preventing an adverse event.

This may be 1 reason that participants in the development of rigorous practice guidelines, including experts in the content area, methodologists, community practitioners, and patients and their representatives, seldom use numbers to identify the value judgments they are making. Still, a rigorous guideline will establish, reflect, and make explicit the community and patient values on which the recommendation is based.

Most practice guidelines fail to systematically summarize the evidence. Even those that meet criteria for evidence accumulation and summarization do not usually make their underlying values explicit. Guidelines that do not meet either set of criteria produce recommendations of low methodologic rigor.

Practice guidelines that meet the criteria in Table 1 provide an alternative to quantitative strategies to arrive at a systematic synthesis.

**Table 4.** Using the Number Needed to Treat (NNT) to Make Treatment Recommendations

| Patient Group | Risk of Stroke Without Treatment, % | Relative Risk Reduction With Warfarin, % | Group's Absolute Risk Reduction, % | Group's NNT (Threshold NNT) |
|---|---|---|---|---|
| Age <65 y, no risk factors | 1 | 68 | 0.68 | 146 (cost omitted: 152; cost considered: 53) |
| Previous thromboembolic event | 8.1 | 68 | 5.5 | 18 (cost omitted: 152; cost considered: 53) |

### Systematic Review of Evidence, Unsystematic Application of Values

Traditionally, investigators provide their results and then make an intuitive recommendation about the action that they believe should follow from their evidence. They may do so without considering all treatment options or all outcomes (Table 2). Even when they consider all relevant treatments and outcomes, they may fail to use community or patient values, or even to make the values they are using explicit. For instance, the authors of a meta-analysis of antithrombotic therapy in atrial fibrillation stated "about one patient in seven in the combined study cohort were at such low risk of stroke (1% per year) that chronic anticoagulation is not warranted."[26] Here, the relative value of stroke and gastrointestinal bleeding is implicit in the recommendation. The nature of the value judgment is not transparent, and we have no guarantee that the implicit values reflect those of our patient or community. Clinicians faced with such recommendations need to take care that they are aware of all relevant outcomes, both reductions in targets and treatment-related adverse events, and are aware of the relative values implied in the treatment recommendations.

### Unsystematic Review, Unsystematic Synthesis

The unsystematic approach represents the traditional strategy of accumulating and summarizing evidence in an unsystematic fashion and then applying implicit preferences to arrive at a treatment recommendation. The approach is open to bias and is likely to lead to consistent, valid recommendations only when the gradient between beneficial and adverse consequences of alternative actions is very large.

### Intermediate Approaches

Both quantitative strategies and practice guidelines, when done rigorously, are very resource-intensive. Investigators may adopt less onerous methods and still provide useful insights. Researchers doing meta-analyses may wish to take the first steps in making treatment recommendations without a formal decision analysis or practice guideline development exercise. If they are to optimize the rigor of these tentative recommendations they will comprehensively identify all options and outcomes and use their meta-analysis to establish the causal links between the two. They may then choose to label values in only a qualitative way, such as: "We value preventing a stroke considerably more highly than incurring a gastrointestinal hemorrhage. Given this value, we would be willing to treat a moderate-to-large number of patients to prevent a single target event and would therefore recommend treating all but those at lowest risk of stroke."

Clinicians may find such recommendations useful, and they have the advantage of highlighting that if one does not share the specified values, one would choose an alternative treatment strategy. They may not, however, reflect community or patient preferences. In addition, they are less specific than the process of placing a number on our values. While quantifying values may make us uncomfortable, we are regularly (if uncon-

sciously) making such judgments in the process of instituting or withholding treatment for our patients.

## ARE TREATMENT RECOMMENDATIONS DESIRABLE AT ALL?

The approaches we have described highlight that patient management decisions are always a function of both evidence and preferences. Clinicians may point out that values are likely to differ substantially between settings. Monitoring of anticoagulant therapy might take on a much stronger negative value in a rural setting where travel distances are large or in a more severely resource-constrained environment in which there is a direct inverse relationship between (for example) the resources available for purchase of antibiotics and those allocated to monitoring levels of anticoagulation.

Patient-to-patient differences in values are equally important. The magnitude of the negative value of anticoagulant monitoring or the relative negative value associated with a stroke vs a gastrointestinal hemorrhage will vary widely between individual patients, even in the same setting. If decisions are so dependent on preferences, what is the point of recommendations?

This line of argument suggests that investigators should systematically search, accumulate, and summarize information for presentation to clinicians. In addition, investigators may highlight the implications of different sets of values for clinical action. The dependence of the decision on the underlying values and the variability of values would suggest that such a presentation would be more useful than a recommendation.

We find this argument compelling. Its implementation is, however, dependent on standard methods of summarizing and presenting information that clinicians are comfortable interpreting and using. Furthermore, it implies clinicians having the time and the methods to ascertain patient values that they can then integrate with the information from systematic reviews of the impact of management decisions on patient outcomes. These requirements are unlikely to be fully met in the immediate future. Moreover, treatment recommendations are likely to remain useful for providing insight, marking progress, highlighting areas where we need more information, and stimulating productive controversy. In any case, clinical decisions are likely to improve if clinicians are aware of the underlying determinants of their actions and are able to be more critical about the recommendations offered to them. Our taxonomy may help to achieve both goals.

## RESOLUTION OF THE SCENARIO

The closest statement to a recommendation relevant to your patient from the original journal article[1] is the following: "Many elderly patients with atrial fibrillation are unable to sustain chronic anticoagulation. Furthermore, the risk of bleeding (particularly intracranial hemorrhage) was increased when elderly patients in our study received anticoagulant therapy." This study neither summarized the available evidence nor explicitly stated its underlying values; therefore, we would classify its recommendation as low in rigor.

The decision analysis uses systematic summaries of the available evidence and specifies the patient values used in developing its conclusion that "Treatment with warfarin is cost-effective in patients with nonvalvular atrial fibrillation and one or more additional risk factors for stroke,"[2] placing it in the high rigor category. Moreover, the patient values used in the analysis appear consistent with your patient's preferences. The only limitation to the decision analysis is that its bottom-line recommendation involves considerations of cost, and you have reservations about including cost considerations in your decision. The practice guideline[3] once again uses a systematic summary of the evidence, and, though making frequent reference to patients' values, does not specify the relative value of stroke and bleeding implied in its strong recommendation that high-risk patients such as ours be offered anticoagulant therapy. Nevertheless, armed with consistent recommendations from a systematic synthesis and a recommendation of intermediate rigor, you feel confident recommending your patient begin taking warfarin.

**Evidence-Based Medicine Working Group:** The original list of members (with affiliations) appears in the first article of this series (*JAMA*. 1993;270:2093-2095). A list of new members appears in the 10th article of the series (*JAMA*. 1996;275:1435-1439). The following members of the Evidence-Based Medicine Working Group contributed to this article: Pat Brill-Edwards, MD, FRCP(C); Les Irwig, MBBCh, PhD; Elizabeth Juniper, MCSP, MSc; Hui Lee, MD, MSc, FRCP (C); Mitchell Levine, MD, MSc, FRCP(C);Virginia Moyer, MD, MPH; John Philbrick, MD; W. Scott Richardson, MD; and John W. Williams, Jr, MD, MHS.
**Cochrane Applicability Methods Working Group:** Jesse Berlin, PhD, Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania School of Medicine, Philadelphia; Dianne O'Connell, BMaths (Hons), PhD, Discipline of Clinical Pharmacology, Newcastle Mater Misericordiae Hospital, Waratah, New South Wales, Australia; Gifford Batstone, MMBS, BSc, MSc, FRCPath, Centre for Postgraduate and Continuing Medical Education, University of Nottingham, England; Luc Bijnens, ScD, Clinical Data Processing and Systems, Janssen Research Foundation, Beerse, Belgium; Graham Colditz, MD, DrPH, Channing Laboratory, Brigham and Women's Hospital and Harvard Medical School, Boston, Mass; Jeremy Grimshaw, MBChB, MRCGP, Health Services Research Unit, Department of Public Health, Aberdeen, Scotland; Francois Gueyffier, MD, PhD, Service de Pharmacologie Clinique (Pr Boissel), Lyon, France; David Henry, MBChB, FRCP, School of Population Health Sciences, The University of Newcastle, New South Wales, Australia; Peter Langhorne, PhD, FRCP, Academic Section of Geriatric Medicine, Royal Infirmary, Glasgow, Scotland; Joseph Lau, MD, Department of Medicine, Tufts University School of Medicine, New England Cochrane Center, Division of Clinical Care Research, New England Medical Center, Boston, Mass; Cynthia Mulrow, MD, MSc, VA Cochrane Center at San Antonio, Audie L. Murphy Memorial Veterans Hospital, San Antonio, Tex; Chris Silagy, MBBS, PhD, FRACGP, FAFPHM, Monash Institute of Public Health and Health Services Research, Monash Medical Centre, Clayton, Victoria, Australia; and Donald E. Stanley, DO, Rutland Regional Medical Center, Rutland, Vt. The following members of the Cochrane Applicability Methods Working Group contributed to this article: Jesse Berlin, PhD, and Dianne O'Connell, PhD.
**Acknowledgment:** The Evidence-Based Medicine Working Group is deeply indebted to Deborah Maddock for her astute and dedicated administrative coordination of the Users' Guides series.

### REFERENCES

1. Stroke Prevention in Atrial Fibrillation Investigators. Warfarin versus aspirin for prevention of thromboembolism in atrial fibrillation: Stroke Prevention in Atrial Fibrillation II Study. *Lancet*. 1994;343:687-691.
2. Gage BF, Cardinalli AB, Albers GW, Owens DK. Cost-effectiveness of warfarin and aspirin for prophylaxis of stroke in patients with nonvalvular atrial fibrillation. *JAMA*. 1995;274:1839-1845.
3. Laupacis A, Albers G, Dalen J, Dunn MI, Jacobson

AK, Singer DE. Antithrombotic therapy in atrial fibrillation. *Chest*. 1998;114:579S-589S.
**4.** Glasziou PP, Irwig LM. An evidence-based approach to individualising treatment. *BMJ*. 1995;311:1356-1358.
**5.** Smith GD, Egger N. Who benefits from medical interventions? *BMJ*. 1993;308:72-74.
**6.** Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts: treatments for myocardial infarction. *JAMA*. 1992;268:240-248.
**7.** Oxman AD, Cook DJ, Guyatt GH, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VI: how to use an overview. *JAMA*. 1994;272:1367-1371.
**8.** Oxman AD, Guyatt GH. A consumer's guide to subgroup analysis. *Ann Intern Med*. 1992;116:78-84.
**9.** Richardson WS, Detsky AS, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VII: how to use a clinical decision analysis, A: are the results of the study valid? *JAMA*. 1995;273:1292-1295.
**10.** Richardson WS, Detsky AS, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VII: how to use a clinical decision analysis, B: what are the results and will they help me in caring for my patients? *JAMA*. 1995;273:1610-1613.
**11.** Drummond MF, Richardson WS, O'Brien BJ, Levine M, Heyland DK, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, XIII: how to use an article on economic analysis of clinical practice, A: are the results of the study valid? *JAMA*. 1997;277:1552-1557.
**12.** O'Brien BJ, Heyland DK, Richardson WS, Levine M, Drummond MF, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, XIII: how to use an article on economic analysis of clinical practice, B: what are the results and will they help me in caring for my patients? *JAMA*. 1997;277:1802-1806.
**13.** Hayward RS, Wilson MC, Tunis SR, Bass EB, Guyatt GH, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VIII: how to use clinical practice guidelines, A: are the recommendations valid? *JAMA*. 1995;274:570-574.
**14.** Wilson MC, Hayward RS, Tunis SR, Bass EB, Guyatt GH, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VIII: how to use clinical practice guidelines, B: what are the recommendations and will they help you in caring for your patients? *JAMA*. 1995;274:1630-1632.
**15.** Vandenbroucke-Grauls CMJ, Vendenbroucke JP. Effect of selective decontamination of the digestive tract on respiratory tract infections and mortality in the intensive care unit. *Lancet*. 1991;338:859-862.
**16.** Selective Decontamination of the Digestive Tract Trialists' Collaborative Group. Meta-analysis of randomised controlled trials of selective decontamination of the digestive tract. *BMJ*. 1993;307:525-532.
**17.** Heyland DK, Cook DJ, Jaeschke R, Griffith L, Lee HN, Guyatt GH. Selective decontamination of the digestive tract. *Chest*. 1994;105:1221-1229.
**18.** Kollef MH. The role of selective digestive tract decontamination on mortality and respiratory tract infections. *Chest*. 1994;105:1101-1108.
**19.** Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *N Engl J Med*. 1988;318:1728-1733.
**20.** Guyatt GH, Sackett DL, Sinclair JC, Hayward RS, Cook DJ, Cook RJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, IX: a method for grading health care recommendations. *JAMA*. 1995;274:1800-1804.
**21.** Caro JJ, Groome PA, Flegel KM. Atrial fibrillation and anticoagulation: from randomised trials to practice. *Lancet*. 1993;341:1381-1384.
**22.** Grootendoorst P, Feeny D, Furlong W. Health utilities index mark 3: evidence of construct validity for stroke and arthritis in a population health survey. *Med Care*. In press.
**23.** Glasziou P, Bromwich LS, Simes RJ. Quality of life six months after myocardial infarction treated with thrombolytic therapy. *Med J Aust*. 1994;161:532-536.
**24.** Solomon NA, Glick HA, Russo CJ, Lee J, Schulman KA. Patient preference for stroke outcomes. *Stroke*. 1994;25:1721-1725.
**25.** Man-Son-Hing M, Laupacis A, O'Connor A, et al. Warfarin for atrial fibrillation: the patient's perspective. *Arch Intern Med*. 1996;156:1841-1848.
**26.** Atrial Fibrillation Investigators. Risk factors for stroke and efficacy of antithrombotic therapy in atrial fibrillation: analysis of pooled data from five randomized controlled trials. *Arch Intern Med*. 1994;154:1449-1457.
**27.** Laupacis A, Feeny D, Detsky AS, Tugwell PX. How attractive does a new technology have to be to warrant adoption and utilization? tentative guidelines for using clinical and economic evaluations. *CMAJ*. 1992;146:473-481.
**28.** Silvestri G, Pritchard R, Welch G. Preferences for chemotherapy in patients with advanced non-small cell lung cancer: descriptive study based on scripted interviews. *BMJ*. 1998;317:771-775.
**29.** Robbins JM, Tilford JM, Jacobs RF, Wheeler JG, Gillaspy SR, Schutze GE. A number-needed-to-treat analysis of the use of respiratory syncytial virus immune globulin to prevent hospitalization. *Arch Pediatr Adolesc Med*. 1998;152:358-366.

> If anyone declares to you that he has actual proof, from his own experience, of something which he requires for the confirmation of his theory,—even though he be considered a man of great authority, truthfulness, earnest words and morality, yet, just because he is anxious for you to believe his theory, you should hesitate.
>
> —Moses ben Maimon (Maimonides) (1135-1204)

# Users' Guides to the Medical Literature: XVI. How to Use a Treatment Recommendation

Gordon H. Guyatt; Jack Sinclair; Deborah J. Cook; et al.

*JAMA*. 1999;281(19):1836-1843 (doi:10.1001/jama.281.19.1836)

http://jama.ama-assn.org/cgi/content/full/281/19/1836

# Users' Guides to the Medical Literature

## XVII. How to Use Guidelines and Recommendations About Screening

Alexandra Barratt, MBBS, MPH, PhD
Les Irwig, MBBCh, PhD
Paul Glasziou, MBBS, PhD
Robert G. Cumming, MBBS, MPH, PhD
Angela Raffle, BSc (Hons), MBChB
Nicholas Hicks, MA, BMBCh
J. A. Muir Gray, CBE, MD
Gordon H. Guyatt, MD, MSc
for the Evidence-Based Medicine Working Group

### CLINICAL SCENARIO

You are a family physician seeing a 47-year-old woman and her husband of the same age. They are concerned because a friend recently found out that she had bowel cancer and has urged them both to undergo screening with fecal occult blood tests (FOBTs) because, she says, prevention is much better than the cure she is now undergoing. Both your patients have no family history of bowel cancer and no change in bowel habit. They ask whether you agree that they should be screened.

You know that trials of FOBT screening have demonstrated that screening can reduce mortality from colorectal cancer (CRC), but you also recall that FOBTs can have a high false-positive rate that then requires investigation by colonoscopy. You are unsure whether screening these relatively young, asymptomatic people at average risk of bowel cancer is likely to do more good than harm. You decide to check the literature to see if there are any guidelines or recommendations about screening for CRC that might help you.

### THE SEARCH

Since you know there is more than 1 randomized controlled trial (RCT), you look first for a systematic review. Your MEDLINE search (using the terms *fecal occult blood test* and *colorectal* or *colonic neoplasms* and *mass screening* and *systematic review*) produces a systematic review by Towler et al.[1] However, there may be ancillary evidence that would influence your decision about whether to recommend screening to your patient (such as the false-positive rate of the test, the adverse effects of subsequent investigation and treatment, and costs) so you also check for a practice guideline. You find the American Gastroenterological Association (AGA) guideline on CRC screening,[2] which is based on the same trials as the systematic review but also provides the additional information you were hoping to find. The full text is provided so you print off a copy to take home and read.

### INTRODUCTION

When assessing a guideline or recommendation about screening you should apply the criteria suggested earlier in this series about assessment of health care interventions.[3-9] You may also consider other criteria for evaluating whether screening is worthwhile.[5-8] Sometimes screening is clearly effective, with large benefits and negligible harms, as is the case with phenylketonuria screening and screening for systolic hypertension (>160 mm Hg) among the elderly.[9] In other situations, clinicians must often weigh the benefits and harms when considering whether to screen.[10] This guide extends earlier approaches by providing a framework for assessing the methodological strength of guidelines on screening and by demonstrating the importance of weighing the benefits and harms of screening when they are closely balanced. The final decision about whether to screen is greatly influenced by the values different individuals place on each of the possible benefits and harms.

Our criteria for reviewing a guideline (or a meta-analysis) about screen-

**Author Affiliations:** Department of Public Health and Community Medicine, University of Sydney, Australia (Drs Barratt, Irwig, and Cumming); Department of Social and Preventive Medicine, University of Queensland, Herston, Australia (Dr Glasziou); Avon Health Authority, Bristol, England (Dr Raffle); Oxfordshire Health Authority, Oxford, England (Dr Hicks); Institute of Health Sciences, University of Oxford, England (Dr Gray); and Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario (Dr Guyatt).

The original list of members (with affiliations) appears in the first article of the series (*JAMA*. 1993; 270:2093-2095). A list of new members appears in the 10th article of the series (*JAMA*.1996;275:1435-1439). The following members of the Evidence-Based Medicine Working Group contributed to this article: Deborah Cook, MD, MSc; Lee Green, MD; Mitchell Levine, MD, MSc, FRCPC; Thomas Newman, MD; and Mark Wilson, MD.

**Corresponding Author and Reprints:** Gordon H. Guyatt, MD, MSc, McMaster University Health Sciences Centre, 1200 Main St W, Room 2C12, Hamilton, Ontario, Canada L8N 3Z5.

**Users' Guides to the Medical Literature Section Editor:** Drummond Rennie, MD, Deputy Editor (West), *JAMA*.

ing follow the Users' Guides for an article about practice guidelines (TABLE 1); in this article we will not review all the Users' Guides for guidelines, but highlight only those issues specific to screening.

TABLE 2 presents the possible consequences of screening. Some people will have true-positive test results with clinically significant disease ($a^0$): a proportion of this group will benefit according to the effectiveness of treatment and the severity of the detected disease. For example, children found to have phenylketonuria will experience large, long-lasting benefits. Other people will have "true"-positive test results with inconsequential disease ($a^1$): they may suffer harms of labeling, investigation, and treatment for a disease or risk factor that would never have affected their lives. Consider, for instance, a man in whom screening reveals low-grade prostate cancer who is destined to die of a heart attack before his prostate cancer becomes clinically manifest. He may suffer unnecessary treatment and associated adverse effects. Persons with false-positive test results (b) may suffer the harms associated with investigation of the screen-detected abnormality. Persons with false-negative test results ($c^0$) may experience harm if false reassurance results in delayed presentation or investigation of symptoms; some may also be angry when they discover they have a disease despite having a negative screening test result. In contrast, persons with "false"-negative test results who have inconsequential disease ($c^1$) are not harmed by their disease being missed because it was never destined to affect them. Persons with true-negative test results (d) may experience benefit associated with an accurate reassurance of being disease free, but may also suffer inconvenience, cost, and anxiety.

The longer the gap between possible detection and clinically important consequences, the greater the number of people in the inconsequential disease category ($a^1$). When screening for risk factors, very large numbers of people need to be screened and treated to prevent 1 adverse event years later,[11] and thus, most people found to have a risk factor at screening will be treated for inconsequential disease.

**Table 1.** Users' Guides for Guidelines and Recommendations About Screening

| |
|---|
| **Are the recommendations valid?** |
| Is there randomized controlled trial evidence that earlier intervention works? |
| Were the data indentified, selected, and combined in an unbiased fashion? |
| **What are the recommendations and will they help you in caring for your patients?** |
| What are the benefits? |
| What are the harms? |
| How do these compare in different people and with different screening strategies? |
| What is the impact of people's values and preferences? |
| What is the impact of uncertainty? |
| What is the cost-effectiveness? |

## ARE THE RECOMMENDATIONS VALID?

### Is There RCT Evidence That Earlier Intervention Works?

Guidelines recommending screening are on strong ground if they are based on RCTs in which screening is compared with conventional care. In the past, many screening programs, some of them effective (such as cervical cancer screening and screening for phenylketonuria), have been implemented on the strength of observational data. When the benefits are enormous and the downsides minimal, there is no need for RCTs. More often, the benefits and harms from screening are more evenly balanced. In these situations, observational studies of screening may be misleading. Survival as measured from the time of diagnosis may be increased, not because patients live longer, but because screening lengthens the time that they know they have disease (*lead-time bias*). Patients whose disease is discovered by screening may also appear to live longer because screening tends to detect slowly progressing disease and may miss rapidly progressive disease that becomes symptomatic between screening rounds (*length-time bias*). Therefore, unless the evidence of benefit is overwhelming, RCT assessment is required.

Investigators may choose 1 of 2 designs to test the impact of a screening process. The trial may assess the entire screening process (early detection and early intervention, FIGURE 1, left), in which case people are randomized to be screened and treated if early abnormality is detected or not screened (and treated only if symptomatic disease occurs). Trials of mammographic screening have used this design.[12-14]

Alternatively, everyone may participate in screening and those with positive test results are randomized to be treated or not treated (Figure 1, right). If those who receive treatment do better, then one can conclude that early treatment has provided some benefit. Investigators usually use this design when screening detects not the disease itself, but factors that increase the risk of disease. Tests of screening programs for hypertension and high cholesterol levels have used this design.[15,16] The principles outlined in this article apply to both screening for occult disease and screening for risk factors for later disease.

**Table 2.** Summary of Benefits and Harms of Screening by Underlying Disease State*

| | Reference Standard Results | | | |
|---|---|---|---|---|
| | Disease or Risk Factor Present | | | Disease or Risk Factor Absent |
| Screening test positive | $a^0$ = True positives (significant disease) | or | $a^1$ = "True" positives (inconsequential disease) | b = False positives |
| Screening test negative | $c^0$ = False negatives (significant disease) | or | $c^1$ = "False" negatives (inconsequential disease) | d = True negatives |

*"$a^0$ indicates disease or risk factor that will cause symptoms in the future (significant disease); $a^1$, disease or risk factor asymptomatic until death (inconsequential disease); b, false positives; $c^0$, missed disease that will be significant in the future; $c^1$, missed disease that will be inconsequential in the future; and d, true negatives. Sensitivity = a/a+c and specificity = d/b+d.

### Were the Data Identified, Selected, and Combined in an Unbiased Fashion?

As for all guidelines, developers must specify the inclusion and exclusion criteria for the studies they choose to consider, conduct a comprehensive search, and assess the methodological quality of the studies they include. Towler et al[1] searched for published and unpublished trials and assessed their quality using criteria recommended by the Cochrane Collaboration. The investigators extracted data from the trials and combined them in a meta-analysis on an intention-to-screen basis.

The AGA guideline[2] on colorectal screening used explicit inclusion and exclusion criteria and a comprehensive search to identify all the RCTs of FOBT screening. The authors include a critical appraisal of the trials and conclude that the trials provide strong evidence of effectiveness, though they are limited in that they do not consider the effect of screening on health-related quality of life.

## WHAT ARE THE RECOMMENDATIONS AND WILL THEY HELP YOU IN CARING FOR YOUR PATIENTS?

A good guideline about a screening program should summarize the trial evidence about benefits and present data about the harms. The guideline should then provide information about how these benefits and harms can vary in subgroups of the population and under different screening strategies.

### What Are the Benefits?

What outcomes need to be measured to estimate the benefits of a screening program?

Benefits will usually be experienced by some of those with positive test results, as either a reduction in mortality or an increase in quality of life. The benefit can be estimated as an absolute risk reduction (ARR) or a relative risk reduction (RRR) in adverse outcomes. (Readers desiring a full discussion of these concepts can refer back to an earlier Users' Guide.[17]) Briefly, the ARR depends on the baseline risk of disease and thus presents a more realistic estimate of the size of the mortality benefit. The RRR, in contrast, is independent of baseline risk and can lead to a misleading impression of benefit (TABLE 3). The number of people needed to screen to prevent an adverse outcome provides another way of presenting benefit.

In addition to prevention of adverse outcomes, people may also regard knowledge of the presence of an abnormality as a benefit as in antenatal screening for Down syndrome. Another potential benefit of screening comes from reassurance afforded by a negative test result, if a person is experiencing anxiety because a family member or friend has developed the target condition or from discussion in the media. However, if the anxiety is a result of the publicity surrounding the screening program itself, we would not view anxiety reduction as a benefit.

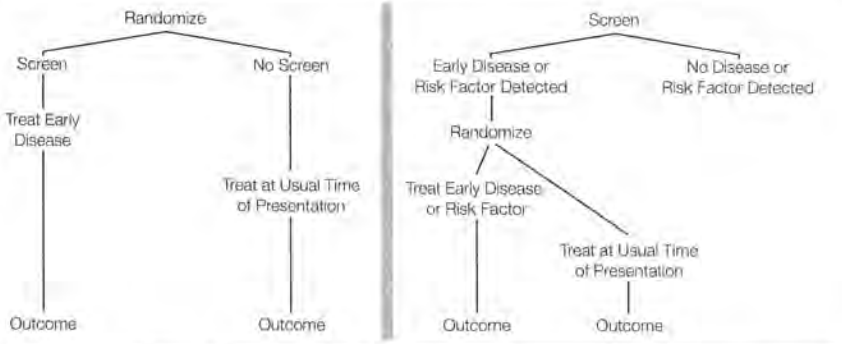
The AGA guideline reports that the RRRs from 3 trials of FOBT screening are 33% (annual screening) and 15% and 18% (biennial screening). An estimate of the uncertainty associated with these estimates (as one would get from the 95% confidence interval [CI] around a pooled RRR) would help the reader appreciate the range within which the true RRR plausibly lies. Based on a computer simulation, the AGA guideline estimates an ARR of 1330 deaths prevented per 100 000 (13.3 per 1000) people screened annually using FOBT from 50 to 85 years of age, assuming 100% participation (TABLE 4).

### What Are the Harms?

Among those with positive test results, harms may include the following:

- complications arising from investigation
- adverse effects of treatment
- unnecessary treatment of persons with true-positive test results who have inconsequential disease
- adverse effects of labeling or early diagnosis
- anxiety generated by the investigations and treatment
- costs and inconvenience incurred during investigations and treatment.

**Figure 1.** Designs for Randomized Controlled Trials of Screening




Left, A randomized controlled trial can assess the entire screening process, in which case participants are randomized to be screened (and treated) or not screened. Right, Alternatively, everyone can participate in the screening, and those with positive results are randomized to be treated or not treated.

**Table 3.** Comparison of Data Presented as Relative and Absolute Risk Reductions and Number Needed to Screen With Varying Baseline Risks of Disease and Constant Relative Risk

| Baseline Risk (Risk in Unscreened Group), % | Risk in Screened Group, % | Relative Risk Reduction, % | Absolute Risk Reduction, % | No. Needed to Screen |
|---|---|---|---|---|
| 4 | 2 | 50 | 2 | 50 |
| 2 | 1 | 50 | 1 | 100 |
| 1 | 0.5 | 50 | 0.5 | 200 |
| 0.1 | 0.05 | 50 | 0.05 | 2000 |

**Table 4.** Clinical Consequences for 1000 People Entering a Program of Annual Fecal Occult Blood Test Screening for Colorectal Cancer at Age 50 Years and Remaining in the Program Until 85 Years of Age or Death*

| Clinical Consequences | No. |
|---|---|
| Harms | |
| Screening tests | 27 030 |
| Diagnostic evaluations (by colonoscopy) | 2263 |
| False-positive screening tests | 2158 |
| Deaths due to colonoscopy complications | 0.5 |
| Bowel perforations from colonoscopy | 3.0 |
| Major bleeding episodes from colonoscopy | 7.4 |
| Minor complications from colonoscopy | 7.7 |
| Benefits | |
| Deaths averted | 13.3 |
| Years of life saved | 123.3 |
| Years of life gained per person whose cancer death was prevented | 9.3 |

*Adapted from Winawer et al.[2]

The AGA guideline reports that of the patients who do not have CRC, 8% to 10% will have false-positive test results (specificity, 90%-92% using rehydrated slides). In the trials, only 2% to 6% of those with positive test results actually had colon cancer (positive predictive value, 2%-6%). Thus, of every 100 screening participants with a positive test result, only 2 to 6 will have cancer, but all 100 will be exposed to colonoscopy and its attendant risks (Table 4). While the colonoscopies will reveal few cancers, they will show many polyps (25% of people aged 50 years or older have polyps, some of which will be judged to need removal depending on the size of the polyp). Part of the benefit of screening will come from removal of the small proportion of polyps that would have progressed to invasive cancer. Part of the harm of screening will come from regular colonoscopies that are recommended for people who have had a benign or inconsequential polyp removed.

Among those with negative test results, harms may include the following:

- anxiety generated by the screening test (waiting for result)
- false reassurance (and delayed presentation of symptomatic disease later)
- costs and inconvenience incurred during the screening test.

Of those who have cancer, FOBT screening using rehydrated slides will correctly identify 90% and miss the other 10% (sensitivity of 90%), according to the AGA guideline. Those who present with symptoms after a false-negative screen may experience a sense of anger and betrayal that they would not suffer in the absence of a screening program.

Using the computer simulation, the AGA guideline presents data on the frequency of some of these harms. These data are summarized in Table 4 for 1000 people participating in annual screening by FOBT from 50 to 85 years of age. The model assumes those who test positive have a colonoscopy.

We now know the magnitude of both benefits and harms (as presented in Table 4). This balance sheet tells us that screening 1000 people annually with FOBT from 50 years of age will prevent 13.3 deaths from CRC, but will cause 0.5 deaths from the complications of investigation and surgery. There will also be 10.4 major complications (perforations and major bleeding episodes) and 7.7 minor complications. The authors provide no data on anxiety, but we could assume that some people will feel anxious prior to colonoscopy. FIGURE 2 presents these data as a flow diagram.

These data assume that the screening programs will deliver the same magnitude of benefit and harms as found in RCTs; this will be true only if the program is delivered to the same standard of quality as in the trials. Otherwise, benefits will be smaller and the harms greater.

## How Do Benefits and Harms Compare in Different People and With Different Screening Strategies?

The AGA guideline recommends that people at average risk and older than 50 years of age be offered screening for CRC. The guideline discusses several screening strategies (FOBT, flexible sigmoidoscopy, barium enema, and colonoscopy) and, in relation to FOBT, recommends offering annual screening. The magnitude of benefits and harms will vary in different patients and under different screening strategies, as the following discussion reveals.

**Risk of Disease.** Assuming that the RRR is constant over a broad range of risk of disease, benefits will be greater for people at higher risk of disease. For example, mortality from CRC rises with age, and the mortality benefit achieved by screening rises accordingly (FIGURE 3, top). But the life years lost in the population to CRC are related both to the age at which mortality is highest and the length of life still available. Thus, the number of life years that can be saved by CRC screening increases with age to about 75 years and then decreases again as life expectancy declines (Figure 3, bottom). The number of deaths averted by screening over 10 years for those aged 40, 50, and 60 years at first screening (0.2, 1.0, and 2.4, respectively, per 1000 people[1]) reflects these differences. Because of a greater benefit, it may be rational for a 60-year-old person to decide screening is worthwhile, while a 40-year-old person (or 80 years old) with smaller potential benefit might decide it is not worthwhile.

Risk of disease, and therefore benefits from screening, may be increased by other factors, such as a family history. The AGA guideline reports that people with 1 or more first-degree relatives (parent, sibling, child) with CRC, but without one of the specific genetic syndromes, have approximately twice the risk of developing CRC as average-risk individuals without a family history. This means that for people aged 40 years who have a first-degree relative with CRC, the incidence of CRC is comparable to that for people aged 50 years without a family history. The guideline also notes that within each age group, the risk is greatest in those whose relatives developed cancer at a younger age.

**Screening Interval.** As the screening interval is shortened, the effectiveness of a screening program will tend to improve, although there is a limit to the amount of improvement that is pos-

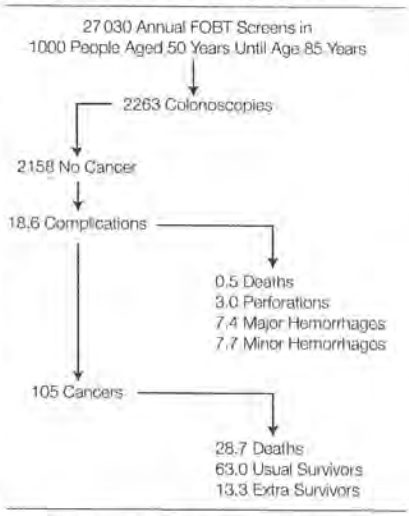
sible. For example, screening twice as often could theoretically double the relative mortality reduction obtainable by screening, but in practice, the effect is usually much less. Cervical cancer screening may, for instance, reduce the incidence of invasive cervical cancer by 64%, 84%, and 94% if screening is conducted at 10-year, 5-year, and annual intervals, respectively.[18]

The frequency of harms will also increase with more frequent screening, potentially directly in proportion to the frequency of screening. Thus, we will see diminishing marginal return as the screening interval is shortened. Ultimately, the marginal harms will outweigh the marginal benefit of further reductions in the screening interval.

**Test Characteristics.** If the sensitivity of a new test is greater than the test used in the trials and is detecting significant disease earlier, the benefit of screening will increase. But it may be that the new, apparently more sensitive, test is detecting more cases of inconsequential disease (for example, by detecting more low-grade prostate cancers or more low-grade cervical epithelial abnormalities[19]), which will increase the harms. On the other hand, if specificity is improved and testing produces fewer false-positive results, net benefit will increase and the test may now be useful in groups in which the old test was not.

Ideally, clinicians would look to RCTs of the new test compared with the old test. However, new tests often appear in profusion, and randomized trials are expensive and often only interpretable after long follow-up. Being pragmatic, we will usually need to accept that the trials have shown that earlier detection works and a comparison of a new vs the old test only needs to examine test characteristics. Returning to CRC screening, since we have RCT data of mortality reduction, we may assume that earlier detection using other methods such as flexible sigmoidoscopy will also reduce mortality from CRC even though there are no published reports of RCTs of screening with flexible sigmoidoscopy.

**Figure 2.** Flow Diagram of the Clinical Consequences for 1000 People Entering a Program of Annual Fecal Occult Blood Test (FOBT) Screening for Colorectal Cancer (CRC) at Age 50 Years and Remaining in the Program Until 85 Years of Age or Death




Usual survivors are those who would have survived with or without screening. Extra survivors are those in whom the earlier detection of cancer averts death. Adapted from Winawer et al.[2]

## What Is the Impact of People's Values and Preferences?

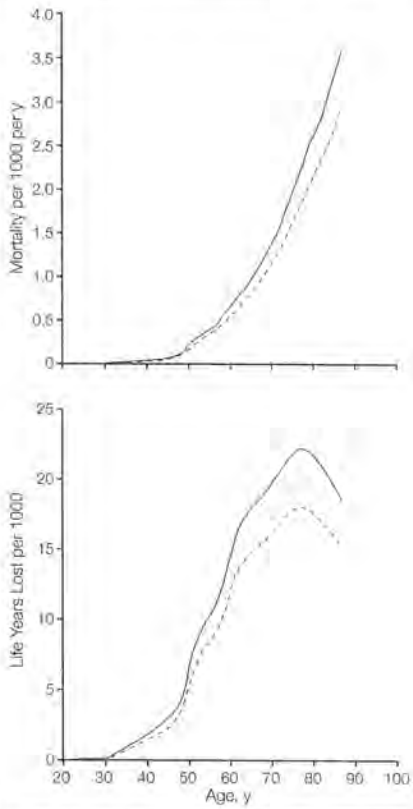
People will value benefits and harms of screening differently. For example, pregnant women who are considering screening for Down syndrome may make different choices depending on the value they place on having a Down syndrome baby vs the risk of iatrogenic abortion from amniocentesis.[20]

Individuals who choose to participate in screening programs are benefiting (in their view) from screening, and other individuals are benefiting (in their view) from not participating. Individuals can only make the right choice for themselves if they have access to high-quality information about the benefits and harms of screening and are able to weigh that information. This probably will require much better educational materials and decision support materials; some examples are already available.[21,22]

## What Is the Impact of Uncertainty Associated With the Evidence?

There is always uncertainty about the benefits and harms of screening. The

**Figure 3.** Mortality From Colorectal Cancer and Years of Life Lost Due to Colorectal Cancer With and Without Screening




Top, Mortality from colorectal cancer. Bottom, Life years lost due to colorectal cancer. Broken lines indicate with screening, and solid lines, without screening. Data from Towler et al.[1]

95% CIs around the magnitude of each benefit and harm provides an indication of the amount of uncertainty in each estimate. Where sample size is limited, the CIs will be wide and clinicians should alert potential screening participants that the magnitude of the benefit or harm could be considerably smaller or greater than the point estimate.

## What Is the Cost-effectiveness?

While clinicians will be most interested in the balance of benefits and harms for their individual patients, policymakers must consider issues of cost-effectiveness and local resources in their decisions. Clinicians can look to previous Users' Guides to help them

evaluate studies addressing these economic issues.[23,24]

The AGA guideline reports that the estimated cost-effectiveness of FOBT screening is approximately $10 000 per life year gained among people older than 50 years (although, like the absolute size of the benefit, it will vary with risk of disease). The AGA guideline also notes that all CRC screening strategies examined (FOBT, flexible sigmoidoscopy, barium enema, colonoscopy) cost less than $20 000 per life year saved.

These cost-effectiveness ratios are within the range of what is currently paid in some countries for the benefits of other screening programs such as mammographic screening for women aged 50 to 69 years (estimated at $21 400 per life year saved[25]), ultrasound screening for carotid stenosis (incremental cost per quality-adjusted life year gained is estimated at $39 495[26]) and ultrasound screening for abdominal aortic aneurysm in men aged 60 to 80 years (estimated $41 550 per life year gained[27]).

## RESOLUTION OF THE SCENARIO

The guideline should quantify the benefit of screening according to age so you can inform your patients as accurately as possible about the benefits of screening for them. The AGA guideline does not provide age-specific mortality reductions attributable to screening; therefore, you cannot easily quantify the benefit for your patients. From the guideline, all you could say is that screening a group of 1000 people with FOBT beginning at 50 years of age and continuing annually to 85 years of age will avert about 13 deaths from CRC. However, we know from the systematic review by Towler et al[1] that the mortality benefit for people between 40 and 50 years of age is about 0.2 to 1.0 deaths averted over 10 years per 1000 people screened. Next you could outline the potential harms of screening. As noted earlier, the harms are mostly related to the colonoscopy. According to the AGA guideline, the risks of colonoscopy are about 0.1 to 0.3 per 1000 for death, and 1 to 3 per 1000 for perforation and hemorrhage. In addition, there would also be issues of cost, inconvenience, and anxiety.

It is up to your patients to weigh whether the benefit of reduced risk of death from CRC is worth the risks. If they feel unable to do this, then you could consider helping them to clarify their values about the possible outcomes. For example, if they are not bothered by the prospect of a colonoscopy, they would probably choose to be screened. But if either of them places a high value on avoiding colonoscopy now, he or she may prefer to reconsider screening in a few years' time when the benefits will be greater.


**REFERENCES**

1. Towler B, Irwig L, Glasziou P, et al. A systematic review of the effects of screening for colorectal cancer using the faecal occult blood test, Hemoccult. *BMJ*. 1998;317:559-565.
2. Winawer SJ, Fletcher RH, Millar L, et al. Colorectal cancer screening: clinical guidelines and rationale. *Gastroenterology*. 1997;112:594-642.
3. Hayward RSA, Wilson MC, Tunis SR, et al, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VIII: how to use clinical practice guidelines, A: are the recommendations valid? *JAMA*. 1995;274:570-574.
4. Wilson MC, Hayward RS, Tunis SR, et al, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VIII: how to use clinical practice guidelines, B: what are the recommendations and will they help you in caring for your patients? *JAMA*. 1995;274:1630-1632.
5. Wilson JMG, Jungner G. *Principles and Practice of Screening for Disease*. Geneva, Switzerland: World Health Organization; 1968.
6. Muir Gray JA. *Evidence-Based Healthcare*. New York, NY: Churchill Livingstone; 1997.
7. Sackett DL, Haynes RB, Tugwell P. *Clinical Epidemiology: A Basic Science for Clinical Medicine*. 2nd ed. Boston, Mass: Little Brown & Co; 1991.
8. Welch HG, Black WC. Evaluating randomized trials of screening. *J Gen Intern Med*. 1997;12:118-124.
9. SHEP Co-operative Research Group. Prevention of stroke by antihypertensive drug treatment in older persons with isolated systolic hypertension: final results of the Systolic Hypertension in the Elderly Program (SHEP). *JAMA*. 1991;265:3255-3264.
10. Eddy DM. Comparing benefits and harms: the balance sheet. *JAMA*. 1990;263:2493, 2498, 2501, 2505.
11. Khaw KT, Rose G. Cholesterol screening programmes: how much benefit? *BMJ*. 1989;299:606-607.
12. Andersson I, Aspegren K, Janzon L, et al. Mammographic screening and mortality from breast cancer: the Malmo mammographic screening trial. *BMJ*. 1988;297:943-948.
13. Tabar L, Fagerberg G, Duffy S, et al. The Swedish two county trial of mammographic screening for breast cancer: recent results and calculation of benefit. *J Epidemiol Commun Health*. 1989;43:107-114.
14. Roberts MM, Alexander FE, Anderson TJ, et al. Edinburgh trial of screening for breast cancer: mortality at seven years. *Lancet*. 1990;335:241-246.
15. Multiple Risk Factor Intervention Trial Research Group. Multiple Risk Factor Intervention Trial: risk factor changes and mortality results. *JAMA*. 1982;248:1465-1477.
16. Frick MH, Elo E, Haapa K, et al. Helsinki Heart Study: primary prevention trial with gemfibrizil in middle-aged men with dyslipidemia. *N Engl J Med*. 1987;317:1237-1245.
17. Guyatt GH, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, II: how to use an article about therapy or prevention, B: what were the results and will they help me in caring for my patients? *JAMA*. 1994;271:59-63.
18. IARC Working Group on Evaluation of Cervical Cancer Screening Programmes. Screening for squamous cervical cancer: duration of low risk after negative results of cervical cytology and its implication for screening policies. *BMJ*. 1986;293:659-664.
19. Raffle AE. New tests in cervical screening. *Lancet*. 1998;351:297.
20. Fletcher J, Hicks NR, Kay JDS, Boyd PA. Using decision analysis to compare policies for antenatal screening for Down's syndrome. *BMJ*. 1995;311:351-356.
21. Wolf A, Nasser J, Wolf AM, Schorling JB. The impact of informed consent on patient interest in prostate-specific antigen screening. *Arch Intern Med*. 1996;156:1333-1336.
22. Flood AB, Wennberg JE, Nease RF, et al. The importance of patient preference in the decision to screen for prostate cancer. *J Gen Intern Med*. 1996;11:342-349.
23. Drummond MF, Richardson WS, O'Brien BJ, Levine M, Heyland D, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, XIII: how to use an article on economic analysis of clinical practice, A: are the results of the study valid? *JAMA*. 1997;277:1552-1557.
24. O'Brien BJ, Heyland D, Richardson WS, Levine M, Drummond MF, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, XIII: how to use an article on economic analysis of clinical practice, B: what are the results and will they help me in caring for my patients? *JAMA*. 1997;277:1802-1806.
25. Salzmann P, Kerlikowske K, Phillips K. Cost-effectiveness of extending screening mammography guidelines to include women 40-49 years. *Ann Intern Med*. 1997;127:955-965.
26. Yin D, Carpenter JP. Cost-effectiveness of screening for asymptomatic carotid stenosis. *J Vasc Surg*. 1998;27:245-255.
27. Frame PS, Fryback DG, Patterson C. Screening for abdominal aortic aneurysm in men ages 60 to 80 years: a cost-effectiveness analysis. *Ann Intern Med*. 1993;119:411-416.

# Users' Guides to the Medical Literature: XVII. How to Use Guidelines and Recommendations About Screening

Alexandra Barratt; Les Irwig; Paul Glasziou; et al.

*JAMA*. 1999;281(21):2029-2034 (doi:10.1001/jama.281.21.2029)

http://jama.ama-assn.org/cgi/content/full/281/21/2029

# Users' Guides to the Medical Literature

## XVIII. How to Use an Article Evaluating the Clinical Impact of a Computer-Based Clinical Decision Support System

Adrienne G. Randolph, MD, MSc
R. Brian Haynes, MD, PhD
Jeremy C. Wyatt, MD
Deborah J. Cook, MD, MSc
Gordon H. Guyatt, MD, MSc

### CLINICAL SCENARIO

It is 7 AM, and medical rounds are starting on university hospital ward 3B. In the past 24 hours of your residency, you have transferred 2 critically ill patients to the intensive care unit; accepted 11 patients to your medical service; examined and revised medication orders for 22 patients; placed 9 intravascular catheters; written 35 notes; and reviewed, categorized, and acted on more than 300 new pieces of laboratory and radiology data. You were planning to ask the infectious disease specialist about a patient, but he seems very busy, and the broad-spectrum antibiotic regimen you prescribed should suffice. You were just told that you ordered total parenteral nutrition for the wrong patient. While deciding which patient should receive parenteral nutrition, you realize that the calculations for the amino acid concentration are erroneous. After the first 5 minutes of your first patient presentation, the senior physician asks you details from the patient's past medical history. You wish you could refer to your admission note, but you couldn't access it before your rounds because a utilization review clerk had the chart.

The chair of medicine keeps promising to install computers to help manage all of this information, but she is limited by the budget squeeze. She needs proof that computerization will improve patient care to justify such a major expense. She asks you to help. You remember reading, in one of the many journals piled up at home, about how computers can be used to provide decision support leading to improved patient outcomes. If you can show that computers improve patient care, maybe the hospital administration will see the expense as an investment that could reduce costs.

### THE SEARCH

When you get home that night, you connect to the Internet and decide to search the medical literature using Internet Grateful Med from the US National Library of Medicine. You type *http://igm.nlm.nih.gov/* into your browser and choose MEDLINE. You quickly realize that you don't know what search terms to use. You enter *decision* then click the button for *Find MeSH/Meta Terms*. From the 31 Medical Subject Headings terms offered, you choose *decision making, computer-assisted*; *therapy, computer assisted*; *diagnosis, computer-assisted*; *drug therapy, computer-assisted*, specifying that they are the major topics of the article. You limit your search to randomized controlled trials in English during the years 1995 to 1998. Browsing through the 45 abstracts from the search, you choose "A Randomized Trial of 'Corollary Orders' to Prevent Errors of Omission." The abstract of this article concludes that "physician work stations, linked to a comprehensive electronic medical record, can be an efficient means for decreasing errors of omissions and improving adherence to practice guidelines."[1]

You order the full article over the Internet from Loansome Doc. In this study[1] conducted on the inpatient general medical wards of an inner-city public hospital, 6 independent services (red service, green service, etc) cared for the inpatients. Each service included a faculty internist, a senior resident, and 2 interns. A different physician team rotated onto each service every 6 weeks, and during a year, 8 different teams worked on each service. At the beginning of the study, the investigators randomly allocated 3 of the 6 services to the intervention group,

**Author Affiliations:** Departments of Pediatrics and Anesthesia, Children's Hospital and Harvard Medical School, Boston, Mass (Dr Randolph); Departments of Clinical Epidemiology and Biostatistics and Medicine, McMaster University, Hamilton, Ontario (Drs Haynes, Cook, and Guyatt); and School of Public Policy, University College London, London, England (Dr Wyatt).

The original list of members (with affiliations) appears in the first article of this series (*JAMA*. 1993; 270:2093-2095). A list of new members appears in the 10th article of the series (*JAMA*. 1996;275:1435-1439). The following members contributed to this article: Anne Holbrook, MD, PharmD, MSc; Virginia Moyer, MD, MPH; W. Scott Richardson, MD; David L. Sackett, MD, MSc.

**Corresponding Author and Reprints:** Gordon H. Guyatt, MD, MSc, McMaster University Health Sciences Centre, 1200 Main St W, Room 2C12, Hamilton, Ontario, Canada L8N 3Z5.

**Users' Guides to the Medical Literature Section Editor:** Drummond Rennie, MD, Deputy Editor (West), *JAMA*.

which had access to a computer-based clinical decision support system (CDSS); the other 3 services served as controls and did not have access to a CDSS. Teams were randomly assigned to the intervention and control services. The CDSS responded to trigger orders by suggesting corollary orders needed to detect or ameliorate adverse reactions and allowed physicians to accept or reject these suggestions. TABLE 1 shows examples of corollary orders and their trigger orders.

## CLINICAL COMPUTER SYSTEMS

Clinicians who manage the care of patients are dependent on computers. Laboratory data management software, pharmacy information management systems, applications for tracking patient location through admission and discharge, mechanical ventilators, and oxygen saturation measurement devices are among the many types of computerized systems that have become an integral part of the modern hospital. These devices and systems capture, transform, display, or analyze data for use in clinical decision making. Using computers to search the medical literature or to improve the legibility, display, and accessibility of information in the patient's chart may produce benefits that can sometimes be related to the care of an individual patient. However, medical literature databases and ordinary patient charting systems do not filter and abstract information from detailed clinical data. We use the term *CDSS* to describe software designed to directly aid in clinical decision making about individual patients. Specifically, detailed individual patient data are input into a computer program that sorts and matches them using programs or algorithms in a knowledge base, resulting in the generation of patient-specific assessments or recommendations for clinicians.[2] TABLE 2 shows functions of decision support systems developed for the following medical purposes: alerting, reminding, critiquing, interpreting, predicting, diagnosing, assisting, and suggesting.[3]

Many alerting, reminding, and critiquing systems are based on simple *if-then* rules that tell the computer what to do when a certain event occurs. Alerting systems monitor a continuous signal or stream of data and generate a message (an alert) in response to items or patterns that might require action on the part of the care provider.[4] A simple example of an alert is the starred (*) or highlighted item (with H or L marking or with BOLD or changed colors on the screen) that alerts the clinician to values that are out of range on computerized laboratory printouts and display screens. Alerting systems draw attention to events as they occur. Reminder systems notify clinicians of important tasks that need to be done before an event occurs. An outpatient clinic reminder system may generate a list of immunizations that each patient on the daily schedule requires. Although the technical rules that generate alerts and reminders are often simple, alerting the right person in a timely fashion is quite complex.

When the clinician has made a decision and the computer evaluates that decision and generates an appropriateness rating or alternative suggestion, the decision support approach is called critiquing. The distinction between assisting and critiquing decision support programs is that assisting programs help formulate the clinical decision, whereas critiquing programs have no part in suggesting the order or plan but evaluate the plan, after it is entered, against an algorithm in the computer.[3] Critiquing systems are commonly applied to physician order entry. For example, a clinician entering an order for a blood transfusion may receive a message stating that the pa-

**Table 1.** Example Trigger and Corollary Orders

| Trigger Orders | Corollary Orders |
|---|---|
| Heparin infusion | Platelet count once before heparin starts, then every 24 h |
| | Activated partial thromboplastin time at start, again 6 h after a dosage change |
| | Prothrombin time once before heparin started |
| | Hemoglobin at start of therapy, then every morning |
| | Test stools for occult blood while administering heparin |
| Intravenous fluids | Place a saline lock when intravenous fluids are discontinued |
| Narcotics (class II) | Docusate if not taking any other stool softener or laxative |
| Nonsteroidals | Creatinine level (if not 1 in previous 10 d); SMA-12,* blood urea nitrogen counted as equivalent |
| Aminoglycosides | Peak and trough levels after dosage changes and every week |
| | Creatinine level twice per week (every Monday and Thursday) |
| Warfarin sodium | Prothrombin time each morning |
| Amphotericin B | Creatinine level twice per week (every Monday and Thursday) |
| | Magnesium level (twice per week while receiving therapy) |
| | Electrolytes (twice per week while receiving therapy) |
| | Acetaminophen (650 mg by mouth 30 min before each dose) |
| | Diphenhydramine hydrochloride (50 mg 30 min before each amphotericin dose) |

*SMA-12 indicates sequential multiple analyzer, measuring glucose, blood urea nitrogen, uric acid, calcium, phosphorus, total protein, albumin, cholesterol, total bilirubin, alkaline phosphatase, serum glutamic oxaloacetic transaminase, and lactate dehydrogenase.

**Table 2.** Functions of Computer-Based Clinical Decision Support Systems

| Function | Example |
|---|---|
| Alerting | Highlighting out-of-range laboratory values |
| Reminding | Reminding the clinician to schedule a mammogram |
| Critiquing | Rejecting an electronic order |
| Interpreting | Interpreting the electrocardiogram |
| Predicting | Predicting risk of mortality from a severity-of-illness score |
| Diagnosing | Listing a differential diagnosis for a patient with chest pain |
| Assisting | Tailoring the antibiotic choices for liver transplantation and renal failure |
| Suggesting | Generating suggestions for adjusting the mechanical ventilator |

tient's hemoglobin level is above the transfusion threshold, and the clinician must justify the order by stating an indication, such as active bleeding.[5] Getting the attention of the person who can take action is one of the most difficult aspects of making alerting, reminding, and critiquing systems effective.

The automated interpretations of electrocardiogram readings[6] and the outcome predictions generated by severity-of-illness scoring systems[7] are examples of decision support systems used for interpreting and predicting, respectively. These systems filter and abstract detailed clinical data and generate a report characterizing the meaning of the data (eg, anterior myocardial infarction).[6]

Computer-aided diagnostic systems assist the clinician with the process of differential diagnosis.[8] When the electrocardiogram results are not definitive, computer systems that try to distinguish between myocardial infarction and other sources of chest pain can sometimes outperform a clinician.[9] These types of systems require pertinent patient information, such as signs, symptoms, past medical history, laboratory values, and demographic characteristics. The programs start generating hypotheses, often prompt the user for more information, and ultimately provide a diagnosis or a list of possible diagnoses ranked probabilistically.

Computerized patient management systems are complex programs that make suggestions about the optimal decision based on the information currently known by the system. These types of systems are often integrated into the physician ordering process. After collecting information on specific patient variables, the assistant program tailors the order to the patient based on prior information in the database regarding appropriate dosages or by implementing specified protocols. The Antibiotic Assistant[10] is a CDSS that implements guidelines to assist physicians with ordering antibiotics. This system recommends the most cost-effective antibiotic regimen taking into account the patient's renal function, drug allergies, the site of infection, the epidemiology of organisms in patients with this infection at this hospital over many years, the efficacy of the antibiotic regimen, and the cost of therapy. A system that instructs caregivers about how to manage the ventilation of patients with adult respiratory distress syndrome[11] is another example.

The primary reason to invest in computer support is to improve quality of care. If a computer system purports to aid clinical decisions, enhance patient care, and improve outcomes, then it should be subject to the same rules of testing as any other health care intervention with similar claims. In this article, we describe how to use articles that evaluate the clinical impact of a CDSS. While the focus of a CDSS may be restricted to diagnosis or prognosis, we will limit our discussion to the situation in which the CDSS is designed to change clinician behavior and patient outcome. Many iterative steps are involved in developing, evaluating, and improving a CDSS before it can progress beyond the laboratory environment and pilot-testing phase and be allowed to have a wider impact on physicians and patients. These evaluations involve social science methods for evaluating human behavior and computer science methods for evaluating technological safety and robustness.[4] We limit our discussion to mature systems that have surpassed initial evaluation and are being implemented to change physician behavior and patient outcome.

## Are the Results of the Study Valid?

When clinicians examine the effect of a CDSS on patient management or outcome, they should use the same criteria appropriate for any other intervention (TABLE 3), whether it be a drug, a rehabilitation program, or an approach to diagnosis or screening.[12] In our Users' Guide to prevention and therapy,[13] the importance of random assignment, blinding of patients and outcome assessors, and complete follow-up were explained. The purpose of our discussion in this article is to highlight issues of particular importance in the evaluation of a CDSS.

**Table 3.** Using Articles Describing Computer-Based Clinical Decision Support Systems (CDSSs)

| |
|---|
| **Are the results of the study valid?** |
| Was the method of participant allocation appropriate? |
| Was the control group uninfluenced by the CDSS? |
| Aside from the CDSS, were the groups treated equally? |
| **What were the results?** |
| What was the effect of the CDSS? |
| **Can you apply the computer-based CDSS in your clinical setting?** |
| What elements of the CDSS are required? |
| Is the CDSS exportable to a new site? |
| Is the CDSS likely to be accepted by clinicians in your setting? |
| Do the benefits of the CDSS justify the risks and costs? |

**Was the Method of Participant Allocation Appropriate?** The validity of the observational study designs often used to evaluate a CDSS is limited. The most common observational design is the before-after study design, in which investigators compare outcomes before a technology is implemented (using a historic control group) with those after the system is implemented. The validity of this approach is threatened by the possibility that changes over time (called secular trends) in patient mix or in aspects of health care delivery may result in changes in behavior that appear to be attributable to the CDSS. Consider a CDSS that assisted physicians with antibiotic ordering[10] in the late 1980s and was associated with improvements in the cost-effectiveness of antibiotic ordering over the next 5 years. Changes in the health care system, including the advent of managed care, were occurring simultaneously during that time. To control for secular trends, the computerized antibiotic practice guideline study investigators[10] compared antibiotic prescribing practices with those of other nonfederal US acute care hospitals for the duration of the study.

One type of time-series design, in which the intervention is turned on and

off multiple times, has been used to control for potential secular trends. Although this provides some protection against bias, random allocation of patients to a concurrent control group remains the strongest study design for evaluating therapeutic or preventive interventions.[13] Use of historical controls may lead to a higher tendency to see positive results. A comparison of the 2 types of studies used to evaluate the same antihypertensive drugs revealed that 80% of historically controlled studies suggested that the new drugs were effective, whereas only 20% of randomized controlled trials confirmed this result.[14] Randomized controlled trials have been successfully used to evaluate more than 70 CDSSs.[2,15-17]

An important issue for CDSS evaluation is the unit of allocation. Investigators in clinical trials usually randomize patients. When evaluating the effect of a CDSS on patient care, the intervention is usually aimed at changing the decision making of the clinician, so investigators may randomize individual clinicians or clinician clusters such as health care teams, hospital wards, or outpatient practices.[18] A common mistake made by investigators is to analyze their data as if they had randomized patients rather than clinicians. This is called a *unit of analysis error*.[19]

To highlight the problem, we will use an extreme example. Investigators randomize study participants to ensure that treatment and control groups are balanced with respect to important predictors of outcome. Randomization often fails to balance groups if sample size is small. Consider a study in which an investigator randomizes one team of clinicians to a CDSS and another to standard practice. During the course of the study, each team sees 10 000 patients. If the investigator analyzes the data as if patients were individually randomized, the sample size appears huge (the unit of analysis error[19]). However, it is very plausible, perhaps even likely, that the 2 teams' performance differed at the start and that this difference persisted through the study independent of the CDSS. Because the base sample size in this study is only 2 (2 teams), the likelihood of imbalance despite randomization is very large.

When investigators randomize physicians and health care teams, obtaining a sample of sufficient size can be difficult. If only a few health care teams are available, stratification of these teams according to important prognostic factors can reduce potential imbalances. If there are many known risk factors, investigators can pair health care teams according to their similarities and randomly allocate the intervention within each matched pair.[20] In addition, investigators can use statistical methods developed specifically for analyzing studies using cluster randomization.[21]

There is one other issue regarding randomization to which clinicians should attend. If some clinicians assigned to CDSS fail to receive the intervention, should these clinicians be included in the analysis?

The answer, counterintuitive to some, is yes. Randomization can accomplish the goal of balancing groups with respect to both known and unknown determinants of outcome only if patients (or clinicians) are analyzed in the groups to which they are randomized. Deleting or moving patients after randomization compromises or destroys the balance that randomization is designed to achieve. An analysis in which patients are included in the groups to which they were randomized, whether or not they received the intervention, is called *intention to treat*.[13]

In the study by Overhage et al,[1] during the course of a year, there were 36 teams randomly assigned to 18 CDSSs and 18 control services. House staff were required to write all orders and were used as the unit of analysis. Each service admitted patients in sequence, so that all 6 services received equal numbers of patients. A total of 86 house staff physicians who each received more than 5 corollary orders during the study cared for 2181 different patients during 2955 different admissions.

Random assignment of teams to CDSS and non-CDSS services increases our belief that the results are valid. However, although investigators did not randomly assign house staff to services, they conducted their analysis at the individual house staff level, comparing 45 intervention physicians with 41 control physicians. They took no steps to ensure that the characteristics of house staff on the intervention and control teams were similar, leaving the study open to biases from baseline differences in house staff performance. Moreover, the use of individual house staff instead of the team as the unit of analysis may have led to false precision in estimating the impact of the intervention because of a falsely inflated sample size.

In the study by Overhage et al,[1] investigators excluded 6 physicians from the intervention group because those physicians received fewer than 5 suggestions about corollary orders. This decision violates the intention-to-treat principle and risks introducing bias, because physicians on the control side who received fewer than 5 suggestions were included. Fortunately, the small number of excluded physicians were mostly off-service physicians covering night calls for 1 or 2 nights and not actually service team members, so the contribution of such physicians to the comparison of CDSS and control is small.

**Was the Control Group Uninfluenced by the CDSS?** One problem with performing a controlled trial randomizing a CDSS across patients is the difficulty in controlling for contamination of the control group by the intervention. Strickland and Hasson[22] randomly allocated patients to have changes in their level of mechanical ventilator support either directed by a computer protocol and implemented through a physician or directed by the physician independently. Because the same physicians and respiratory therapists who used the computer protocol managed the care of patients not assigned to the protocol, it is possible clinicians remembered and applied protocol algorithms in control patients. When the control group is influenced by the intervention, the effect of the CDSS may be diluted. Contamination

may spuriously decrease, or even eliminate, a true intervention effect.

One method of preventing exposure of the control group to the CDSS is to assign individual clinicians to use or not use the CDSS. This is often problematic because of cross-coverage of patients. Comparing the performance of wards or hospitals that do or do not use the CDSS is another possibility. Unfortunately, it usually is not feasible to enroll a sufficient number of hospitals in a study to avoid the problem we described earlier—when sample size is small, randomization may fail to ensure prognostically similar groups.

In the study by Overhage et al,[1] physicians whose teams were assigned to a control service had the CDSS guidelines available on paper but did not receive assistance when ordering. To control for the risk that cross-coverage of patients could expose the control group to the CDSS, the investigators had the chief medical resident construct the residents' evening call schedule to separate coverage for patients based on patients' study status. If switches in the schedule were made, control physicians provided call coverage only for non-CDSS patients, and intervention physicians covered only CDSS patients. Furthermore, to avoid contamination that could occur if intervention physicians cared for control patients, the computer suggested orders only when the patient had been assigned to a physician in the CDSS group, and corollary order suggestions were suppressed if the patient was assigned to the control group. If physicians returned for a second rotation and changed study status, the investigators excluded data from their second rotation. All of these efforts were to prevent contamination of the control group by the CDSS.

**Aside From the CDSS, Were the Groups Treated Equally?** The results of studies evaluating interventions aimed at therapy or prevention are more believable if patients, their caregivers, and study personnel are blind to the treatment.[13] Unblinded study personnel who are measuring outcomes may provide different interpretations of marginal findings or differential encouragement during performance tests.[23] Blinding also diminishes the placebo effect,[13] which, in the case of CDSS, may be the tendency of patients or clinicians to ascribe positive attributes to use of a computer workstation.[4] Although blinding the clinicians, patients, and study personnel to the presence of the computer-based CDSS may prevent this type of bias, blinding is sometimes not possible.

Interventions other than the treatment being studied that can influence the outcome are called *cointerventions*. They frequently occur because most patients receive multiple therapies aimed at improving their outcome. A problem arises when cointerventions are differentially applied to the treatment and control groups. This situation is more likely to arise in unblinded studies, particularly if the use of very effective nonstudy treatments is permitted at physicians' discretion.[13] Clinicians' concerns regarding lack of blinding are ameliorated if investigators describe permissible cointerventions and their differential use and/or standardize cointerventions[24] to ensure that their application is similar in both treatment and control groups.

It is also important to ensure that the evaluation of the outcome for each group is not biased. In some studies, the computer system may be used as a data collection tool to evaluate the outcome in the CDSS group. The "data completeness bias" can occur when the information system is used to log episodes in the treatment group and a manual system is used to log episodes in the non-CDSS group.[4] Because the computer may log more episodes than the manual system, it may appear that the CDSS group had more events, which could bias the outcome in favor of or against the CDSS group. To prevent this bias, outcomes should be logged similarly in both groups.

In the study by Overhage et al,[1] although faculty were proscribed from writing orders except during emergencies, physicians practiced within teams, and the faculty influenced the residents through their teaching. Faculty could rotate with different house staff on different rotations during the study, further complicating this situation. To allow for this clustering of physicians within teams, the investigators used generalized estimating equations to control for potential cointervention.

## What Are the Results?

**What Is the Effect of the CDSS?** A CDSS is often aimed at preventing adverse events or health outcomes or at improving compliance with a treatment regimen. (See our Users' Guide for prevention or therapy[13] for a discussion of relative risk and relative risk reductions, risk differences and absolute risk reductions, and confidence intervals.) In the study by Overhage et al,[1] intervention physicians ordered the corollary orders suggested by the CDSS much more frequently than control physicians spontaneously ordered them. This was true when measured by immediate compliance (46.3% vs 21.9%; relative increase, 2.11; $P<.0001$), 24-hour compliance (50.4% vs 29.0%; relative increase, 1.74; $P<.0001$), or hospital-stay compliance (55.9% vs 37.1%; relative increase, 1.51; $P<.0001$). Because the numerators and denominators are not reported for the total numbers of corollary orders complied with and not complied with for each group, we cannot calculate the confidence intervals for the risk difference for the increase in compliance. However, because the $P$ values are very small, we know that the lower boundary of the confidence interval is appreciably greater than 1, and the confidence interval is therefore relatively narrow.

Length of stay and hospital charges did not differ significantly. Pharmacists made 105 interventions with the CDSS group of physicians and 156 with control physicians (2-tailed $P = .003$) for errors considered to be life-threatening, severe, or significant.

## Can You Apply the CDSS in Your Clinical Setting?

**What Elements of the CDSS Are Required?** Investigators should specify the

intervention that they are evaluating. Two of the major elements of a CDSS, the logic and the computer interface used to present the logic, could each be evaluated as a separate intervention. However, sometimes it is not possible to separate these 2 elements and achieve the same result. For example, we mentioned a randomized controlled trial that compared a computerized protocol for managing patients diagnosed as having adult respiratory distress syndrome with standard clinical care using extracorporeal carbon dioxide removal as rescue therapy.[11] The computerized protocol group without rescue therapy did as well as the rescue therapy group. Was this due to the logic in the protocol, the use of the computer, or both interacting together?

To test whether the computer is needed requires that one group apply the protocol logic as written on paper and the other group use the same logic implemented by the computer. Sometimes the protocol logic is so complex that use of a computer may be required for implementation.

The CDSS may have a positive impact for unintended reasons. The impact of structured data collection forms and performance evaluations (the Checklist Effect and the Feedback Effect,[4] respectively) on decision making can equal that of computer-generated advice.[25] The CDSS intervention itself may be administered by research personnel or by paid clinical staff who receive scant mention in the published report but without whom the impact of the system is seriously undermined.

The CDSS in the study by Overhage et al of corollary orders[1] and in the adult respiratory distress syndrome study[11] had 3 components: a knowledge base defining which corollary orders were required for each trigger order, a database that stored the trigger orders, and an inference engine that compared the database with the knowledge base when a trigger order was received and sent a list of suggested corollary orders to the computer terminal for display.

**Is the CDSS Exportable to a New Site?** For a CDSS to be exported to a new site, it has to be able to integrate with existing software, users at the new site must be able to maintain the system, and users must accept the system. Double-charting occurs when systems require staff (usually nurses) to enter the data twice—into the computer and again on a flow sheet. Systems that require double-charting increase staff time devoted to documentation, frustrate users, and divert time that could be devoted to patient care. In general, such systems fail in clinical use.

Successful systems usually have automatic electronic interfaces to existing data-producing systems. Unfortunately, building interfaces to diverse computer systems is often challenging and sometimes impossible.

The program described in the study by Overhage et al[1] was implemented using the Regenstrief Medical Record System developed at Indiana University School of Medicine. This system provides an electronic medical record system and a physician order entry system. While it may be possible to use the knowledge built into the system in a health care environment in which the patient population is similar, the inference engine used to compare the rules with the order entered into the database is not easily exported to other locations. If, after critically appraising the article, you are convinced that a CDSS for implementing guidelines would be useful, you would need sufficient resources to rebuild the system at your own site.

**Is the CDSS Likely to Be Accepted by Clinicians in Your Setting?** A CDSS may not be accepted if the clinicians differ in important ways from those who participated in the study. The choice of evaluative group may limit the generalizability of the conclusions if recruitment is based upon enthusiasm, demographics, or a zest for new technology. Clinicians in a new setting may be surprised when their colleagues do not use a CDSS with the same avidity as the original participants.

The user interface is an important component of the effectiveness of a CDSS. The CDSS interface should be developed on the basis of potential users' capabilities and limitations, the users' tasks, and the environment in which those tasks are performed.[26] One of the main difficulties with alerting systems is notifying the individual with decision-making capability as rapidly as possible that there is an abnormal laboratory value or other potential problem. A group of investigators tried a number of different alerting methods, from a highlighted icon on the computer screen to a flashing yellow light placed on the top of the computer.[27] These investigators later gave the nurses pagers to alert them to abnormal laboratory values.[28] The nurses could then decide how to act on the information and when to alert the physician.

To ensure user acceptance, users must feel that they can depend on the system to be available whenever they need it. The amount of downtime needed for data backup, troubleshooting, and upgrading should be minimal. The response time must be fast, data integrity must be maintained, and data redundancy must be minimized. If systems have been functioning at other sites for a period of time, major problems or software bugs may have been eradicated, decreasing downtime and improving acceptance. Investigators should also assess the amount of training required for users to feel comfortable with the system. If users become frustrated, system performance will be suboptimal.

Many computer programs may function well at the site where the program was developed; unfortunately, the staff at your own institution may have objections to the approaches taken elsewhere. For example, an expert system for managing ventilated patients who have adult respiratory distress syndrome may use continuous positive airway pressure trials to wean patients off the ventilator, whereas clinicians at your institution may prefer pressure-support weaning. Syntax, laboratory coding and phrasing of diagnoses, and therapeutic interventions can vary markedly among institutions. Customizing the application to the environ-

ment may not be feasible, and additional expense may be invoked when mapping vocabulary to synonyms unless a mechanism to do so is already programmed into the system. To ensure user acceptance, the needs and concerns of users should be considered, and users should be included in decision making and implementation stages.

The logic in the Regenstrief Order Entry system[1] was based on the expertise of a hospital committee of staff physicians and pharmacists. Although the investigators used reference texts, the degree to which they applied an evidence-based approach is unclear. Use of solid evidence[29] from the literature could enhance clinician acceptance by convincing physicians that the rules positively affect patient outcomes. However, gaining consensus even with evidence-based practices can be difficult and a method for gaining consensus must be integrated into the local processes and culture of care. Furthermore, physicians will need some time to become acquainted with any new system, especially an order entry system.

When the study by Overhage et al began, all physicians on the medical wards had been entering all inpatient orders directly into physician workstations for 12 months. Because the order entry program was developed over time and refined by user input, it was tailored to the needs of the clinicians at that hospital. Whether this system would be easily accepted in a new environment by clinicians who had nothing to do with its development is open to question.

**Do the Benefits of the CDSS Justify the Risks and Costs?** Does the report reveal the behind-the-scenes costs? The real cost of the CDSS is usually much higher than the initial hardware, software, interface, training, maintenance, and upgrade costs (which may not be in the report). Often the CDSS is designed and maintained by staff whose actions are critical to the success of the intervention. An institution might not want to pay for the time of such people in addition to the cost of the computer software and hardware. Indeed, it can be very difficult to estimate the costs of purchasing or building and implementing an integrated CDSS.

**Are CDSSs Beneficial?** Human performance may improve when participants are aware that their behavior is being observed (the Hawthorne effect[30]); the same behavior may not be exhibited when the monitoring of outcomes has stopped. Taking into account the influence of a study environment, a recently updated[17] published systematic review of studies assessing CDSSs used in inpatient and outpatient clinical settings by health care providers[2] showed that the majority of CDSSs studied were beneficial. The review assessed patient-related outcomes (eg, mortality, length of hospital stay, decrease in infections) or health care process measures (eg, compliance with reminders or with evidence-based processes of care). A total of 68 prospective trials using concurrent control groups have reported the effects of using CDSSs on drug dosing, diagnosis, preventive care, and active medical care. Forty-three (66%) of 65 studies showed that CDSSs improved physician performance. These included 9 of 15 studies on drug dosing systems, 1 of 5 studies on diagnostic aids, 14 of 19 preventive care systems, and 19 of 26 studies evaluating CDSSs for active medical care. Six (43%) of 14 studies showed that CDSSs improved patient outcomes, 3 studies showed no benefit, and the remaining studies lacked sufficient power to detect a clinically important effect.

Health care processes are more often evaluated than patient health outcomes because process events occur more frequently. For example, a trial designed to show a 25% improvement (from 50% to 62.5%) in the proportion of patients who are compliant with a certain medication regimen would need to enroll 246 patients per group. A trial designed to show that this medication reduces mortality by 25% (from 5% to 3.75%) would need to enroll 4177 patients per group. Furthermore, long follow-up periods are required to show that preventive interventions improve patient health outcomes.

Fortunately, evaluation of health care processes will adequately infer benefit if the care processes being monitored are already known to improve outcomes.[31] We could conclude that a CDSS that increased the frequency with which aspirin, β-blockers, and angiotensin-converting enzyme inhibitors were administered to appropriate patients after myocardial infarction was beneficial, because large, well-designed randomized trials have demonstrated the benefit of these 3 interventions. Unfortunately, the link between processes and outcomes is often weak or unknown.

The study by Overhage et al[1] was able to demonstrate that a physician workstation, when linked to an order entry system able to run a series of rules, is an efficient means for decreasing errors of omission and improving adherence to practice guidelines. It is unclear how many of the rules in the system were based on solid evidence and thus how likely it is that compliance with rules will improve outcomes. Therefore, it is unclear whether the benefits are worth the cost of purchasing, configuring, installing, and maintaining the CDSS.

## RESOLUTION OF THE SCENARIO

A computer-based CDSS evaluation involves the interplay between 3 elements: 1 or more human intermediaries, an integrated computerized system and its interface, and the knowledge in the decision support system. This makes evaluation of a computer-based CDSS a complex undertaking. Systematic reviews[32] of the impact of a CDSS on provider behavior and patient outcome have shown evidence of benefit.[2,15-17] Because the evaluation process for these reviews was not standardized, it is difficult to compare the results.

We have described a process of evaluating articles that aims to measure the impact of a computer-based CDSS on provider decisions or patient out-

comes. Despite the complexity of evaluation, clinicians can use basic principles of evidence-based care to evaluate CDSSs. A study evaluating a CDSS is more believable if there is a concurrent control group with a random allocation of subjects. Randomization of clinicians by clusters can prevent the cross-contamination of the control group by the intervention that could mask the effect of the CDSS. When using multilevel designs (composed of the physician or physician group and their respective patients) investigators should treat the physician or group, not the patients, as the unit of analysis. Because most studies evaluating CDSSs are not blinded, we stressed the importance of controlling for cointerventions that could bias the outcome.

Even if the study is valid and a positive effect is shown, CDSSs have special applicability issues that must be considered. Is the computer essential to deployment of the knowledge in the CDSS? Can the CDSS be exported to a new site? Will clinicians accept the CDSS? And, finally, is it possible to accurately evaluate the cost of the CDSS when assessing risks and benefits?

## REFERENCES


1. Overhage JM, Tierney WM, Zhou XH, McDonald CJ. A randomized trial of "corollary orders" to prevent errors of omission. *J Am Med Inform Assoc*. 1997; 4:364-375.
2. Johnston ME, Langton KB, Haynes RB, Mathieu A. Effects of computer-based clinical decision support systems on clinician performance and patient outcome: a critical appraisal of research. *Ann Intern Med*. 1994; 120:135-142.
3. Pryor TA. Development of decision support systems. *Int J Clin Monitoring Computing*. 1990;7:137-146.
4. Friedman CP, Wyatt JC. *Evaluation Methods in Medical Informatics*. New York, NY: Springer-Verlag; 1997.
5. Lepage EF, Gardner RM, Laub RM, Golubjatnikov OK. Improving blood transfusion practice: role of a computerized hospital information system. *Transfusion*. 1992;32:253-259.
6. Weinfurt PT. Electrocardiographic monitoring: an overview. *J Clin Monitoring*. 1990;6:132-138.
7. Knaus WA, Wagner DP, Draper EA, et al. The APACHE III prognostic system: risk prediction of hospital mortality for critically ill hospitalized adults. *Chest*. 1991;100:1619-1636.
8. Berner ES, Webster GD, Shugerman AA, et al. Performance of four computer-based diagnostic systems. *N Engl J Med*. 1994;330:1792-1796.
9. Kennedy RL, Harrison RF, Burton AM, et al. An artificial neural network system for diagnosis of acute myocardial infarction (AMI) in the accident and emergency department: evaluation and comparison with serum myoglobin measurements. *Comput Methods Programs Biomed*. 1997;52:93-103.
10. Evans RS, Pestotnik SL, Classen DC, et al. A computer-assisted management program for antibiotics and other antiinfective agents. *N Engl J Med*. 1998;338:232-238.
11. Morris AH, Wallace CJ, Menlove RL, et al. Randomized clinical trial of pressure-controlled inverse ratio ventilation and extracorporeal $CO_2$ removal for adult respiratory distress syndrome. *Am J Respir Crit Care Med*. 1994;149(2 pt 1):295-305.
12. Spiegelhalter DJ. Evaluation of clinical decision-aids, with an application to a system for dyspepsia. *Stat Med*. 1983;2:207-216.
13. Guyatt GH, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, II: how to use an article about therapy or prevention, A: are the results of the study valid? *JAMA*. 1993;270:2598-2601.
14. Sacks H, Chalmers TC, Smith H Jr. Randomized versus historical controls for clinical trials. *Am J Med*. 1982;72:233-240.
15. Shea S, DuMouchel W, Bahamonde L. A meta-analysis of 16 randomized controlled trials to evaluate computer-based clinical reminder systems for preventive care in the ambulatory setting. *J Am Med Inform Assoc*. 1996;3:399-409.
16. Balas EA, Austin SM, Mitchell JA, Ewigman BG, Bopp KD, Brown GD. The clinical value of computerized information services: a review of 98 randomized clinical trials. *Arch Fam Med*. 1996;5:271-278.
17. Hunt DL, Haynes RB, Hanna SE, Smith K. Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. *JAMA*. 1998;280:1339-1346.
18. Cornfield J. Randomization by group: a formal analysis. *Am J Epidemiol*. 1978;108:100-102.
19. Whiting-O'Keefe QE, Henke C, Simborg DW. Choosing the correct unit of analysis in medical care experiments. *Med Care*. 1984;22:1101-1114.
20. Klar N, Donner A. The merits of matching in community intervention trials: a cautionary tale. *Stat Med*. 1997;16:1753-1764.
21. Thompson SG, Pyke SD, Hardy RJ. The design and analysis of paired cluster randomised trials: an application of meta-analysis techniques. *Stat Med*. 1997; 16:2063-2079.
22. Strickland JH Jr, Hasson JH. A computer-controlled ventilator weaning system: a clinical trial. *Chest*. 1993;103:1220-1226.
23. Guyatt GH, Pugsley SO, Sullivan MJ, et al. Effect of encouragement on walking test performance. *Thorax*. 1984;39:818-822.
24. Morris AH, East TD, Wallace CJ, et al. Standardization of clinical decision making for the conduct of credible clinical research in complicated medical environments. *Proc AMIA Annu Fall Symp*. October 1996:418-422.
25. Adams ID, Chan M, Clifford PC, et al. Computer aided diagnosis of acute abdominal pain: a multicentre study. *BMJ*. 1986;293:800-804.
26. Salvemini AV. Improving the human-computer interface: a human factors engineering approach. *MD Comput*. 1998;15:311-315.
27. Bradshaw KE, Gardner RM, Pryor TA. Development of a computerized laboratory alerting system. *Comput Biomed Res*. 1989;22:575-587.
28. Tate KE, Gardner RM, Scherting K. Nurses, pagers, and patient-specific criteria: three keys to improved critical value reporting. *Proc Annu Symp Comput Applications Med Care*. October 1995:164-168.
29. Guyatt GH, Sackett DL, Sinclair JC, et al. Users' guides to the medical literature, IX: a method for grading health care recommendations. *JAMA*. 1995;274: 1800-1804.
30. Roethligsburger FJ, Dickson WJ. *Management and the Worker*. Cambridge, Mass: Harvard University Press; 1939.
31. Mant J, Hicks N. Detecting differences in quality of care: the sensitivity of measures of process and outcome in treating acute myocardial infarction. *BMJ*. 1995;311:793-796.
32. Oxman AD, Cook DJ, Guyatt GH, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VI: how to use an overview. *JAMA*. 1994;272:1367-1371.

# Users' Guides to the Medical Literature: XVIII. How to Use an Article Evaluating the Clinical Impact of a Computer-Based Clinical Decision Support System

Adrienne G. Randolph; R. Brian Haynes; Jeremy C. Wyatt; et al.

*JAMA*. 1999;282(1):67-74 (doi:10.1001/jama.282.1.67)

http://jama.ama-assn.org/cgi/content/full/282/1/67

# Users' Guides to the Medical Literature

## XIX. Applying Clinical Trial Results
## A. How to Use an Article Measuring the Effect of an Intervention on Surrogate End Points

Heiner C. Bucher, MD, MPH
Gordon H. Guyatt, MD, MSc
Deborah J. Cook, MD, MSc
Anne Holbrook, MD, MSc
Finlay A. McAlister, MD
for the Evidence-Based Medicine Working Group

### CLINICAL SCENARIO

You are a physician seeing a 62-year-old woman with postmenopausal osteoporosis. Her bone mineral density, as measured by dual-energy x-ray absorptiometry, is 2.5 SDs below the mean value in premenopausal women. Although she does not have back pain, a spinal radiograph shows an old vertebral fracture. The patient has not yet experienced problems as a result of her vertebral fracture, but she is disturbed by the prospect that she may end up like her mother whose osteoporotic fractures have resulted in severe, long-term back pain.

The patient has reflux esophagitis and a past endoscopy revealed nonspecific gastritis. A specialist had prescribed alendronate, which the patient had to stop taking after several weeks because of dyspepsia. She searched the Web and discovered a new drug, raloxifene, and wonders whether this drug might be an alternative. You know that this drug has been licensed for the prevention of postmenopausal osteoporosis. You promise to examine the literature and to get back to her.

**See also pp 786 and 790.**

### THE SEARCH

Using MEDLINE you identify a study of raloxifene for the treatment of osteoporosis demonstrating an effect on bone mineral density.[1] You are wondering whether this warrants administration to lower your patient's risk of osteoporotic fracture.

### INTRODUCTION

Ideally, clinicians making treatment decisions should refer to methodologically strong clinical trials examining the impact of therapy on clinically important outcomes. By clinically important outcomes we mean outcomes that are important to patients: health-related quality of life, morbid end points such as stroke or myocardial infarction, or death. Often, however, conducting these trials requires such a large sample size, or long-term patient follow-up, that researchers or drug companies look for alternatives. Substituting surrogate end points for the target event allows conduct of shorter and smaller trials, thus offering an apparent solution to the dilemma.

A surrogate end point may be defined as "a laboratory measurement or a physical sign used as a substitute for a clinically meaningful end point that measures directly how a patient feels, functions or survives."[2] Surrogate end points include physiologic variables (such as bone mineral density as a surrogate for long-bone fractures, blood pressure for stroke, low-density lipoprotein cholesterol levels for myocardial infarction, and CD4 cell count for acquired immunodeficiency syndrome [AIDS] and AIDS-related mortality) or measures of subclinical disease (such as degree of atherosclerosis on coronary angiography).

The use of surrogate end points is indispensable for drug evaluation in phase 2 and early phase 3 trials geared to establishing a drug's promise of benefit. In many countries, companies may obtain drug approval by demonstrating a positive impact on surrogate end points. The use of surrogate end points for regulatory purposes reflects drug approval decisions that regulators must make in the face of public health exigencies.

Reliance on surrogate end points may be beneficial or harmful. On the one hand, use of the surrogate end point

**Author Affiliations:** Medizinische Universitäts-Poliklinik, Kantonsspital Basel, Basel, Switzerland (Dr Bucher); Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario (Drs Guyatt, Cook, and Holbrook); and Division of General Internal Medicine, University of Alberta Hospital, Edmonton (Dr McAlister).

The original list of members (with affiliations) appears in the first article of the series (*JAMA*. 1993; 270:2093-2095). A list of new members appears in the 10th article of the series (*JAMA*. 1996;275:1435-1439). The following members of the Evidence-Based Medicine Working Group contributed to this article: Antonio Dans, MD, Leonilla Dans, MD, Pat Brill-Edwards, MD, Daren Heyland, MD, Les Irwig, MBBCh, PhD, FFPHM, Roman Jaeschke, MD, MSc, Hui Lee, MD, MSc, Mitchell Levine, MD, MSc, Virginia Moyer, MD, MPH, and David Naylor, MD, DPhil.

**Corresponding Author and Reprints:** Gordon H. Guyatt, MD, MSc, McMaster University Health Sciences Centre, 1200 Main St W, Room 2C12, Hamilton, Ontario, Canada L8N 3Z5.

**Users' Guides to the Medical Literature Section Editor:** Drummond Rennie, MD, Deputy Editor (West), *JAMA*.

**Table 1.** Users' Guide for a Surrogate End Point Trial

| |
|---|
| Are the results valid? |
| • Necessary, but not sufficient: is there a strong, independent, consistent association between the surrogate end point and the clinical end point? |
| • Is there evidence from randomized trials in other drug classes that improvement in the surrogate end point has consistently led to improvement in the target outcome?* |
| • Is there evidence from randomized trials in the same drug class that improvement in the surrogate end point has consistently led to improvement in the target outcome?* |
| What were the results? |
| • How large, precise, and lasting was the treatment effect? Effect should be large, precise, and lasting to consider a surrogate trial as possible basis for offering patients the intervention. |
| Will the results help me in caring for my patients? |
| • Are the likely treatment benefits worth the potential harms and costs? Offer intervention on basis of surrogate data only if patient's risk of the target outcome is high, patient places a high value on avoiding the target outcome, and if there are no satisfactory alternative therapies. |

*Answers to one or both of these questions should be "yes" for surrogate trial to be an adequate guide for clinical action.

may lead to the rapid and appropriate dissemination of new treatments. For example, the Food and Drug Administration's decision to approve new antiretroviral drugs based on information from trials using surrogate end points recognized the enormous need for effective therapies for patients with human immunodeficiency virus (HIV) infection. Subsequently, several of these drugs have proved effective in randomized trials focusing on clinically important outcomes.[3-6]

On the other hand, reliance on surrogate end points may lead to excess morbidity and mortality. For example, while cardiac inotropes may improve short-term cardiac hemodynamic function in patients with heart failure, randomized clinical trials have demonstrated excess mortality with a number of these agents.[7] In particular, flosequinan was widely prescribed after its release, but had to be withdrawn after a trial revealed its deleterious effects on survival.[8]

How are clinicians to distinguish between these 2 situations? Surrogate outcome will be consistently reliable only if there is a causal connection between change in surrogate and change in the clinically important outcome. Thus, the surrogate must be in the causal pathway of the disease process and an intervention's entire effect on the clinical outcome of interest should be fully captured by a change in the surrogate. This Users' Guide builds on previous discussions of how one can establish a causal relationship[9] and presents an approach to critical appraisal of studies using surrogate end points and application of their results to manage individual patients.

As our discussion will make evident, the clinician needs to assess far more than a single study to make the decision about the adequacy of a surrogate. Evaluation may require a comprehensive review of observational studies of the relationship between the surrogate and the target, and of some or all of the randomized trials that have evaluated treatment impact on both the surrogate and the target. While most clinicians would hesitate to conduct such an investigation, our guidelines will allow them to evaluate the arguments made by experts or the pharmaceutical industry for prescribing treatments on the basis of their effect on surrogate end points.

## THE GUIDES

In this guide, we follow the framework of previous articles in the series[10] and ask 3 sorts of questions: are the results valid; what were the results; and will the results help me in caring for my patients? (TABLE 1). When we consider the validity of a surrogate, we must address 2 issues. First, to be consistently reliable, the surrogate must be in the causal pathway from the intervention to the outcome. Second, in considering a particular intervention, we must be confident that there are no important effects of that intervention on the outcome of interest that are not mediated through, or captured by, the surrogate. Our guides for validity (Table 1) bear directly on these 2 issues.

### Are the Results Valid? Is There a Strong, Independent, Consistent Association Between the Surrogate End Point and the Clinical End Point?

To provide a valid substitute for an important target outcome, the surrogate must be associated or correlated with that target. In general, researchers choose surrogate end points because they have found a correlation between a surrogate and a target outcome in observational studies, and their understanding of the biology makes it plausible that changes in the surrogate will invariably lead to changes in the important outcome. The stronger the association, the more likely the causal link between the surrogate and the target. The strength of an association is reflected in statistics such as relative risk (RR) or odds ratio. We have presented a full discussion of statistics reflecting the strength of association in another article.[11] Many biologically plausible surrogates are only weakly associated with clinically important outcomes. For example, measures of respiratory function in patients with chronic lung disease, or conventional exercise tests in patients with heart and lung disease, are only weakly correlated with capacity to undertake activities of daily living.[12,13] When correlations are low, the surrogate is likely to be a poor substitute for the target outcome.

In addition to the strength of the association, one's confidence in the validity of the association depends on whether it is consistent across different studies and after adjustment for known confounders. For example, ecologic studies such as the Seven Countries Study[14] suggested a strong correlation between serum cholesterol levels and coronary heart disease mortality even after adjusting for other predictors such as age, smoking, and systolic blood pressure. Subsequent cohort studies confirmed this association and suggested that long-term reductions in serum cholesterol levels of 0.6 mmol/L (23 mg/dL) would lower the risk of coronary heart disease by approximately 30%. When a surrogate is associated with an outcome after ad-

justing for multiple other potential prognostic factors we call the association *independent*.

Similarly, cohort studies have consistently revealed that a single measurement of plasma viral load predicts the subsequent risk of AIDS or death in patients infected with HIV.[15-20] For example, in 1 study the proportion of patients that progressed to AIDS after 5 years in the lowest through the highest quartiles of viral load was 8%, 26%, 49%, and 62%, respectively.[20] Moreover, this association retained its predictive power after adjustment for other potential predictors such as CD4 cell count.[15-19]

Returning to the scenario, you are wondering if you can substitute bone mineral density for fractures or health-related quality of life in considering whether to recommend raloxifene. A large cohort study investigated risk factors for hip fracture.[21] Postmenopausal women with a calcaneal bone density in the highest third had a hip fracture rate of 9.4/1000 woman-years while women in the middle and lowest third had a fracture rate per 1000 woman-years of 14.7 and 27.3, respectively. Furthermore, after considering other risk factors for osteoporotic hip fractures including maternal history of hip fracture, previous fractures from any site, poor self-rated health, use of long-acting benzodiazepines, impaired visual function, and reduced physical activity, bone mineral density continued to predict the risk of hip fracture.[21] These findings are consistent across studies looking at the association between bone density and fracture risk.[22,23] Thus, bone mineral density is a moderately strong, independent predictor of fracture, and meets our first criterion for an acceptable surrogate end point.

While meeting this first criterion is necessary, it is not sufficient to support reliance on a surrogate outcome. As we will emphasize below (Table 1), before offering an intervention on the basis of effects on a surrogate outcome, the clinician should note a consistent relationship between surrogate and target in randomized trials; the effect of the intervention on the surrogate must be large, precise, and lasting, and the benefit-risk trade-off must be clear.

### Is There Evidence From Randomized Trials in Other Drug Classes That Improvement in the Surrogate End Point Has Consistently Led to Improvement in the Target Outcome?

Given the possibility of effects unrelated to the surrogate end point, pathophysiologic studies, ecological studies, and cohort studies are insufficient to establish that the link between surrogate and clinically important outcomes is ironclad. We can confidently rely on surrogate end points only when long-term randomized trials have consistently demonstrated that modification of the surrogate is associated with concomitant modifications in the target outcome of interest. For example, although ventricular ectopic beats are associated with adverse prognosis in patients with myocardial infarction[24] and class 1 antiarrhythmic agents effectively suppress ventricular arrhythmias in animals and humans,[25] these drugs have proved to increase mortality when evaluated in randomized trials.[26] In this case, reliance on the surrogate end point of suppression of nonlethal arrhythmias led to the deaths of tens of thousands of patients.[27]

The treatment of heart failure provides another instructive example. Trials of angiotensin-converting enzyme inhibitors in heart failure treatment have demonstrated parallel increases in exercise capacity[28-31] and decreases in mortality,[32] suggesting that clinicians may be able to rely on exercise capacity as a valid surrogate. Milrinone[33] and epoprostenol[34] have both demonstrated improved exercise tolerance in patients with symptomatic heart failure. However, when these drugs were evaluated in randomized controlled trials both showed an increase in cardiovascular mortality that in one instance was statistically significant,[35] and in the second case led to the early termination of the study.[36] Thus, exercise tolerance is inconsistent in predicting improved mortality and is therefore an unsatisfactory substitute. Other suggested surrogate end points in heart failure have included ejection fraction, heart-rate variability, and markers of autonomic function.[37] The dopaminergic agent ibopamine positively influences all 3 surrogate end points, and yet a randomized trial demonstrated that the drug increases mortality in heart failure.[38]

An example of a surrogate end point is CD4 cell count, which has been validated in randomized trials. A number of trials comparing different classes of antiretroviral therapies have demonstrated that patients randomized to more potent drug regimens had higher CD4 cell counts and were less likely to progress to AIDS or death.[6,39] While there is no guarantee that the next trial using a different class of drugs will show the same pattern, these results greatly strengthen our inference that if therapy for HIV infection increases the CD4 count, a reduction in AIDS-related mortality will result.

Returning to our scenario, trials of etidronate[40,41] and alendronate[42] for the prevention of osteoporotic fractures in postmenopausal women have shown parallel increases in bone mineral density and reduced incidences of new vertebral fractures. This would suggest that clinicians might rely on bone density to evaluate new drugs in osteoporosis in making the assumption that if they saw increases in bone density, decreases in fractures would follow.

However, another secondary prevention trial in postmenopausal women using sodium fluoride showed divergent results.[43] Although sodium fluoride increased bone mineral density at the lumbar spine by 35% over 5 years, more vertebral and nonvertebral fractures occurred in the intervention group than in the placebo group (163 and 72 in 101 women with sodium fluoride vs 136 and 24 in 101 women with placebo). In another randomized trial, fluoride again showed a large increase in bone density without any change in fracture rate.[44] Inferences on the basis of unchanged bone density may also be problematic. A study of calcium and vitamin D in the elderly showed virtually no change in bone density, but a reduction in fracture risk of approximately 50%.[45] Thus, increase in

bone mineral density as a surrogate end point has shown an inconsistent relationship to osteoporotic fractures.

### Is There Evidence From Randomized Trials in the Same Drug Class That Improvement in the Surrogate End Point Has Consistently Led to Improvement in the Target Outcome?

Clinicians are in a stronger position to rely on surrogate end points if the new drug they are considering is from a class of drugs in which the relationship between changes in the surrogate and changes in the target has been verified in randomized trials. For instance, thiazide diuretics and β-blockers have both been shown to reduce blood pressure and clinically important outcomes such as stroke in patients with hypertension. Thus, we would be much more comfortable relying on reduction in blood pressure to justify administering a new β-blocker or thiazide diuretic than to justify offering a novel antihypertensive agent from another class.[46]

For example, although 1 dihydropyridine calcium channel blocker has been shown to reduce clinically important outcomes in patients with hypertension,[47] 4 other trials have shown that these agents are less efficacious than thiazides or angiotensin-converting enzyme inhibitors in preventing hard clinical end points despite exerting similar degrees of blood pressure lowering.[48-51]

We will consider the example of cholesterol reduction as a surrogate for cardiovascular outcomes such as myocardial infarction and death in part B of this Users' Guide.[52] Briefly, several large trials of primary and secondary prevention of coronary heart disease with statins have consistently shown that these drugs reduce cardiovascular outcomes.[53]

We could therefore make the assumption that a new statin with a similar low-density lipoprotein cholesterol–lowering potency may also reduce clinically important outcomes. However, we would be reluctant to generalize to another class of lipid-lowering agents since trials of 1 such class (the fibrates) have shown that these drugs reduce the incidence of myocardial infarction but increase the risk of mortality from other causes (with no impact on overall mortality).[53-55]

These examples highlight the point we made earlier: confidence in a surrogate outcome depends on the assumption that the treatment captures any relationship between the treatment and the outcome.[56,57] This assumption can be violated in 2 ways. First, treatment may have a beneficial mechanism of effect on the outcome independent of its effect on the surrogate. For instance, 1 explanation for the superior effect of angiotensin-converting enzyme inhibitors vs calcium antagonists on clinically important outcomes is that angiotensin-converting enzyme inhibition has biological effects independent of lowering blood pressure that reduce risk of stroke or death and that calcium antagonists do not share these effects.

Second, treatment may have deleterious effects on the outcome that are not mediated through the surrogate. Mortality-increasing effects of fibrates rather than inability to lower morbidity and mortality through cholesterol reduction probably explain the lack of effect of fibrates on clinically important outcomes. That such additional effects are less likely across classes of drugs than within classes is what makes us more inclined to rely on within-class evidence from surrogate outcomes.

This criterion is complicated by the variable definitions of drug class. A manufacturer of a drug related to a class of agents with a consistently positive association between modification of a surrogate end point and modification of the target (such as a β-blocker) will naturally argue for a broad definition of class. Manufacturers of agents that are related to drugs with known or suspected adverse effects on target events (clofibrate, or some calcium antagonists) are likely to argue, on the other hand, that the chemical or physiological connection is not sufficiently close to consider the new drug to be in the same class as the harmful agent. Part B will address these issues more fully.[52]

Returning to the scenario, we have established that because of the inconsistent relationship between increase in bone mineral density and fracture reduction we would be reluctant to offer patients a new antiosteoporotic agent solely on the basis of evidence of its effect on the surrogate end point. Raloxifene, the drug we are considering for our patient, is a nonsteroidal benzothiophene, a selective estrogen-receptor modulator representing a new class of drugs for the prevention of osteoporosis-related bone fractures. Thus, it is likely that the mechanisms of action will be considerably different from bisphosphonates and the conclusion that similar reductions in loss of bone density will lead to parallel reductions in clinical fractures is questionable. In TABLE 2, we apply our validity criteria to a number of controversial examples of the use of surrogate end points.

### What Were the Results? How Large, Precise, and Lasting Was the Treatment Effect?

We are interested not only in whether an intervention alters a surrogate end point, but also in the magnitude, precision, and duration of the effect. If an intervention shows large reductions in the surrogate end point, the 95% confidence intervals (CIs) around those large reductions are narrow, and the effect persists over a sufficiently long period, our confidence that the target outcome will be favorably affected increases. Positive effects that are smaller, with wider CIs, and shorter duration of follow-up leave us less confident.

We have already cited evidence suggesting that CD4 cell counts may be an acceptable surrogate for mortality in patients with HIV infection. A randomized controlled trial of immediate vs delayed zidovudine therapy in HIV-infected asymptomatic individuals declared a positive result for immediate therapy, largely on the basis of a greater proportion of treated patients with CD4 cell counts above 435 × $10^6$/L at a median follow-up of 1.7 years.[38] Subsequently, the Concorde study addressed the same question in a randomized trial

with a median follow-up of 3.3 years.[59] The Concorde investigators found a continuous decline in CD4 cell counts in both treated and control groups, but the median difference of 30 × $10^6$/L in favor of treated patients at study termination was statistically significant. However, the study showed no effect of zidovudine in terms of reduced progression to AIDS or death. The median CD4 cell count difference was insufficient to have an impact on clinically important outcomes. The Concorde authors made the following conclusion: the small, but highly significant persistent difference in CD4 cell counts between the groups was not translated into a significant clinical benefit and "called into question the uncritical use of CD4 cell counts as a surrogate endpoint." Had the Concorde analysis showed significantly shorter times to reach a CD4 cell count of 350 × $10^6$/L in the control group and been regarded as fundamental, the trial might have been stopped early with a false-positive result.

Returning to our scenario, the trial of raloxifene in women with osteoporosis demonstrated that after 2 years of treatment, raloxifene-treated patients in the group receiving the highest dosage showed an increase in bone mineral density at the lumbar spine of 2.2% (SE, 0.3%) compared with a slight decrease in the control group 0.8% (SE, 0.3%). This difference in change over time was statistically significant ($P<.03$). Ideally, the investigators would have provided us with a CI around the 3% difference in percentage change in bone mineral density in the treatment and control groups. As we will illustrate when we consider weighing benefits and harms, the magnitude of the effect on the surrogate may (or may not) help us estimate the size of a possible affect on the target outcome.

### Will the Results Help in Caring for My Patients?

The questions clinicians should ask themselves in applying the results are the same ones we have suggested for any issue of therapy or prevention[60] and elaborated on in our Users' Guide regarding applicability.[61] These 3 questions have to do with whether the results can be applied to your patient's care, whether all important outcomes were considered, and whether the likely benefits are worth the down sides of treatment.

"Can the results be applied to my patient's care" refers to the extent to which the patient before you is similar to those who participated in the published studies under consideration, and the extent to which the therapy, and the associated technologies for monitoring and responding to complications, are available in your setting. "Were all important outcomes considered" relates to the focus of this Users' Guide, and all the issues we have raised thus far: was the primary outcome really the one in which patients will be interested?

This second criterion also draws issues of adverse intervention effects to our attention. Applying the third cri-

**Table 2.** Selected Examples of Applied Validity Criteria for the Critical Evaluation of Studies Using Surrogate End Points

| Types of Intervention | Criterion: Is There a Strong, Independent, Consistent Association Between the Surrogate End Point and the Clinical End Point? | Criterion: Is There Evidence From Randomized Trials in Other Drug Classes That Improvement in the Surrogate End Point Has Consistently Led to Improvement in the Target Outcome? | Criterion: Is There Evidence From Randomized Trials in the Same Drug Class That Improvement in the Surrogate End Point Has Consistently Led to Improvement in the Target Outcome? | Surrogate End Point | End Point |
|---|---|---|---|---|---|
| Nonsteroidal benzothiophene<br>Raloxifene[1] | Yes[21-23] | No[62,64] | No[1,62] | Bone mineral density | Osteoporotic fractures |
| Protease inhibitor*<br>Nelfinavir[63] | Yes[16-19] | Yes[64-66] | Yes[6,39] | Human immunodeficiency virus plasma load | Acquired immunodeficiency syndrome or death |
| Reverse transcriptase inhibitor<br>Abacavir[67] | Yes[16-19] | Yes[65,66,68] | Yes[64,66] | Human immunodeficiency virus viral plasma load | Acquired immunodeficiency syndrome or death |
| Protease inhibitor*<br>Nelfinavir[63] | Yes[15-19] | Yes[64-66] | Yes[6,39] | CD4 cell count | Acquired immunodeficiency syndrome or death |
| Reverse transcriptase inhibitor<br>Abacavir[67] | Yes[15-19] | Yes[65,66,68] | Yes[64,66] | CD4 cell count | Acquired immunodeficiency syndrome or death |
| Antihypertensive drugs<br>Dihydropyridine calcium antagonist<br>New thiazide diuretic | Yes[69,70]<br>Yes[69,70] | Yes[71]<br>Yes[71] | No[48,51]<br>Yes[71] | Blood pressure reduction | Stroke, myocardial infarction, cardiovascular mortality |
| Antilipidemic drugs<br>Atorvastatin[72,73]<br>Bezafibrate[75,76] | Yes[74,76]<br>Yes[74,76] | No[53]<br>No[53] | Yes[53]<br>No[53] | Cholesterol or low-density lipoprotein cholesterol | Myocardial infarction, death from myocardial infarction |

*In combination therapy with 2 reverse transcriptase inhibitors.

terion, judging whether the benefits are worth the down sides of treatment, presents particular challenges when investigators have focused on surrogate end points, and we will discuss this criterion in some detail.

### Are the Likely Treatment Benefits Worth the Potential Harms and Costs?

To know whether to offer a treatment to their patients, clinicians must be able to estimate the magnitude of the likely benefit. When the data available are limited to the effect on a surrogate end point, estimating the extent to which treatment will reduce clinically important outcomes becomes a challenge.

One approach is to extrapolate from 1 or more randomized trials assessing a related intervention in a similar patient population that provides both surrogate end point and clinical outcome data. For example, until recently there were little long-term data on the efficacy of lovastatin in reducing clinically important outcomes. However, one could extrapolate from short-term dose efficacy studies assessing the surrogate end point of cholesterol lowering. Thus, since 40 mg of lovastatin produced a similar degree of lowering of low-density lipoprotein cholesterol as 40 mg of pravastatin (31% vs 34% reduction) in the CURVES Study,[77] one could theorize that lovastatin would have similar long-term benefits to pravastatin. Subsequently, the AFCAPS/TexCAPS Trial (a 5-year trial assessing the efficacy of lovastatin in the primary prevention of ischemic heart disease)[78] confirmed that this agent had a beneficial profile similar to pravastatin (as determined by the 5-year, primary prevention WOSCOPS Trial)[79]: the RR reductions (and 95% CIs) for myocardial infarction were 40% (17%-57%) and 31% (17%-43%), respectively. However, this approach is likely to be seriously flawed when one is extrapolating from trials of another class of drugs.

Returning to our scenario, to estimate the magnitude of the fracture reduction we might expect with raloxifene (in which we have only surrogate end point data), we could (recognizing the limitations of this approach pointed out above) examine the results of randomized controlled trials of alendronate (a drug from a different class for which we have data on the same surrogate end point as well as clinical end points such as fracture reduction). While alendronate appears to improve vertebral bone density by 7.5% over 2 years (vs control),[42] raloxifene is associated with only a 3.0% improvement over the same time frame. A systematic overview of the alendronate trials[80] reported a 29% reduction in RR of nonvertebral fracture over 2 years. Only 1 trial looked at symptomatic vertebral fractures in women with decreased bone density and an existing vertebral fracture.[81] This study demonstrated an RR reduction of 55% with alendronate and suggested that our patient's risk over 3 years of a nonvertebral fracture would be approximately 15%; symptomatic vertebral fracture would be about 5%. Given the RR reductions with alendronate, one would need to treat approximately 25 women to prevent a nonvertebral fracture and 40 women to prevent a symptomatic vertebral fracture over a 3-year period.

Since the improvement in bone mineral density with raloxifene is at best 50% of the effect of alendronate, we would anticipate a considerably lower reduction in fracture risk with raloxifene. However, interim analysis of an ongoing raloxifene trial[62] reported a 46% RR reduction with this therapy (despite less of an increase in bone mineral density than seen with the alendronate trials). This serves to emphasize the dangers of extrapolating results across classes when it is uncertain that the effects on clinically important outcomes are mediated in the same fashion by the 2 comparison drugs.

In deciding whether the likely magnitude of the treatment effect warrants offering patients the intervention, clinicians must consider not only the uncertainty associated with that estimate, but the trade-off with potential toxic effects and costs of therapy. In addition, clinicians must ponder the consequences of not treating, and the available management alternatives. The deadly and usually relentless progression of HIV infection, and the paucity of alternative therapies, has contributed to the readiness of patients, clinicians, and regulatory agencies to accept evidence from surrogate end points in instituting novel therapies in patients infected with HIV. In osteoporosis, in which the consequences of the condition are less immediately devastating, and a variety of agents are available, the case for relying on surrogate end points is far less compelling.

## RESOLUTION OF THE SCENARIO

We have found a strong, consistent, independent, and biologically plausible association between bone mineral density and vertebral and nonvertebral fractures. Randomized trials, however, have failed to show a consistent association between increased bone density and reduction in fracture across all drug classes.

Because our patient is at substantial risk of fracture over the short term, the number needed to treat to prevent both nonvertebral and vertebral fractures is moderate, as is the absolute benefit she might expect. Moreover, she is interested in longer-term fracture prevention, and her risk will grow over time. One might offer her alternative interventions, including hormone replacement therapy, calcium and vitamin D, bisphosphonates, or calcitonin.

While there is strong evidence from randomized trials supporting the use of bisphosphonates to decrease osteoporotic fractures, randomized trial data showing fracture reduction in populations similar to our patient with the other agents is limited. Our patient is concerned about her long-term risk. Raloxifene was well tolerated during this 2-year trial but no information is available about long-term adverse effects including cardiovascular disease, venous thromboembolism, breast and endometrial cancer, and menopausal symptoms. While a number of options (including a trial of etidronate, offering hormone replacement therapy, calcium and vitamin D, calcitonin, or suggesting only a bal-

anced diet and exercise) might be reasonable, ideally the clinician would subject these options to the same scrutiny applied to raloxifene.

Data indicating a reduction in fracture rate would greatly strengthen the case for including raloxifene as the preferred option. Just as you are about to see the patient (and, for us, just before this article went to press) you pick up a few of your latest editions of *JAMA* from the pile in the corner of your office, and find 2 highly relevant randomized trials.[82,83] The results show that, in women like your patient with a prevalent vertebral fracture, raloxifene decreased radiological vertebral fracture risk (for 60 mg: number needed to treat = 16 [RR, 0.7; 95% CI, 0.6-0.9]; and for 120 mg: number needed to treat = 10 [RR, 0.5; 95% CI, 0.4-0.7]), but did not decrease the incidence of nonvertebral fracture. In helping your patient to decide on the right course of action, you realize you will have to consider other effects of raloxifene: the *JAMA* articles also show a 76% RR reduction of breast cancer as detected by mammography (number needed to treat, 126), a 3-fold increase in the risk of venous thromboembolism, and an increased incidence of hot flashes, leg cramps, influenzalike syndromes, and peripheral edema.

When we use surrogate end points to make inferences about expected benefit, we are making assumptions regarding the link between the surrogate end point and the target outcome. We have outlined criteria clinicians can use to decide when these assumptions might be appropriate. Even if a surrogate end point meets all of these criteria, inferences about a treatment benefit may still prove misleading. Thus, treatment recommendations based on surrogate outcome effects can never be strong. Furthermore, difficulties in estimating the magnitude of effects on clinically important end points compromises economic analysis examining the cost-effectiveness of alternative management strategies.

These considerations emphasize that waiting for randomized trials investigating the effect of the intervention on outcomes of unequivocal importance to patients is the only ironclad solution to the surrogate outcome dilemma. When clinicians must choose between alternative interventions, trials should make head-to-head comparisons between competing treatments rather than restricting comparisons of treatment to control or placebo. We expand on this issue in Part B of this Users' Guide. However, when patients' risk of serious morbidity or mortality are high, this "wait-and-see" strategy may pose problems for many patients and their physicians.

We encourage clinicians to critically question therapeutic interventions in which the only proof of efficacy is from surrogate end point data. When the surrogate end point meets all our validity criteria, the effect of the intervention on the surrogate end point is large, the patient's risk of the target outcome is high, the patient places a high value on avoiding the target outcome, and there are no satisfactory alternative therapies, clinicians can recommend therapy on the basis of randomized trials evaluating only surrogate end points. In other situations, clinicians must carefully consider the known adverse effects and cost of therapy, and the possibility of unanticipated adverse effects, before recommending an intervention solely on the basis of surrogate end point data.

**Acknowledgment:** We are grateful to Cliff Rosen, MD, for his helpful comments concerning the scenario and the associated discussion. Deborah Maddock provided invaluable coordination for the EBM Working Group in the development of this article.

## REFERENCES

1. Delmas PD, Bjarnason NH, Mitlak BH, et al. Effects of raloxifene on bone mineral density, serum cholesterol concentrations, and uterine endometrium in postmenopausal women. *N Engl J Med.* 1997;337:1641-1647.
2. Temple RJ. A regulatory authority's opinion about surrogate endpoints. In: Nimmo WS, Tucker GT, eds. *Clinical Measurement in Drug Evaluation.* New York, NY: John Wiley & Sons Inc; 1995:57.
3. Hammer SM, Katzenstein DA, Hughes MD, et al. A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *N Engl J Med.* 1996;335:1081-1090.
4. Delta Coordinating Committee. Delta: a randomised double-blind controlled trial comparing combinations of zidovudine plus didanosine or zalcitabine with zidovudine alone in HIV-infected individuals. *Lancet.* 1996;348:283-291.
5. Saravolatz LD, Winslow DL, Collins G, et al. Zidovudine alone or in combination with didanosine or zalcitabine in HIV-infected patients with the acquired immunodeficiency syndrome or fewer than 200 CD4 cells per cubic millimeter. *N Engl J Med.* 1996; 335:1099-1106.
6. Hammer SM, Squires KE, Hughes MD, et al. A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200 per cubic millimeter or less. *N Engl J Med.* 1997;337:725-733.
7. Niebauer J, Coats AJ. Treating chronic heart failure: time to take stock. *Lancet.* 1997;349:966-967.
8. Massie BM, Berk MR, Brozena SC, et al. Can further benefit be achieved by adding flosequinan to patients with congestive heart failure who remain symptomatic on diuretic, digoxin, and an angiotensin converting enzyme inhibitor? results of the flosequinan-ACE inhibitor trial (FACET). *Circulation.* 1993;88:492-501.
9. How to read clinical journals, IV: to determine etiology or causation. *CMAJ.* 1981;124:985-990.
10. Oxman AD, Sackett DL, Guyatt GH. Users' guides to the medical literature, I: how to get started. *JAMA.* 1993;270:2093-2095.
11. Guyatt G, Walter S, Shannon H, Cook D, Jaeschke R, Heddle N. Basic statistics for clinicians, IV: correlation and regression. *CMAJ.* 1995;152:497-504.
12. Guyatt GH, Thompson PJ, Berman LB, et al. How should we measure function in patients with chronic heart and lung disease? *J Chronic Dis.* 1985;38:517-524.
13. Mahler DA, Weinberg DH, Wells CK, Feinstein AR. The measurement of dyspnea: contents, interobserver agreement, and physiologic correlates of two new clinical indexes. *Chest.* 1984;85:751-758.
14. Verschuren WM, Jacobs DR, Bloemberg BP, et al. Serum total cholesterol and long-term coronary heart disease mortality in different cultures. *JAMA.* 1995; 274:131-136.
15. Mellors JW, Rinaldo CR Jr, Gupta P, et al. Prognosis in HIV-1 infection predicted by the quantity of virus in plasma. *Science.* 1996;272:1167-1170.
16. Mellors JW, Kingsley LA, Rinaldo CR, et al. Quantitation of HIV-1 RNA in plasma predicts outcome after seroconversion. *Ann Intern Med.* 1995;122:573-579.
17. Ruiz L, Romeu J, Clotet B, et al. Quantitative HIV-1 RNA as a marker of clinical stability and survival in a cohort of 302 patients with a mean CD4 cell count of 300 × 10(6)/1. *AIDS.* 1996;10:F39-F44.
18. O'Brien TR, Blattner WA, Waters D, et al. Serum HIV-1 RNA levels and time to development of AIDS in the Multicenter Hemophilia Cohort Study. *JAMA.* 1996;276:105-110.
19. Yerly S, Perneger TV, Hirschel B, et al. A critical assessment of the prognostic value of HIV-1 RNA levels and CD4+ cell counts in HIV-infected patients. *Arch Intern Med.* 1998;158:247-252.
20. Ho DD. Viral counts in HIV infection. *Science.* 1996;272:1124-1125.
21. Cummings SR, Nevitt MC, Browner WS, et al. Risk factors for hip fracture in white women. *N Engl J Med.* 1995;332:767-773.
22. Marshall D, Johnell O, Wedel H. Meta-analysis of how well measures of bone mineral density predict occurrence of osteoporotic fractures. *BMJ.* 1996; 312:1254-1259.
23. Huang C, Ross PD, Wasnich RD. Short-term and long-term fracture prediction by bone mass measurements. *J Bone Miner Res.* 1998;13:107-113.
24. Bigger JT Jr, Fleiss JL, Kleiger R, et al. The relationships among ventricular arrhythmias, left ventricular dysfunction, and mortality in the 2 years after myocardial infarction. *Circulation.* 1984;69:250-258.
25. McAlister FA, Teo KK. Antiarrhythmic therapies for the prevention of sudden cardiac death. *Drugs.* 1997;54:235-252.
26. Echt DS, Liebson PR, Mitchell LB, et al, and the Cardiac Arrhythmia Suppression Trial. Mortality and

morbidity in patients receiving encainide, flecainide, or placebo. *N Engl J Med.* 1991;324:781-788.
**27.** Moore TJ. *Deadly Medicine.* New York, NY: Simon & Schuster; 1995.
**28.** Drexler H, Banhardt U, Meinertz T, et al. Contrasting peripheral short-term and long-term effects of converting enzyme inhibition in patients with congestive heart failure: a double-blind, placebo-controlled trial. *Circulation.* 1989;79:491-502.
**29.** Lewis GR. Comparison of lisinopril versus placebo for congestive heart failure. *Am J Cardiol.* 1989; 63:12D-16D.
**30.** Giles TD, Fisher MB, Rush JE. Lisinopril and captopril in the treatment of heart failure in older patients. *Am J Med.* 1988;85:44-47.
**31.** Riegger GA. Effects of quinapril on exercise tolerance in patients with mild to moderate heart failure. *Eur Heart J.* 1991;12:705-711.
**32.** Garg R, Yusuf S. Overview of randomized trials of angiotensin-converting enzyme inhibitors on mortality and morbidity in patients with heart failure. *JAMA.* 1995;273:1450-1456.
**33.** Di Bianco R, Shabetai R, Kostuk W, et al. A comparison of oral milrinone, digoxin, and their combination in the treatment of patients with chronic heart failure. *N Engl J Med.* 1989;320:677-683.
**34.** Sueta CA, Gheorghiade M, Adams KF, et al, and the Epoprostenol Multicenter Research Group. Safety and efficacy of epoprostenol in patients with severe congestive heart failure. *Am J Cardiol.* 1995;75:34A-43A.
**35.** Packer M, Carver JR, Rodeheffer RJ, et al, for the Promise Study Research Group. Effect of oral milrinone on mortality in severe chronic heart failure. *N Engl J Med.* 1991;325:1468-1475.
**36.** Califf RM, Adams KF, McKenna WJ, et al. A randomized controlled trial of epoprostenol therapy for severe congestive heart failure. *Am Heart J.* 1997; 134:44-54.
**37.** Yee KM, Struthers AD. Can drug effects on mortality in heart failure be predicted by any surrogate measure? *Eur Heart J.* 1997;18:1860-1864.
**38.** Hampton JR, van Veldhuisen DJ, Kleber FX, et al. Randomised study of effect of ibopamine on survival in patients with advanced severe heart failure. *Lancet.* 1997;349:971-977.
**39.** Cameron DW, Heath-Chiozzi M, Danner S, et al, and the Advanced HIV Disease Ritonavir Study Group. Randomised placebo-controlled trial of ritonavir in advanced HIV-1 disease. *Lancet.* 1997;351:543-549.
**40.** Watts NB, Harris ST, Genant HK, et al. Intermittent cyclical etidronate treatment of postmenopausal osteoporosis. *N Engl J Med.* 1990;323:73-79.
**41.** Storm T, Thamsborg G, Steiniche T, Genant HK, Sorensen OH. Effect of intermittent cyclical etidronate therapy on bone mass and fracture rate in women with postmenopausal osteoporosis. *N Engl J Med.* 1990;322:1265-1271.
**42.** Liberman UA, Weiss SR, Broll J, et al, and the Alendronate Phase III Osteoporosis Treatment Study Group. Effect of oral alendronate on bone mineral density and the incidence of fractures in postmenopausal osteoporosis. *N Engl J Med.* 1995;333:1437-1443.
**43.** Riggs BL, Hodgson SF, O'Fallon WM, et al. Effect of fluoride treatment on the fracture rate in postmenopausal women with osteoporosis. *N Engl J Med.* 1990;322:802-809.
**44.** Meunier PJ, Sebert J-L, Reginster J-Y, et al. Fluoride salts are no better at preventing new vertebral fractures than calcium-vitamin D in postmenopausal osteoporosis. *Osteoporos Int.* 1998;8:4-12.
**45.** Dawson-Hughes B, Harris SS, Krall EA, Dallal GE. Effect of calcium and vitamin D supplementation on bone density in men and women 65 years of age and older. *N Engl J Med.* 1977;337:670-676.
**46.** McAlister FA, Straus S, Sackett DL. Randomized clinical trials of antihypertensive drugs: all that glitters is not gold. *CMAJ.* 1998;159:488-490.
**47.** Staessen JA, Fagard R, Thijs L, et al. Randomised double-blind comparison of placebo and active treatment for older patients with isolated systolic hypertension. *Lancet.* 1997;350:757-764.
**48.** Psaty BM, Siscovick DS, Weiss NS, et al. Hypertension and outcomes research: from clinical trials to clinical epidemiology. *Am J Hypertens.* 1996;9:178-183.
**49.** Borhani NO, Mercuri M, Borhani PA, et al. Final outcome results of the multicenter isradipine diuretic atherosclerosis study (MIDAS): a randomized controlled trial. *JAMA.* 1996;276:785-791.
**50.** Tatti P, Pahor M, Byington RP, et al. Outcome results of the fosinopril versus amlodipine cardiovascular events randomized trial (FACET) in patients with hypertension and NIDDM. *Diabetes Care.* 1998;21:597-603.
**51.** Estacio RO, Jeffers BW, Hiatt WR, Biggerstaff SL, Gifford N, Schrier RW. The effect of nisoldipine as compared with enalapril on cardiovascular outcomes in patients with non-insulin-dependent diabetes and hypertension. *N Engl J Med.* 1998;338:645-652.
**52.** McAlister FA, Laupacis A, Wells GA, Sackett DL, for the Evidence-Based Medicine Working Group. Users' guide to the medical literature, XIX: applying clinical trial results part B: guidelines for determining whether a drug is exerting (more than) a class effect. *JAMA.* In press.
**53.** Bucher HC, Griffith LE, Guyatt GH. Systematic review on risk and benefit of different cholesterol lowering interventions. *Arterioscler Thromb Vasc Biol.* 1999;19:187-195.
**54.** Muldoon MF, Manuck SB, Matthews KA. Lowering cholesterol concentration and mortality. *BMJ.* 1990;301:309-314.
**55.** Smith GD, Song F, Sheldon TA. Cholesterol lowering and mortality. *BMJ.* 1993;306:1367-1373.
**56.** Prentice RL. Surrogate endpoints in clinical trials. *Stat Med.* 1989;8:431-440.
**57.** Fleming TR. Surrogate markers in AIDS and cancer trials. *Stat Med.* 1994;13:1423-1435.
**58.** Cooper DA, Gatell JM, Kroon S, et al, and the European-Australian Collaborative Group. Zidovudine in persons with asymptomatic HIV infection and CD4+ cell counts greater than 400 per cubic millimeter. *N Engl J Med.* 1993;329:297-303.
**59.** Concorde Coordinating Committee. Concorde: MRC/ANRS randomised double-blind controlled trial of immediate and deferred zidovudine in symptom-free HIV infection. *Lancet.* 1994;343:871-881.
**60.** Guyatt GH, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, II: how to use an article about therapy or prevention A: are the results of the study valid? *JAMA.* 1993;270:2598-2601.
**61.** Dans AL, Dans LF, Guyatt GH, Richardson S, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, XIV: how to decide on the applicability of clinical trial results to your patient. *JAMA.* 1998;279:545-549.
**62.** Ettinger B, Black D, Cummings S, et al. Raloxifene reduces the risk of incident vertebral fractures: 24 month interim analyses. *Osteoporos Int.* 1998;8 (suppl 3):11.
**63.** Clendeneninn N, Quart B, Anderson R, Knowles M, Chang Y. Analysis of long-term virologic data from the viracept (nelfinavir) 511 protocol using 3 HIV-RNA assays. In: Abstracts from the 5th Conference on Retroviruses and Opportunistic Infections; Chicago, Ill; 1998.
**64.** Brun-Vezinet F, Boucher C, Loveday C, et al, and the Delta Virology Working Group and Coordinating Committee. HIV-1 viral load, phenotype, and resistance in a subset of drug-naive participants from the Delta trial: the national virology groups. *Lancet.* 1997; 350:983-990.
**65.** Montaner JS, Reiss P, Cooper D, et al. A randomized, double-blind trial comparing combinations of nevirapine, didanosine, and zidovudine for HIV-infected patients: the Netherlands, Canada, and Australia (INCAS) study. *JAMA.* 1998;279:930-937.
**66.** CEASAR Coordinating Committee. Randomised trial of addition of lamivudine or lamivudine plus loviride to zidovudine-containing regimens for patients with HIV-1 infection: the CAESAR trial. *Lancet.* 1997;349:1413-1421.
**67.** Fischl M, Greenberg S, Clumeck N, et al. Safety and activity of abacavir (1592, ABC) with 3TC/ZDV in antiretroviral naive subjects. In: Abstracts from the 12th World AIDS Conference: Geneva, Switzerland; June 28-July 3, 1998.
**68.** Katzenstein DA, Hammer SM, Hughes MD, et al, and the AIDS Clinical Trials Group Study 175 Virology Study Team. The relation of virologic and immunologic markers to clinical outcomes after nucleoside therapy in HIV-infected adults with 200 to 500 CD4 cells per cubic millimeter. *N Engl J Med.* 1996;335:1091-1098.
**69.** Collins R, Peto R, MacMahon S, et al. Blood pressure, stroke, and coronary heart disease, part 2: short-term reductions in blood pressure; overview of randomised drug trials in their epidemiological context. *Lancet.* 1990;335:827-838.
**70.** MacMahon S, Peto R, Cutler J, et al. Blood pressure, stroke, and coronary heart disease, part 1: prolonged differences in blood pressure; prospective observational studies corrected for the regression dilution bias. *Lancet.* 1990;335:765-774.
**71.** Psaty BM, Smith NL, Siscovick DS, et al. Health outcomes associated with antihypertensive therapies used as first-line agents: a systematic review and meta-analysis. *JAMA.* 1997;277:739-745.
**72.** Heinonen TM, Stein E, Weiss SR, et al. The lipid-lowering effects of atorvastatin, a new HMG-coA reductase inhibitor: results of a randomized, double-masked study. *Clin Ther.* 1996;18:853-863.
**73.** Bakker-Arkema RG, Davidson MH, Goldstein RJ, et al. Efficacy and safety of a new HMG-coA reductase inhibitor, atorvastatin, in patients with hypertriglyceridemia. *JAMA.* 1996;275:128-133.
**74.** Law MR, Wald NJ, Thompson SG. By how much and how quickly does reduction in serum cholesterol concentration lower risk of ischaemic heart disease? *BMJ.* 1994;308:367-372.
**75.** Winocour PH, Durrington PN, Bhatagnar D, et al. The effect of bezafibrate on very low density lipoprotein (VLDL), intermediate density lipoprotein (IDL), and low density lipoprotein (LDL) composition in type 1 diabetes associated with hypercholesterolaemia or combined hyperlipidaemia. *Atherosclerosis.* 1992;93:83-94.
**76.** Jones IR, Swai A, Taylor R, Miller M, Laker MF, Alberti KG. Lowering of plasma glucose concentrations with bezafibrate in patients with moderately controlled NIDDM. *Diabetes Care.* 1990;13:855-863.
**77.** Jones P, Kafonek S, Laurora I, Hunninghake D. Comparative dose efficacy study of atorvastatin versus simvastatin, pravastatin, lovastatin, and fluvastatin in patients with hypercholesterolemia (the CURVES study). *Am J Cardiol.* 1998;81:582-587.
**78.** Downs JR, Clearfield M, Weis S, et al. Primary prevention of acute coronary events with lovastatin in men and women with average cholesterol levels. *JAMA.* 1998;279:1615-1622.
**79.** Shepherd J, Cobbe SM, Ford I, et al. Prevention of coronary heart disease with pravastatin in men with hypercholesterolemia. *N Engl J Med.* 1995;333:1301-1307.
**80.** Karpf DB, Shapiro DR, Seeman E, et al. Prevention of nonvertebral fractures by alendronate: a meta-analysis. *JAMA.* 1997;277:1159-1164.
**81.** Black DM, Cummings SR, Karpf DB, et al, and the Fracture Intervention Trial Research Group. Randomised trial of effect of alendronate on risk of fracture in women with existing vertebral fractures. *Lancet.* 1996;348:1535-1541.
**82.** Cummings SR, Eckert S, Krueger KA, et al. The effect of raloxifene on risk of breast cancer in postmenopausal women. *JAMA.* 1999;281:2189-2197.
**83.** Ettinger B, Black DM, Mitlak BH, et al. Reduction of vertebral fracture risk in postmenopausal women with osteoporosis treated with raloxifene: results from a 3-year randomized trial. *JAMA.* 1999;282:637-645.

# Users' Guides to the Medical Literature: XIX. Applying Clinical Trial Results; A. How to Use an Article Measuring the Effect of an Intervention on Surrogate End Points

Heiner C. Bucher; Gordon H. Guyatt; Deborah J. Cook; et al.

*JAMA*. 1999;282(8):771-778 (doi:10.1001/jama.282.8.771)

http://jama.ama-assn.org/cgi/content/full/282/8/771

# Users' Guides to the Medical Literature

## XIX. Applying Clinical Trial Results
## B. Guidelines for Determining Whether a Drug Is Exerting (More Than) a Class Effect

Finlay A. McAlister, MD, FRCPC
Andreas Laupacis, MD, MSc, FRCPC
George A. Wells, MSc, PhD
David L. Sackett, FRSC, MD, FRCP
for the Evidence-Based Medicine Working Group

MOST CLASSES OF DRUGS include multiple compounds. The opinions of clinicians, manufacturers, and purchasers may differ as to whether a particular drug is more efficacious, safer, or more cost-effective than others in its class.[1] In this article, we review the types of evidence commonly cited to support the prescribing of a particular drug rather than another of the same class and provide a hierarchy for grading studies that compare a drug with another of the same class, expanding on our discussion in part A of this Users' Guide.[2]

## CLINICAL SCENARIOS

### The Clinician

As a clinician, you care for many patients with elevated serum cholesterol levels. A speaker at a recent continuing medical education event reviewed the benefits of cholesterol-lowering therapy, particularly with 3-hydroxy-3-methylglutaryl coenzyme A reductase inhibitors (statins), in the primary and secondary prevention of ischemic heart disease but did not recommend a particular statin. You decide to consider statin therapy for all your patients with elevated cholesterol levels, but are uncertain which of the statins on the market is best. You ask a general internist, cardiologist, and endocrinologist for their opinions, and each suggests a different statin, citing different reasons. You contact pharmaceutical representatives to provide you with evidence that their statins are better than those of their competitors. Although you use the *JAMA* series on Users' Guides to the Medical Literature to assess the validity of published studies, faced with a variety of competing claims, you realize that you need a framework for grading the strength of these studies.

### The Policymaker

Your colleague, a purchaser for a large health maintenance organization (HMO), is faced with a similar dilemma when she is asked to consider replacing the statin on her HMO's formulary with a newer one. She wonders whether there is enough evidence to support the contention that the new statin is as good as, or better than, the one currently on formulary. While the new statin is cheaper, it has been evaluated only in short-term trials, with cholesterol lowering as the solitary end point.

## DRUG CLASSES

Although there is no uniformly accepted definition of a drug class—and some argue that it cannot be defined at all—drugs are generally said to belong to the same class for 1 of 3 reasons (TABLE 1). Herein, we define a drug class as those drugs that share a similar structure and mechanism of action. Most classes of drugs include multiple compounds, and because of their similar mechanisms of action, they are generally thought to confer similar pharmacologic effects and clinical outcomes (*class effects*). This assumption is a key medical heuristic[3] and underlies clinical practice guidelines in which evidence from studies involving 1 or more drugs within a class is extrapolated to other drugs of the same class. For example, it is recommended that β-blockers be prescribed for survivors of myocardial infarction or angiotensin-converting enzyme inhibitors to patients with heart failure. In this circumstance, clinicians are likely to be interested in the drug within each class

**Author Affiliations:** National Health Service Research and Development Centre for Evidence-Based Medicine, John Radcliffe Hospital, Oxford, England (Drs McAlister and Sackett); Division of General Internal Medicine, Ottawa Hospital (Dr Laupacis), and Clinical Epidemiology Unit, Loeb Health Research Institute (Drs Laupacis and Wells), Ottawa, Ontario.

Dr McAlister is currently at the Division of General Internal Medicine, University of Alberta Hospital, Edmonton.

The original list of members (with affiliations) appears in the first article of this series (*JAMA*. 1993; 270:2093-2095). A list of new members appears in the 10th article of the series (*JAMA*. 1996;275:1435-1439). The following members contributed to this article: Gordon Guyatt, MD, MSc, Virginia Moyer, MD, MPH, and David Naylor, MD, DPhil.

**Corresponding Author:** Finlay A. McAlister, MD, FRCPC, Division of General Internal Medicine, 2E3.24 WMC, University of Alberta, Canada Hospital, 8440 112 St, Edmonton, Alberta, Canada T5G 2R7.

**Reprints:** Gordon Guyatt, MD, MSc, McMaster University Health Sciences Centre, 1200 Main St W, Room 2C12, Hamilton, Ontario, Canada L8N 3Z5.

**Users' Guides to the Medical Literature Section Editor:** Drummond Rennie, MD, Deputy Editor (West), *JAMA*.

**Table 1.** Definitions of Drug Classes*

| Definition | Example |
|---|---|
| Drugs with similar chemical structure | Dihydropyridine CCBs have dihydropyridine rings |
| Drugs with similar mechanism of action | CCBs block the voltage-dependent calcium channels on the surfaces of cell membranes |
| Drugs with similar pharmacologic effects | Antihypertensives (eg, CCBs, ACE inhibitors, β-blockers, thiazides, α-blockers) lower blood pressure |

*CCB indicates calcium channel blocker; ACE, angiotensin-converting enzyme.

with the most attractive efficacy-to-safety ratio; purchasers, in the most cost-effective drug from a class; and manufacturers, in the most frequent prescribing of their drugs.

The absolute treatment effects seen with a drug (defined by the absolute risk reduction or number needed to treat) are influenced by the baseline risk or control event rate of those patients in whom it is used. Thus, the absolute risk reduction varies considerably among different groups of patients. On the other hand, the relative treatment effect of a drug (defined by the relative risk reduction [RRR]) is often (but not always[4]) similar, irrespective of the baseline risk of trial participants.[5,6] If 2 drugs are tested in separate placebo-controlled trials, only proportional effects such as the RRR resulting from each drug can be compared (and then only under the assumption of constant RRR over different control event rates). Although the point estimates of effect size vary, a class effect is considered to be present when drugs with similar mechanisms of action generate RRRs (or odds ratios [ORs]) that are similar in direction and magnitude. For example, the Collaborative Group on ACE Inhibitor Trials[7] suggested that there is a class effect for angiotensin-converting enzyme inhibitors in patients with symptomatic heart failure, despite the fact that the OR point estimates for effects on total mortality ranged from 0.14 (95% confidence interval [CI], 0-7.6) for perindopril (1 trial, 125 patients) to 0.78 (95% CI, 0.67-0.91) for enalapril (7 trials, 3381 patients). We are confident in this class effect, because the overall OR in 32 trials involving 7105 patients was 0.77 (95% CI, 0.67-0.88), the CIs for each of the angiotensin-converting enzyme inhibitors overlapped, and there was no statistical heterogeneity between trials of different agents.

### Risks of Assuming a Class Effect

Although drugs of the same class typically exhibit similar pharmacological effects and clinical outcomes, this may not always be the case. Note the current controversy regarding the safety of sotalol hydrochloride in myocardial infarction survivors with congestive heart failure after the publication of the SWORD Trial,[8] which suggested an increase in mortality with sotalol, compared with the decrease in mortality with other β-blockers. It is useful to recall a previous controversy regarding the efficacy of β-blockers with intrinsic sympathetic activity (ISA) in patients with myocardial infarction. Although a meta-analysis[9] suggested that the treatment effect was greater with non-ISA β-blockers, subsequent trials[10] failed to confirm this, and the evidence[11] suggests there is little difference between β-blocker subgroups. It would seem reasonable to accept a priori that drugs within the same class exert similar effects, unless there is clear evidence of important differences.

However, this assumption can lead to 2 important errors of extrapolation with major clinical consequences. First, when agents in a class of drugs (such as the thiazide diuretics) all produce similar pharmacological effects (blood pressure lowering) and similar clinical effects (stroke reduction), a second class of drugs (for example, the calcium channel blockers) that produce the same pharmacological effects might be assumed to produce the same clinical benefits. In the absence of randomized trials verifying that final assumption, this type of extrapolation may be erroneous. For example, consider the issue raised in part A of this Guide—some calcium channel blockers have unfavorable effects on total mortality.[12] Second, even within the same class, individual drugs may have physiologic effects other than the mechanism of action that defined them as being from the same class. It therefore may be inaccurate to extrapolate the clinical outcomes shown in randomized trials of 1 drug in a class to another member of that class that has not been subjected to similar outcome-centered trials. For example, some authors have argued that, although all of the statins act on the 3-hydroxy-3-methylglutaryl coenzyme A reductase enzyme, they may have different nonlipid effects on the atherothrombotic process that may influence their clinical efficacy.[13]

To reduce the risk of faulty extrapolation and to maximize the optimal selection of treatments within a class of drugs, it may be useful to develop and apply a hierarchy of evidence when making decisions about the comparative clinical efficacy and safety of drugs within a class. As pointed out in part A of this Users' Guide, no matter how strong the pathophysiologic rationale or indirect evidence, the efficacy and safety of a new drug must be established in clinical outcome studies that test more than just biological plausibility.

### Levels of Evidence

Levels of evidence are increasingly used by groups that make recommendations about patient care,[14-16] and we have used some of them to develop guidelines for comparing 1 drug with other drugs in the same class (TABLE 2). This comparison should occur as part of a systematic review of all the relevant evidence on the effects of a treatment, identified and assessed by thorough and clear methods such as those used in the Cochrane Collaboration [Update Software, Oxford, England; 1998]. We will describe each level in turn, using the choice of statin drugs as an example to illustrate their use (TABLE 3).

**Level 1.** Level 1 includes randomized clinical trials providing head-to-head comparisons of the drug of interest with other drugs of the same class for their effects on clinically important outcomes. This would generate the strongest evidence for the decision maker; however, there are potential

threats to validity (Table 2) and several methodologic issues unique to these trials. First, at least 1 of the drugs should have been shown to have a clinically important impact vs placebo in previous trials carried out in a population similar to that of the current trial. Second, the choice of appropriate dosage for each drug is a complicated issue, as this will affect the outcomes and safety profiles for both drugs. Finally, one must carefully consider the trial size and methods before concluding equivalence of 2 drugs—equivalence trials require much larger sample sizes than standard trials,[17] and any laxity in trial conduct or patient compliance will tend to mask any real differences between drugs.

The choice of clinically important outcomes for level 1 studies depends on the target intervention. In the case of therapies designed to prevent or arrest atherosclerosis (such as statins), this implies long-term efficacy data on events such as myocardial infarction, stroke, and all-cause mortality. On the other hand, for interventions designed to treat symptomatic diseases (such as gastroesophageal reflux disease), clinically important outcomes could include symptom scores and other quality-of-life measures.

Although there are examples of level 1 evidence in other branches of medicine,[18,19] they are rare in the cardiovascular literature. Our literature search failed to find any level 1 evidence for statins.

**Level 2.** Level 2 includes randomized clinical trials providing head-to-head comparisons of the drug of interest with other drugs of the same class for their effects on validated surrogate outcomes or comparisons across 2 or more placebo-controlled trials for effects on clinically important outcomes or validated surrogate outcomes. Part A of this Users' Guide discussed criteria for deciding whether to accept results of trials based on surrogate outcomes. Ecologic studies, cohort studies, and randomized clinical trials with prestatin lipid-lowering agents were supportive of the lipid-lowering hypothesis[20] (that lowering low-density lipoprotein [LDL] cholesterol levels lowers the risk of atherosclerotic heart disease); however, it was not until the publication of the large-scale statin trials[21-25] (Table 3) consistently linking reductions in LDL cholesterol to reductions in morbidity and mortality that we agreed to accept the surrogate end point of LDL cholesterol lowering as a proxy for clinically important outcomes. Thus, to accept head-to-head comparisons for surrogate outcomes as level 2 evidence, at least 1 of the comparators must have demonstrated efficacy in long-term trials with clinically important outcomes.

Whereas a randomized trial[26] comparing 4 statins for their effects on LDL cholesterol, high-density lipoprotein cholesterol, and triglycerides during an 8-week period would be an example of level 2 evidence, it also is important to incorporate considerations of the size and duration of trials in the decision-making process.

On the other hand, a number of level 2 comparisons can be made between various statins—for example, one can compare the treatment effects seen with simvastatin vs pravastatin in secondary prevention trials (such as the 4S[21] and LIPID[25] studies [Table 3]). Although consistency of effects in such comparisons would be strong evidence for the presence of a class effect, these comparisons are less useful in determining whether a drug is more efficacious than another,

**Table 2.** Levels of Evidence for Comparing the Efficacy of Drugs Within the Same Class*

| Level | Comparison | Study Patients | Outcomes | Threats to Validity |
|---|---|---|---|---|
| 1 | Within a head-to-head RCT | Identical (by definition) | Clinically important | Failure to conceal randomization scheme<br>Failure to achieve complete follow-up<br>Failure to achieve double-blinding<br>Soundness of outcome assessment |
| 2 | Within a head-to-head RCT | Identical (by definition) | Validated surrogate | Those of level 1 *plus* validity of surrogate outcome for clinically important outcomes |
| 2 | Across RCTs of different drugs vs placebo | Similar or different (in disease and risk factor status) | Clinically important or validated surrogate | Those of level 1 *plus* differences between trials in:<br>Methodologic quality (adequacy of blinding, allocation concealment, etc)<br>End point definitions<br>Compliance rates<br>Baseline risk of outcomes |
| 3 | Across subgroup analyses from RCTs of different drugs vs placebo | Similar or different | Clinically important or surrogate | Those of level 1 (*plus or minus* those of level 2) *plus:*<br>Multiple comparisons, posthoc data dredging<br>Underpowered subgroups<br>Misclassification into subgroups |
| 3 | Across RCTs of different drugs vs placebo | Similar or different | Unvalidated surrogate | Surrogate outcomes may not capture all of the effects (beneficial or hazardous) of a therapeutic agent |
| 4 | Between nonrandomized studies (observational studies and administrative database research) | Similar or different | Clinically important | Confounding by indication, compliance, and/or calendar time<br>Unknown/unmeasured confounders<br>Measurement error<br>For outcomes research: limited databases, coding systems not suitable for research |

*Clinically important outcomes refer to long-term efficacy data, and the particular end points depend on the condition being treated. For statins used to prevent or treat atherosclerotic disease, clinically important outcomes would include all-cause mortality, myocardial infarction, and stroke. Surrogate outcomes are considered validated only when the relationship between the surrogate outcome and clinically important outcomes has been established in long-term randomized clinical trials (RCTs).

because the advantages of randomization are lost, and the comparison is essentially that between 2 or more cohorts. In addition to the potential biases outlined in Table 2, there is also the possibility of confounding a subject's risk or responsiveness with exposure to a particular treatment in those situations in which subjects from different trials have different risk statuses. For example, if one were to compare the statin used in a primary prevention trial (such as lovastatin in AFCAPS/TexCAPS[24]) with another statin tested in a secondary prevention trial (such as simvastatin in 4S[21]), such a comparison would only be valid if the drug efficacy is known to be independent of baseline risk, an assumption that appears valid to make in some situations (such as antiplatelet[6] or antihypertensive[5] therapy) but has been questioned for the statins.[27-32]

It is theoretically possible to compare the efficacy of 2 drugs tested in separate placebo-controlled trials. As outlined by Bucher et al,[33] an indirect estimate of the association between drugs A and B can be obtained by comparing the OR (or relative risk) from studies of drug A vs placebo (p) and the OR from studies comparing drug B vs placebo: $OR_{A \text{ vs } B} = OR_{A \text{ vs } p} / OR_{B \text{ vs } p}$. However, this assumes that none of the potential biases outlined in Table 2 are operative and that an intervention's treatment effect is consistent across different patient subgroups. Furthermore, these indirect estimates may provide substantially different effect-size estimates than direct comparisons of drug A against drug B. For example, a systematic overview of strategies to prevent *Pneumocystis cari-*

**Table 3.** Features of Randomized, Placebo-Controlled Statin Trials Designed to Detect Differences in Clinically Important End Points*

| | Trial | | | | |
|---|---|---|---|---|---|
| | 4S[21] | WOSCOPS[22] | CARE[23] | AFCAPS/TexCAPS[24] | LIPID[25] |
| Study design | Secondary prevention, multicenter | Primary prevention, single center | Secondary prevention, multicenter | Primary prevention, multicenter | Secondary prevention, multicenter |
| Treatment (dose once daily) | Simvastatin (20 mg) | Pravastatin (40 mg) | Pravastatin (40 mg) | Lovastatin (40 mg) | Pravastatin (40 mg) |
| Patient inclusion criteria† | Age 35-70 y, prior angina or AMI, fasting total cholesterol 5.5-8.0 mmol/L | Age 45-64 y, no prior AMI, fasting LDL cholesterol 4.0-6.0 mmol/L | Age 21-75 y, prior AMI, fasting LDL cholesterol 3.0-4.5 mmol/L | Age 45-73 y (males) or 55-73 y (females), no prior AMI, fasting LDL cholesterol 3.4-4.9 mmol/L | Age 31-75 y, prior AMI or unstable angina, fasting total cholesterol 4.0-7.0 mmol/L |
| Cointerventions, % | | | | | |
| Aspirin | 37 | None | 83 | None | 82 |
| β-Blockers | 57 | None | 40 | None | 47 |
| Duration of follow-up, y | 5.4 (Median) | 4.9 (Mean) | 5.0 (Median) | 5.2 (Mean) | 6.1 (Mean) |
| Patients | | | | | |
| No. | 4444 | 6595 | 4159 | 6605 | 9014 |
| Mean age, y | 58.6 | 55.2 | 59 | 58 | 62 |
| Males, % | 81 | 100 | 86 | 85 | 83 |
| Smokers, % | 26 | 44 | 21 | 12 | 10 |
| Diabetes mellitus, % | 5 | 1 | 15 | 2 | 9 |
| Baseline cholesterol, mean mmol/L† | | | | | |
| Total | 6.8 | 7.0 | 5.4 | 5.7 | 5.6 |
| LDL | 4.9 | 5.0 | 3.6 | 3.9 | 3.9 |
| Control event rates, % | | | | | |
| Death | 11.5 | 4.1 | 9.4 | 0.44 | 14.1 |
| AMI | 22.6 | 7.9 | 10 | 0.56 | 10.3 |
| Treatment effects | | | | | |
| Change in lipids (active treatment vs placebo), % | −25 (Total cholesterol)<br>−35 (LDL cholesterol)<br>+8 (HDL cholesterol)<br>−10 (Triglycerides) | −20 (Total cholesterol)<br>−26 (LDL cholesterol)<br>+5 (HDL cholesterol)<br>−12 (Triglycerides) | −20 (Total cholesterol)<br>−28 (LDL cholesterol)<br>+5 (HDL cholesterol)<br>−14 (Triglycerides) | −18 (Total cholesterol)<br>−25 (LDL cholesterol)<br>+6 (HDL cholesterol)<br>−15 (Triglycerides) | −18 (Total cholesterol)<br>−25 (LDL cholesterol)<br>+5 (HDL cholesterol)<br>−11 (Triglycerides) |
| Relative risk reductions, % (95% CI) | | | | | |
| Death | 30 (15 to 42) | 22 (0 to 40) | 9 (−12 to 26) | −4 (Not given) | 22 (13 to 31) |
| AMI | 27 (20 to 34) | 31 (17 to 43) | 25 (8 to 39) | 40 (17 to 57) | 29 (18 to 38) |
| Number needed to treat‡ | | | | | |
| To prevent 1 death | 27 (5 y) | 111 (5 y) | 125 (5 y) | 5000 to harm§ | 32 (6 y) |
| To prevent 1 AMI | 10 (5 y) | 42 (5 y) | 40 (5 y) | 435 (5 y) | 34 (6 y) |

*4S indicates Scandinavian Simvastatin Survival Study; WOSCOPS, West of Scotland Coronary Prevention Study; CARE, Cholesterol and Recurrent Events Trial; AFCAPS/TexCAPS, Air Force/Texas Coronary Atherosclerosis Prevention Study; LIPID, Long-term Intervention with Pravastatin in Ischaemic Disease Study; AMI, acute myocardial infarction; LDL, low-density lipoprotein; HDL, high-density lipoprotein; and CI, confidence interval.
†To convert cholesterol levels to milligrams per deciliter, divide by 0.02586.
‡Point estimates only. Years in parentheses indicate number of years needed to treat that number of patients to prevent 1 event.
§Since all-cause mortality was nonsignificantly increased in the active treatment arm, results are presented as number needed to treat to cause 1 death.

*jiri* pneumonia in human immunodeficiency virus–positive patients documented that the indirect comparison of trimethoprim-sulfamethoxazole vs a combination of dapsone and pyrimethamine suggested a much larger effect size from trimethoprim-sulfamethoxazole (OR, 0.37; 95% CI, 0.21-0.65) than was seen in the direct comparisons (overall OR, 0.64 in the 9 trials of trimethoprim-sulfamethoxazole vs dapsone and pyrimethamine; 95% CI, 0.45-0.90).[33] Thus, the strength of inference from indirect comparisons is limited.

**Level 3.** Level 3 includes comparisons across subgroups from different placebo-controlled trials or comparisons across placebo-controlled trials in which outcomes are restricted to unvalidated surrogate markers. In addition to the biases that affect higher-level studies, comparisons based on subgroup analysis are potentially flawed (Table 2). Both simple statistics and experience have taught us that many initial subgroup conclusions (especially those that result from data-dredging) are subsequently disproven.[34,35] An example of such a comparison would be looking at the efficacy of simvastatin in the 4S subgroup with the lowest lipid levels (241 patients with total cholesterol levels of 5.5-6.24 mmol/L [213-241 mg/dL])[28] vs the efficacy of pravastatin in the CARE subgroup with comparable lipid profiles (2087 patients with total cholesterol levels of 5.4-6.21 mmol/L [209-240 mg/dL]).[23]

Level 3 evidence may also include the use of surrogate markers that, although they may lie along a recognized pathogenetic pathway from mechanisms of action to important clinical outcomes, have not been validated in long-term randomized clinical trials. To return to an example cited in part A of this Users' Guide, this would involve making inferences about reductions in fractures from the effects on bone density of 2 different bisphosphonates in 2 independent randomized trials.

**Level 4.** Level 4 includes comparisons involving or confined to nonrandomized evidence. This type of evidence is only possible for conditions in which there are a large number of potential treatments commonly used by practitioners. Nonrandomized evidence can include cohort or case-control studies, modeling studies (using risk-prediction equations such as those derived from the Framingham data[36]), and/or outcomes research using administrative databases. Although these types of analyses can provide useful insights (particularly with respect to dose-response relationships),[37] they are best viewed as exercises in hypothesis-generation. In particular, outcomes research studies, originally developed to determine whether the efficacy of interventions proven in randomized trials have their anticipated impacts at a population level, have sometimes been used to pursue the primary determination of efficacy—a purpose for which they were not intended. When used to establish efficacy, they present, in addition to other limitations (Table 2), unique problems in interpretation that restrict the validity of inferences drawn from them about the relative efficacy of medications from the same class.[38]

An example of level 4 evidence is a recent reanalysis of the WOSCOPS database, designed to infer whether pravastatin's efficacy exceeds that expected of other statins.[29] Using the constellation of risk factors and mean on-treatment cholesterol levels seen in the trial, the observed coronary event rates in pravastatin-treated patients were compared with those predicted from the Framingham coronary risk equation to determine whether the treatment benefit with pravastatin exceeded that expected from the degree of cholesterol lowering achieved.

Level 3 and 4 studies have numerous flaws as outlined above and are best viewed as exercises in hypothesis generation.

## Other Considerations

**Amount of Efficacy Evidence.** While we have thus far focused on the validity of the evidence, the number, size, and duration of studies are essential factors to be considered in the decision-making process. Certainly, the superiority of 1 drug within a class can only be definitively established with level 1 evidence. However, while level 1 evidence would be ideal for establishing that a group of drugs exert a class effect (by showing narrow confidence limits around the difference between drugs), we recognize that it is rarely available and is unlikely to ever be available for many classes of drugs because of difficulties in funding and conducting trials so large that they are unlikely to appeal to researchers, manufacturers, or funders. In this situation, the amount of level 2 evidence becomes important. For instance, one would feel more comfortable in concluding that a drug produced a class effect if there were a number of placebo-controlled trials demonstrating that various drugs from the same class had similar treatment effects. However, our goal is not to set a level that must be achieved before a drug can be claimed to be superior to others in its class or before a class effect can be established. Those are decisions that individual clinicians or policymakers must make, taking into account their local circumstances and individual comfort levels.

**Safety.** In the past decade, there have been numerous examples of drugs within the same class that have been shown to have different safety profiles. Although not our primary focus, considerations of drug safety are part of any treatment or purchasing decision, so we offer a set of levels of evidence for determining drug safety in TABLE 4. Phase 1 drug studies in humans are designed to determine the maximally tolerated dose, and clinical trials are generally designed to determine the efficacy of the drug. As such, the sample sizes of neither are adequate to detect uncommon adverse effects. The inverse rule of 3 states that to be 95% sure of seeing at least 1 adverse drug reaction that occurs once in every given number of patients, you need to follow up 3 times that many patients.[39] Given the size and duration of most clinical trials, adverse effects that occur in fewer than 1 in 1000 participants or that take more than 6 months to appear will generally remain undetected.[3] However, randomized clinical trials are still the strongest design for detecting real differences in adverse effects (such as the different rates of intracranial bleeding with different thrombolytic agents[40,41]), and meta-analyses of such

trials can give unbiased estimates of excess hazards. In the absence of clinical trials, premarketing safety data must be considered preliminary, and large, phase 4 studies or systematic postmarketing surveillance data are necessary to confirm the safety of new drugs.

**Convenience/Compliance.** While once-a-day medications are more convenient and usually have higher compliance rates, evidence about drug compliance derived from trials may translate poorly in clinical practice. For instance, while compliance with the various statins described in Table 3 ranged from 90% to 94% during the course of the trials, analyses of administrative databases in Canada and the United States[42] revealed that only half of statin-treated patients were still taking their medication 1 year after it was prescribed.

**Cost.** Faced with a decision as to whether a new drug from a class should be offered to eligible patients within the population, clinicians and policymakers have different perspectives. For clinicians, this decision usually hinges on the efficacy, safety, convenience or compliance, cost of the new drug vs the old, and the applicability of the trial evidence to their patients.[43] However, for policymakers, these issues form only 1 piece of the puzzle. They also must evaluate the efficiency, affordability, and opportunity costs of any new drugs. The efficiency of any intervention is determined by formal economic analyses, and the Users' Guides series offers criteria for evaluating methodological quality.[44] Although cost-minimization analysis is the simplest and least controversial of the economic analysis techniques, it requires proof that the outcomes resulting from both alternatives are the same. As this rarely exists, the policymaker must rely on other types of analyses (cost-effectiveness, cost-benefit, or cost-utility analyses) that involve varying degrees of assumption and guesswork. As pointed out by Naylor and colleagues,[45] economic analyses should be viewed as "promising, clearly helpful, still in need of refinement and open, like any new technology, to both wise use and well-intentioned abuse."

The decision as to whether a new drug is efficient enough to warrant its adoption depends critically on the social, political, and economic realities of the particular health care setting, complicating the policymaker's task. Thus, attempts to establish universal cut-points (using cost or quality-adjusted life-year ratios) have been largely unsuccessful.[46] Although there are occasions for which there is compelling evidence for a new drug's adoption (the new drug is as effective or more effective than others of its class and is less costly) or rejection (the new drug is less effective than others of its class and is more costly), the policymaker operates most often in a cost-utility gray zone between these 2 extremes.[45]

**Table 4.** Levels of Evidence for Comparing the Safety of Drugs Within the Same Class

| Level | Type of Study | Advantages | Threats to Validity |
|---|---|---|---|
| 1 | Randomized clinical trial(s) | Only design that permits the detection of adverse effects when the adverse effect is similar to the event that treatment is trying to prevent | Underpowered for detecting adverse effects |
| 2 | Cohort | Prospective data collection, defined cohort | Critically depends on follow-up, classification, and measurement accuracy |
| 3 | Case-control | Cheap and fast to perform | Selection and recall bias; temporal relationship may not be clear |
| 4 | Phase 4 | If sufficiently large, can detect rare but important adverse effects | No, or unmatched, control group; critically depends on follow-up, classification, and measurement accuracy |
| 5 | Case series | Cheap and fast to perform | Small sample size; selection bias; no control group |
| 6 | Case report(s) | Cheap and fast to perform | Small sample size; selection bias; no control group |

## RESOLUTION OF SCENARIOS

### The Clinician

Given the qualitative consistency of the RRR for acute myocardial infarction in patients treated with 3 of the statins in large trials with clinically important outcomes (Table 3) and the convincing nature of LDL cholesterol lowering as a surrogate outcome,[20,30,47-49] our clinician concludes that there is a class effect of statin drugs on the occurrence of ischemic heart disease. In the apparent absence of differences in safety or compliance profile between the various statins, he decides to pursue a cost-minimization strategy. While the newer statin has been evaluated only for cholesterol-lowering efficacy in a short-term trial (<6 months), he decides to prescribe it because it is the cheapest statin in his local setting.

### The Policymaker

The policymaker agrees with the clinician that the statins appear to exert a class effect in terms of efficacy. However, she is concerned that the efficacy of the newer statin has not been evaluated in long-term trials with clinically important outcomes or validated surrogate outcomes. Thus, she decides to keep the older (and more expensive) statin on her formulary until level 1 or long-term level 2 evidence is available that proves that the newer statin is as good as or better than the currently provided statin.

## CONCLUSION

While it would be preferable that every drug in each class (and indeed every dose and every formulation) be evaluated in randomized clinical trials with active comparators from the same class for its effects on clinically important outcomes, this has not been accomplished for several important classes of drugs. We believe that advocates of newer drugs within a class must provide evidence of equivalence (or superiority) to the older agents and "randomized comparative trials . . . remain the preferred evidentiary standard."[50] Recognizing that this criterion standard is not always attainable (in the case of

the statins, such randomized clinical trials would require very large sample sizes and long follow-up to detect significant differences in myocardial infarction or death between 2 different statins), we suggest that discussions about class effects will benefit from citing the levels of evidence behind the arguments and recognizing the strengths and weaknesses inherent in each study design.

**Funding/Support:** Dr McAlister is supported by the Medical Research Council of Canada and the Alberta Heritage Foundation for Medical Research, Dr Laupacis is supported by the Medical Research Council of Canada, and Dr Sackett is supported by the Research and Development Programme of the National Health Service, United Kingdom.

**Acknowledgment:** We gratefully acknowledge the input of various members of the University of Ottawa Clinical Epidemiology Unit and attendees of the EQUINOX symposium for earlier discussions of this topic, J. Glennie, PharmD, for providing information on the approaches of regulators to these issues, and Ian Chalmers, MB, ChB, and F. Lawson, MB, ChB, for reviewing earlier versions of this article.

## REFERENCES

1. Skolnick AA. Drug firm suit fails to halt publication of Canadian Health Technology Report. *JAMA*. 1998;280:683-684.
2. Bucher HC, Guyatt GH, Cook DJ, Holbrook A, McAlister FA, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, XIX: applying clinical trial results, A: how to use an article measuring the effect of an intervention on surrogate end points. *JAMA*. 1999;282:771-778.
3. McDonald CJ. Medical heuristics. *Ann Intern Med*. 1996;124:56-62.
4. Rothwell PM. Can overall results of clinical trials be applied to all patients? *Lancet*. 1995;345:1616-1619.
5. Collins R, Peto R, MacMahon S, et al. Blood pressure, stroke, and coronary heart disease, part 2: short-term reductions in blood pressure. *Lancet*. 1990;335:827-838.
6. Antiplatelet Trialists' Collaboration. Collaborative overview of randomised trials of antiplatelet therapy—I [published correction appears in *BMJ*. 1994;308:1540]. *BMJ*. 1994;308:81-106.
7. Garg R, Yusuf S, for the Collaborative Group on ACE Inhibitor Trials. Overview of randomized trials of angiotensin-converting enzyme inhibitors on mortality and morbidity in patients with heart failure [published correction appears in *JAMA*. 1995;274:462]. *JAMA*. 1995;273:1450-1456.
8. Waldo AL, Camm AJ, de Ruyter H, et al, for the SWORD Investigators. Effect of d-sotalol on mortality in patients with left ventricular dysfunction after recent and remote myocardial infarction [published correction appears in *Lancet*. 1996;348:416]. *Lancet*. 1996;348:7-12.
9. Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction. *Prog Cardiovasc Dis*. 1985;27:335-371.
10. Boissel JP, Leizorovicz A, Picolet H, Peyrieux JC. Secondary prevention after high-risk acute myocardial infarction with low-dose acebutolol. *Am J Cardiol*. 1990;66:251-260.
11. McAlister FA, Teo KK. Antiarrhythmic therapies for the prevention of sudden cardiac death. *Drugs*. 1997;54:235-252.
12. Furberg CD, Psaty BM. Calcium antagonists. *Am J Hypertens*. 1996;9:122-125.
13. Rosenson RS, Tangney CC. Antiatherothrombotic properties of statins. *JAMA*. 1998;279:1643-1650.
14. Guyatt GH, Sackett DL, Sinclair JC, Hayward RC, Cook DJ, Cook RJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, IX: a method for grading health care recommendations [published correction appears in *JAMA*. 1995;275:1232]. *JAMA*. 1995;274:1800-1804.
15. Canadian Task Force on the Periodic Health Examination. The Periodic Health Examination, 2. 1987 update. *CMAJ*. 1988;138:618-626.
16. Cook DJ, Guyatt GH, Laupacis A, Sackett DL, Goldberg RJ. Clinical recommendations using levels of evidence for antithrombotic agents. *Chest*. 1995;108(4 suppl):227S-230S.
17. Donner A. Approaches to sample size estimation in the design of clinical trials. *Stat Med*. 1984;3:199-214.
18. Naguib M, el Bakry AK, Khoshim MH, et al. Prophylactic antiemetic therapy with ondansetron, tropisetron, granisetron and metoclopramide in patients undergoing laparoscopic cholecystectomy. *Can J Anaesth*. 1996;43:226-231.
19. Korttila K, Clergue F, Leeser J, et al. Intravenous dolasetron and ondansetron in prevention of postoperative nausea and vomiting. *Acta Anaesthesiol Scand*. 1997;41:914-922.
20. Law MR, Wald NJ, Thompson SG. By how much and how quickly does reduction in serum cholesterol concentration lower risk of ischaemic heart disease? *BMJ*. 1994;308:367-373.
21. Scandinavian Simvastatin Survival Study Group. Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease. *Lancet*. 1994;344:1383-1389.
22. Shepherd J, Cobbe SM, Ford I, et al, for the West of Scotland Coronary Prevention Study Group. Prevention of coronary heart disease with pravastatin in men with hypercholesterolemia. *N Engl J Med*. 1995;333:1301-1307.
23. Sacks FM, Pfeffer MA, Moye LA, et al, for the Cholesterol and Recurrent Events Trial Investigators. The effect of pravastatin on coronary events after myocardial infarction in patients with average cholesterol levels. *N Engl J Med*. 1996;335:1001-1009.
24. Downs JR, Clearfield M, Weis S, et al, for the AFCAPS/TexCAPS Research Group. Primary prevention of acute coronary events with lovastatin in men and women with average cholesterol levels: results of AFCAPS/TexCAPS. *JAMA*. 1998;279:1615-1622.
25. The Long-term Intervention with Pravastatin in Ischaemic Disease (LIPID) Study Group. Prevention of cardiovascular events and death with pravastatin in patients with coronary heart disease and a broad range of initial cholesterol levels. *N Engl J Med*. 1998;339:1349-1357.
26. Jones P, Kafonek S, Laurora I, Hunninghake D. Comparative dose efficacy study of atorvastatin versus simvastatin, pravastatin, lovastatin, and fluvastatin in patients with hypercholesterolemia [published correction appears in *Am J Cardiol*. 1998;82:128]. *Am J Cardiol*. 1998;81:582-587.
27. Sacks FM, Moye LA, Davis BR, et al. Relationship between plasma LDL concentrations during treatment with pravastatin and recurrent coronary events in the Cholesterol and Recurrent Events Trial. *Circulation*. 1998;97:1446-1452.
28. Scandinavian Simvastatin Survival Study Group. Baseline serum cholesterol and treatment effect in the Scandinavian Simvastatin Survival Study (4S). *Lancet*. 1995;345:1274-1275.
29. West of Scotland Coronary Prevention Study Group. Influence of pravastatin and plasma lipids on clinical events in the West of Scotland Coronary Prevention Study (WOSCOPS). *Circulation*. 1998;97:1440-1445.
30. Fager G, Wiklund O. Cholesterol reduction and clinical benefit: are there limits to our expectations? *Arterioscler Thromb Vasc Biol*. 1997;17:3527-3533.
31. Smith GD, Song F, Sheldon TA. Cholesterol lowering and mortality [published correction appears in *BMJ*. 1993;306:1648]. *BMJ*. 1993;306:1367-1373.
32. Sacks FM, Gibson CM, Rosner B, Pasternak RC, Stone PH, for the Harvard Atherosclerosis Reversibility Project Research Group. The influence of pretreatment low density lipoprotein cholesterol concentrations on the effect of hypocholesterolemic therapy on coronary atherosclerosis in angiographic trials. *Am J Cardiol*. 1995;76:78C-85C.
33. Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *J Clin Epidemiol*. 1997;50:683-691.
34. Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. *Ann Intern Med*. 1992;116:78-84.
35. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA*. 1991;266:93-98.
36. Anderson KM, Wilson PW, Odell PM, Kannel WB. An updated coronary risk profile: a statement for health professionals. *Circulation*. 1991;83:356-362.
37. Psaty BM, Siscovick DS, Weiss NS, et al. Hypertension and outcomes research: from clinical trials to clinical epidemiology. *Am J Hypertens*. 1996;9:178-183.
38. Marshall WJ. Administrative databases: fact or fiction? *CMAJ*. 1998;158:489-490.
39. Sackett DL, Haynes RB, Gent M, et al. Compliance. In: Inman WH, ed. *Monitoring for Drug Safety*. Philadelphia, Pa: Lippincott; 1980:427-438.
40. Gruppo Italiano per lo Studio della Sopravvivenza nell' Infarto Miocardico. GISSI-2: a factorial randomised trial of alteplase versus streptokinase and heparin versus no heparin among 12,490 patients with acute myocardial infarction. *Lancet*. 1990;336:65-71.
41. Third International Study of Infarct Survival Collaborative Group. ISIS-3: a randomised comparison of streptokinase vs tissue plasminogen activator vs anistreplase and of aspirin plus heparin vs aspirin alone among 41,299 cases of suspected acute myocardial infarction. *Lancet*. 1992;339:753-770.
42. Avorn J, Monette J, Lacour A, et al. Persistence of use of lipid-lowering medications: a cross-national study. *JAMA*. 1998;279:1458-1462.
43. Dans AL, Dans LF, Guyatt GH, Richardson S, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, XIV: how to decide on the applicability of clinical trial results to your patient. *JAMA*. 1998;279:545-549.
44. Drummond MF, Richardson WS, O'Brien BJ, Levine M, Heyland D, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, XIII: how to use an article on economic analysis of clinical practice, A: are the results of the study valid? *JAMA*. 1997;277:1552-1557.
45. Naylor CD, Williams JI, Basinski A, Goel V. Technology assessment and cost-effectiveness analysis: misguided guidelines? *CMAJ*. 1993;148:921-924.
46. Laupacis A, Feeny D, Detsky AS, Tugwell PX. How attractive does a new technology have to be to warrant adoption and utilization? *CMAJ*. 1992;146:473-481.
47. Gotto AM Jr. Cholesterol management in theory and practice. *Circulation*. 1997;96:4424-4430.
48. Grover SA, Paquet S, Levinton C, Coupal L, Zowall H. Estimating the benefits of modifying risk factors of cardiovascular disease [published correction appears in *Arch Intern Med*. 1998;158:1228]. *Arch Intern Med*. 1998;158:655-662.
49. Lacour A, Derderian F, LeLorier J. Comparison of efficacy and cost among lipid-lowering agents in patients with primary hypercholesterolemia. *Can J Cardiol*. 1998;14:355-361.
50. Tu JV, Naylor CD. Choosing among drugs of different price for similar indications [published correction appears in *Can J Cardiol*. 1998;14:662]. *Can J Cardiol*. 1998;14:349-351.

# Users' Guides to the Medical Literature: XIX. Applying Clinical Trial Results; B. Guidelines for Determining Whether a Drug Is Exerting (More Than) a Class Effect

Finlay A. McAlister; Andreas Laupacis; George A. Wells; et al.

*JAMA*. 1999;282(14):1371-1377 (doi:10.1001/jama.282.14.1371)

http://jama.ama-assn.org/cgi/content/full/282/14/1371

# Users' Guides to the Medical Literature

## XX. Integrating Research Evidence With the Care of the Individual Patient

Finlay A. McAlister, MD
Sharon E. Straus, MD
Gordon H. Guyatt, MD
R. Brian Haynes, MD
for the Evidence-Based Medicine Working Group

**Clinicians can use research results to determine optimal care for an individual patient by using a patient's baseline risk estimate, clinical prediction guidelines that quantitate an individual patient's potential for benefit, and published articles. We propose that when clinicians are determining the likelihood that treatment will prevent the target event (at the expense of adverse events) in a patient that they also incorporate the patient's values. The 3 main elements to joint clinical decision making are disclosure of information about the risks and benefits of therapeutic alternatives, exploration of the patient's values about both the therapy and potential outcomes, and the actual decision. In addressing the patient's risk of adverse events without treatment and risk of harm with therapy, clinicians must recognize that patients are rarely identical to the average study patient. Differences between study participants and patients in real-world practice tend to be quantitative (differences in degree of risk of the outcome or responsiveness to therapy) rather than qualitative (no risk or adverse response to therapy). The number needed to treat and number needed to harm can be used to generate patient-specific estimates relative to the risk of the outcome event. Clinicians must consider a patient's risk of adverse events from any intervention and incorporate the patient's values in clinical decision making by using information about the risks and benefits of therapeutic alternatives.**

*JAMA. 2000;283:2829-2836* www.jama.com

### CLINICAL SCENARIO

You are the attending physician on an internal medicine service who, one night, admits 2 patients with strokes (patient A, a 65-year-old woman; patient B, a 65-year-old man). On examination, both have mild weakness of the right arm and left carotid bruits. Patient A has a history of hypertension and an admission blood pressure of 200/110 mm Hg; neither patient has other relevant medical history or physical examination findings.

Aware that carotid bruits are not highly specific for identifying carotid artery stenosis, you send both patients for Doppler ultrasonography.[1] Since your radiology department, in a recent audit, demonstrated that their ultrasonographic interpretations are highly correlated with angiographic results,[2] you feel confident from their findings that both patients have moderate stenoses (50%-69% by North American Symptomatic Carotid Endarterectomy Trial criteria) with no irregularity or ulceration of the plaque surface.[3]

Aware of the recent flurry of literature concerning surgical vs medical therapy for patients with symptomatic carotid stenoses, you decide to review the literature to guide your management of these patients. You formulate the question: "In a patient with a mild stroke and moderate ipsilateral carotid stenosis, would a carotid endarterectomy (compared with best medical therapy) reduce the likelihood of subsequent severe stroke or death?"

### THE SEARCH

A systematic review of randomized trials comparing carotid endarterectomy with standard medical therapy (aspirin in your practice setting) in patients with recent mild stroke would provide the best evidence to answer your question. Through your hospital library, you have access to Ovid Evidence-Based Medicine Reviews, allowing you to search both *Best Evidence* (which includes the contents of *ACP Journal Club* and *Evidence-Based Medicine*) and the Cochrane Database of Systematic Reviews with a single search engine. Using the search terms *stroke* and *carotid endarterectomy*, you don't find

**Author Affiliations:** Division of General Internal Medicine, University of Alberta Hospital, Edmonton (Dr McAlister); Division of General Internal Medicine, Mount Sinai Hospital–University Health Network, Toronto (Dr Straus), and Department of Clinical Epidemiology and Biostatistics and Department of Medicine, McMaster University, Hamilton, Ontario (Drs Guyatt and Haynes).
**The original list of members** (with affiliations) appears in the first article of this series (*JAMA*. 1993;270:2093-2095). A list of new members appears in the 10th article of the series (*JAMA*. 1996;275:1435-1439). A full list of the EBM Working Group members, including institutional affiliations and career awards, was presented in the Introduction to this series and in Users' Guide X.
**Corresponding Author:** Sharon E. Straus, MD, Mount Sinai Hospital, Suite 431, 600 University Ave, Toronto, Ontario, Canada, M5G 1X5 (e-mail: sstraus@mtsinai.on.ca).
**Reprints:** Gordon H. Guyatt, MD, McMaster University Health Sciences Centre, Room 2C12, 1200 Main St W, Hamilton, Ontario, Canada L8N 3Z5.
**Users' Guides to the Medical Literature Section Editor:** Drummond Rennie, MD, Deputy Editor

any relevant reviews in the Cochrane Database but you retrieve 18 citations from *Best Evidence*. Scanning these citations you find one that looks relevant to your question[4] and after reviewing the abstract and commentary from *Best Evidence*, you link to the full-text article for further details.

Investigators in this trial randomized 2267 patients with moderate carotid stenosis (<70%) and ipsilateral transient ischemic attacks or nondisabling stroke within 180 days to carotid endarterectomy or medical care alone.[4] After 5 years of follow-up, significantly fewer patients in the carotid endarterectomy arm (vs the medical care arm) had suffered a recurrent disabling stroke (5.3% vs 10.3%; 49% relative risk reduction [RRR]; 95% confidence interval [CI], 14% to 83%]) or death (13% vs 15%; 13% RRR; 95% CI, −18% to 44%). The size of the treatment effect was such that 20 patients (95% CI, 12 to 70) would have to undergo carotid endarterectomy to prevent 1 disabling stroke that would occur with medical therapy alone. Although encouraged by these results, you are concerned about the wide CIs and the potential for perioperative complications (1.4% excess risk of disabling stroke or death within the first month of surgery), and you question how to apply the results to your patients.

## INTRODUCTION

While randomized trials provide the most valid estimates of the true effects (both beneficial and harmful) of an intervention, they necessarily report average treatment effects. Whether these results are derived from a homogeneous group of high-risk, highly responsive patients (as in efficacy trials) or a heterogeneous group of "all-comers" (as in effectiveness trials),[5] clinicians must decide how to extrapolate the results to individual patients. In this article, we will build on previous Users' Guides[6-9] that assessed the validity and applicability of therapeutic studies to outline a framework that clinicians might use to integrate research results (whether from single trials or systematic reviews) with patient values to determine the optimal care for an individual patient.

## DETERMINING THE APPLICABILITY OF THE EVIDENCE TO AN INDIVIDUAL PATIENT

Previous Users' Guides and other articles have dealt extensively with issues of determining the applicability of evidence to individual patients.[7-10] We will not repeat all of the key principles here, but will emphasize that differences between study participants and patients in real-world practice tend to be quantitative (differences in degree of risk of the outcome or responsiveness to therapy) rather than qualitative (no risk or adverse response to therapy).[8,10] These variations may be unimportant (eg, angiotensin converting enzyme inhibitors appear to exhibit similar beneficial effects in patients with systolic congestive heart failure regardless of cause, severity of symptoms, age, or sex)[11] or easily remediable (eg, drug dosages can be adjusted based on individual patient responsiveness).

Restricting efficacious therapies to "ideal patients" may result in significant harm to those excluded. For example, while β-blockers are prescribed to only a minority of patients with acute myocardial infarction, myocardial infarction patients with concomitant conditions that might lead clinicians to withhold treatment (such as peripheral vascular disease, diabetes mellitus, heart failure, or chronic obstructive pulmonary disease) derive substantial survival benefits from β-blocker therapy.[12] This message is a consistent theme emerging from cardiovascular outcomes research.[13]

A key element to consider in extrapolating the results of the carotid endarterectomy trial that you identified is local surgical expertise because the net benefits in the trial were highly sensitive to perioperative complication rates. In fact, the benefits from carotid endarterectomy in this trial (expressed as RRR in disabling stroke) would be reduced by 20% for each 2% absolute increase in the rate of perioperative stroke and death.[14] Moreover, surgical teams whose complication rates and operative volumes would have rendered them ineligible for the trial perform the majority of endarterectomies in North America.[15] Thus, as has been pointed out by others, "caution should be exercised in drawing conclusions about the effectiveness of carotid endarterectomy in the general population on the basis of trials of clinical efficacy conducted at highly selected facilities."[15]

### Individualizing Treatment Decision

The process of individualizing research evidence to the care of a particular patient incorporates 2 components: determining the likelihood that treatment will prevent the target event (at the expense of adverse events) in that patient and incorporating the patient's values. We will now consider both of these steps in some depth.

### Determining the Benefit-Risk Ratio in an Individual Patient

Although we can summarize the results of randomized trials with binary outcomes in a number of ways, the number of patients that would need to be treated to prevent 1 additional adverse event (number needed to treat [NNT])[16] has gained widespread acceptance as 1 clinically relevant format.[17,18] The NNT is the inverse of the difference in absolute event rates between the experimental and control arms and thus reflects baseline risk as well as treatment effect.[17] For example, the NNT to prevent 1 disabling stroke in patients with moderate carotid artery stenosis is 20, calculated as follows: control event rate (10.3%) minus experimental event rate (5.3%) equals absolute risk reduction (5%). The NNT is the inverse of the absolute risk reduction (1/0.05=20).[4]

Analogous to the NNT, the number needed to harm (NNH) is an expression of the number of patients who would need to receive an intervention

to cause 1 additional adverse event. The NNH is the inverse of the absolute difference in adverse event rates between the experimental and control arms. For example, a meta-analysis of 51 studies of carotid endarterectomy in patients with symptomatic carotid stenosis found that the absolute perioperative mortality rate was 1.6% higher with endarterectomy than with medical treatment: this translates into an NNH to cause 1 additional death in the perioperative period with carotid endarterectomy of 63 compared with withholding surgery.[19]

While one can easily calculate NNT when investigators report event rates and relative risks (RRs), difficulties arise when investigators report only odds ratios (ORs). Since the OR is not always an accurate estimate of the RR (particularly as disease incidence increases above 10%),[20] the clinician must employ standard formulas[18] to derive the NNT or NNH from the OR (TABLE). Alternatively, a nomogram has been developed for converting ORs to RRs.[21]

The average NNT (or NNH) reported in a trial or systematic review may not be directly applicable to an individual patient (because of differences in baseline risk and/or RRR across subgroups), and the clinician is faced with 3 questions in extrapolating to his or her patient: Is my patient's RRR likely to be different from the group average? What is my patient's baseline risk of the target event? What is my patient's risk of harm from the treatment?

Although we often assume that RRRs are constant across the limited range of susceptibilities normally encountered in clinical practice,[22-24] recently published studies have demonstrated that while this is often the case,[25-33] it may not always be.[31-35] Thus, the clinician must carefully scrutinize the reports of trials or systematic reviews for information on the relative treatment effects in different subgroups and should use available criteria for evaluating subgroup analyses.[24] In situations where RRR does appear to differ across subgroups, clinicians should employ the RRR from the subgroup most similar to their patient.

Returning to our clinical scenario, the RRR for stroke with carotid endarterectomy does differ by degree of stenosis and presurgical symptom status.[14] Because our patients have symptomatic stenoses of 50%-69%, it would be inappropriate to extrapolate directly the results from either a trial of symptomatic patients with high-grade stenoses (>70%)[36] or a trial of asymptomatic patients with moderate stenoses[37] to their situation. However, it is possible to extrapolate from the previously identified study[4] that enrolled symptomatic patients with similar degrees of stenoses as our patients.

We will now outline 2 approaches to addressing the latter 2 questions, our patient's risk of adverse events without treatment and our patient's risk of harm with therapy.[22] Both approaches that are described below require time, but with the explosion in the development of electronic evidence resources, this obstacle may be ameliorated in the near future.

**Table.** Deriving the Number Needed to Treat and Number Needed to Harm From the Odds Ratio*

| Control Event Rate | Therapeutic Intervention (OR) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 0.50 | 0.55 | 0.60 | 0.65 | 0.70 | 0.75 | 0.80 | 0.85 | 0.90 |
| | Deriving NNT† | | | | | | | | |
| 0.05 | 41 | 46 | 52 | 59 | 69 | 83 | 104 | 139 | 209 |
| 0.1 | 21 | 24 | 27 | 31 | 36 | 43 | 54 | 73 | 110 |
| 0.2 | 11 | 13 | 14 | 17 | 20 | 24 | 30 | 40 | 61 |
| 0.3 | 8 | 9 | 10 | 12 | 14 | 18 | 22 | 30 | 46 |
| 0.4 | 7 | 8 | 9 | 10 | 12 | 15 | 19 | 26 | 40 |
| 0.5 | 6 | 7 | 8 | 9 | 11 | 14 | 18 | 25 | 38 |
| 0.7 | 6 | 7 | 9 | 10 | 13 | 16 | 20 | 28 | 44 |
| 0.9 | 12 | 15 | 18 | 22 | 27 | 34 | 46 | 64 | 101 |
| Control Event Rate | Therapeutic Intervention (OR) | | | | | | | | |
| | 1.1 | 1.2 | 1.3 | 1.4 | 1.5 | 2.0 | 2.5 | 3.0 | 3.5 |
| | Deriving NNH‡ | | | | | | | | |
| 0.05 | 212 | 106 | 71 | 54 | 43 | 22 | 15 | 12 | 9 |
| 0.1 | 112 | 57 | 38 | 29 | 23 | 12 | 9 | 7 | 6 |
| 0.2 | 64 | 33 | 22 | 17 | 14 | 8 | 5 | 4 | 4 |
| 0.3 | 49 | 25 | 17 | 13 | 11 | 6 | 5 | 4 | 3 |
| 0.4 | 43 | 23 | 16 | 12 | 10 | 6 | 4 | 4 | 3 |
| 0.5 | 42 | 22 | 15 | 12 | 10 | 6 | 5 | 4 | 4 |
| 0.7 | 51 | 27 | 19 | 15 | 13 | 8 | 7 | 6 | 5 |
| 0.9 | 121 | 66 | 47 | 38 | 32 | 21 | 17 | 16 | 14 |

*Adapted from McQuay and Moore.[18] OR indicates odds ratio; NNT, number needed to treat; CER, control event rate; and NNH, number needed to harm. Data are presented as number.
†The formula for determining NNT is [1−(CER × (1 − OR))]/[(1 − CER) × CER × (1 − OR)].
‡The formula for determining NNH is 1 + [CER × (OR − 1)]/[(1 − CER) × (CER) × (OR − 1)].

### Approach 1: Generation of Patient-Specific Baseline Risks

Recognizing that patients are rarely identical to the average study patient, clinicians can derive estimates of the patient's baseline risk from various sources. First, if the study reports risk in various subgroups, clinicians can use the baseline risk for the subgroup most like their patient. However, most trials are not large enough to allow generation of precise estimates of baseline risk in various patient subgroups, and the clinician may have to search for systematic reviews (particularly those including individual patient data)[38] to glean useful information. For example, the Atrial Fibrillation Investigators pooled the individual patient data from all randomized trials testing antithrombotic therapy in nonvalvular atrial fibrillation and were able to provide estimates of prognosis for patients in clinically important subgroups.[25]

Second, as an extension of the subgroup approach, one can use clinical prediction guides to quantitate an individual patient's potential for benefit (and harm) from therapy.[33,39,40] Returning to our example, a prognostic model that could identify patients with carotid stenosis most likely to benefit from endarterectomy would be useful. Such a model would need to incorporate the risk of stroke without surgery (and thus the potential benefit from surgery) with the risk of stroke or other adverse outcomes from surgery. Using the European Carotid Surgery Trial database,[41] investigators have developed a preliminary version of just such a model.[42] However, our enthusiasm for applying this clinical prediction guide should be tempered until it has been prospectively validated in a different group of patients (and preferably with different clinicians).[39]

Third, clinicians could derive an estimate of their patient's baseline risk from published articles (preferably population-based cohort studies)[43] that describe the prognosis of similar (untreated) patients. For example, analysis of the Malmo Stroke Registry demonstrated that in the 3 years after a stroke, patients have a 6% risk of recurrent nonfatal stroke and a 43% risk of death; these risks were higher in older patients or those with diabetes mellitus or cardiac disease.[44]

Analogous to the estimation of patient-specific baseline risk, clinicians can use these same sources of information to determine an individual patient's likelihood of harm from treatment. For example, a systematic review of 36 studies relating the risk of perioperative complications from carotid endarterectomy to various preoperative clinical characteristics revealed that women were at higher risk than men (OR, 1.44; 95% CI, 1.14 to 1.83; absolute rate, 5.2%).[45]

The final step in generating a patient-specific NNT (or NNH) involves the formula: NNT = 1/(PEER × RRR) (where PEER is the patient's estimated event rate or baseline risk).[21] Given the 3-year risk of recurrent disabling stroke in diabetic patients from the Malmo Stroke Registry (8.4%)[44] and the 49% RRR expected with carotid endarterectomy,[4] the patient-specific NNT in a 65-year-old patient with diabetes, ipsilateral carotid stenosis, and a minor stroke would be calculated as NNT = 1/(0.084 × 0.49) = 24. Clinicians who know a patient's baseline risk and RRR can also use a nomogram to calculate the NNT.[46]

### Approach 2: Clinical Judgment

Alternately, the clinician can use the NNT and NNH directly from a study to generate patient-specific estimates. This method involves only 2 steps and is less time-consuming than the previous method (because, depending on the experience of the clinician, it may not require a detailed literature review).

First, the clinician estimates the patient's risk of the outcome event relative to that of the average control patient in the study and converts this risk to a decimal fraction (labeled $f_t$, "for treatment").[47] Patients judged to be at less risk than those in the trials will be assigned an $f_t$ less than 1 and those thought to be at greater risk will be assigned an $f_t$ greater than 1. There are several sources that a clinician can use to obtain a value for $f_t$. The best estimate would come from a systematic review of all available data about the prognosis of similar patients; individual studies about prognosis would provide the next best estimates. Alternatively, the clinician could use clinical expertise in assigning a value to $f_t$. While this may appear to be overly subjective, preliminary data suggest that experienced clinicians may be accurate in estimating relative differences in baseline risk (ie, $f_t$) between patients (far exceeding our abilities to judge absolute risks).[48]

Second, the clinician calculates the patient-specific NNT by dividing the average NNT by $f_t$. Thus, if the clinician felt that patient A was at one fifth ($f_t$=,2) the risk of the average patient in the trial (based on the reduced baseline risk for women demonstrated in the subgroup analyses reported by the investigators),[4] her patient-specific NNT for the prevention of 1 disabling stroke would be 100 (20/0.2).

In addition to considering the benefits from therapy, the clinician needs to consider a patient's risk of adverse events from any intervention. Patients A and B need to be informed that carotid endarterectomy does carry with it a risk of perioperative death. To individualize your patient's risk of death, you can use the $f$ method just described (labeled $f_h$, "for harm"). For example, patient A may be assumed to be at twice the risk ($f_h$=2) of perioperative death as patients in the control group of the study because of her gender, hypertension, and the fact that she has left-sided carotid artery stenosis.[4,45] You can adjust the NNH using $f_h$, assuming the RR increase is constant across the spectrum of susceptibilities (an assumption that, as we've noted for RRR, may or may not hold depending on the particular therapy being considered). Thus, patient A's NNH is estimated to be approximately 32 (63/2).

## INCORPORATING PATIENT VALUES AND PREFERENCES

We have determined the risks of benefit and harm for the individual, but we must still incorporate patient values into

the decision-making process. As outlined in a previous Users' Guide,[9] systematically constructed decision analyses and practice guidelines that include an explicit statement of values can be used to integrate the evidence on benefit or harm with patient values to reach treatment recommendations or establish threshold NNTs.[9,49] Although this situation would be ideal, such evidence is often not available (we could not, for instance, identify a relevant decision analysis for our scenario). Moreover, as there is often substantial variation in values between individuals,[50-52] decision analyses that rely on group averages for values may not always be applicable to a particular patient, although close examination of the utility sensitivity analyses of a decision analysis may provide some guidance.[53-55]

While active patient involvement in decision making can improve outcomes and reported quality of life and possibly reduce health care expenditures,[56-62] the initial step in this process is to determine the extent to which your patient wants to be involved in decision making (recognizing that this may vary with each clinical decision).

### How Much Do Patients Want to Participate?

There are 3 main elements to clinical decision making: the disclosure of information (about the risks and benefits of therapeutic alternatives); the exploration of the patient's values about both the therapy and the potential health outcomes; and the actual decision. Each patient varies in desired level of involvement in these steps, and clinicians may not accurately gauge the degree to which an individual patient wants to be involved.[63-68] Some patients may want all available information provided to them and may want to make the decision themselves, with the clinician's role being that of information provider. Other patients may want all the information provided but may want the clinician to make the final decision. Still others may want to collaborate with their clinician in the entire process. These differences emphasize the need for clinicians to accurately assess patient preferences for information, discussion, and decision making and tailor their approach to the individual.

Regardless of whether the clinician, the patient, or both in partnership will make the decision, clinicians must explore patients' values about the therapy and the potential health outcomes. You can elicit your patient's values in informal ways during exploratory discussions or by more formal (and time-consuming) methods such as the time trade-off, standard gamble, or rating scale techniques.[69]

### Decision Aids

If your patient's goal is shared decision making, there are several models available for providing shared decision-making support. First, formal clinical decision analysis, incorporating the patient's likelihood of the outcome events with his or her own values for each health state, could be used to guide the decision. Performing a clinical decision analysis for each patient would be too time-consuming for the busy clinician, and this approach therefore currently relies on finding an existing decision analysis. To be able to use the existing decision analysis, either our patient's values must approximate those in the analysis, or the decision analysis must provide information about the impact of variation in patient values on the results of the decision analysis. Computer models available at the bedside may broaden the scope of decision analysis applicability and permit wider use with individual patients.[70]

Second, investigators have developed numerical methods of presenting information to patients that incorporate calculated patient values.[40,71] However, these methods have not been fully tested and are not yet feasible for widespread use. Here too, computer models may be useful in the future. Third, clinicians can use "decision aids" that present descriptive and probabilistic information about the disease, treatment options, and potential outcomes.[72-73] Most commonly, these decision aids present the outcome data in terms of the percentage of people with a certain condition who do well without intervention compared with the percentage who do well with intervention. While each of these methods has considerable merit, they sometimes fall short in terms of comprehensibility, applicability, and efficiency for use in busy clinical services. Making well-validated decision aids available on the Internet could improve their clinical usefulness.

### The Likelihood of Being Helped or Harmed

Although the NNT and NNH are useful for clinicians to describe the benefits and harms of therapy, they may be less informative for individual patients who want to know their unique risk of these events. One recently developed method of expressing information to patients that incorporates patient values, can be applied to any clinical decision, and that preliminary evidence suggests may be useful in busy clinical services is the likelihood of being helped vs harmed. (S.E.S., unpublished data, 2000). The first step in this method is the exploration of patient values about receiving the treatment (vs not receiving it) and the severity of adverse events that might be caused by the treatment (vs the severity of the target event that we hope to avoid with the treatment). To answer these questions, patients are provided with brief descriptions of both the target event to be prevented and the potential adverse event from the treatment (BOX).

Following review of the description of the target event, the clinician presents the patient with a rating scale (anchored at 0 [death] and 1 [full health]) and asks him or her to mark the value of the target event.

During your discussions with patient A, you discover that she is a fiercely independent newspaper journalist who lives alone and previously cared for her father after he suffered a disabling stroke. She believes that a disabling stroke is as bad as immediate death and assigns it a value of 0. Similarly, you give your patient the descrip-

**Sample Descriptions of Stroke and Death**

A stroke can result in weakness and loss of function in one side of your body. With a disabling stroke, you are admitted to a hospital for initial treatment (which would include some rehabilitation therapy) and then transferred to a rehabilitation hospital for at least 2 months of intense rehabilitation. You regain some movement in your arm and leg but are left with a permanent weakness in that side of your body and require assistance with activities of daily living such as getting dressed, taking a bath, cooking, eating, and using a toilet. You have trouble getting the words out when you speak.

A surgical procedure called carotid endarterectomy can decrease the risk of disabling stroke but can result in death. This surgery involves repairing one of the major blood vessels in your neck that supplies blood to your brain. It must be performed by a surgeon with experience in this procedure. Death is most likely to occur in the first 30 days after this surgical procedure.

tion of the adverse event that could result from the therapy (death within 30 days of surgery) and ask her to assess this using the rating scale (she assigned a value of 0.25 since death may not necessarily be immediate). Using the 2 ratings, you infer that she believes a disabling stroke to be 1.3 times worse than death within the next month [(1-0)/1-0.25)]. This exercise should be repeated on another occasion to confirm that her values are stable.

In contrast, during your conversation with patient B, you find that he is a former truck driver who recently retired to the country with his wife so that he could be near his daughter and grandson. When you explore his values, he decides that death is 5 times worse than having a disabling stroke.

How can you now incorporate your individual patients' values into the description of therapy? The average patient with a hemispheric stroke and ipsilateral moderate carotid stenosis has a 10.3% chance of having a disabling stroke over 5 years, but this can be decreased to 5.3% with carotid endarterectomy.[4] The average NNT for such patients is 20. The absolute risk increase for death for patients having carotid endarterectomy is 1.6%,[19] which translates to an average NNH of 63 (1/0.02). You work in a hospital where the vascular surgeons have a perioperative mortality rate of 2%, and therefore you can apply this study NNH to your patients.

To calculate the likelihood of being helped vs harmed (LHH), 1/NNT (absolute risk reduction [ARR]) and 1/NNH (absolute risk increase [ARI]) are combined into an aggregate ratio. (Note that although we use 1/NNT and 1/NNH here, alternatively we could use ARR and ARI in these calculations. In a pilot study, we found that physicians made fewer errors in calculation when using NNT/NNH vs ARR/ARI, and many of the errors were in decimal placement.) For both patients, the first approximation of the LHH is LHH = (1/NNT) : (1/NNH) = (1/20) : (1/63) = 3 to 1 in favor of surgery. As a first approximation, both patients can be told that "carotid endarterectomy is 3 times as likely to help you as harm you."

However, this first approximation ignores both patients' individual risks of, and values relating to, stroke and perioperative death. You can particularize the LHH for each patient using the $f$ factors we described previously. As discussed above, women have a lower risk of stroke and the $f_t$ for patient A can be estimated at approximately 0.2.[3] This study (and a systematic review of other studies[4,5]) found that women, patients with left-sided carotid disease, and patients with a history of hypertension have increased risks of perioperative deaths (RRs, 1.4-2.3). Thus, patient A is at an increased risk of death from surgery ($f_h = 2$). Her risk-adjusted LHH is: $LHH_A = [(1/NNT) \times f_t : [(1/NNH) \times f_h] = [(1/20) \times 0.2] : [(1/63) \times 2] = 3$ to 1 in favor of medical therapy. Similarly, the LHH for patient B can be individualized for his unique risks. Men had a greater risk of stroke in the trial[1,4] and you can estimate from the reported subgroup analyses that patient B's $f_t$ is approximately 1.25. Patient B also has left-sided carotid disease, suggesting that his risk of perioperative death is increased ($f_h = 2$). His risk-adjusted LHH is: $LHH_B = [(1/20) \times 1.25] : [(1/63) \times 2] =$ 2 to 1 in favor of surgery.

These risk-adjusted LHHs still ignore each patient's values. Patient A ranked a disabling stroke as 1.3 times worse than death, and this number (the $s$ or *severity* factor) can be used to adjust the LHH as follows: $LHH_A = [(1/NNT) \times f_t \times s] : [(1/NNH) \times f_h = [(1/20) \times 0.2 \times 1.3] : [(1/63) \times 2] = 2$ to 1 in favor of medical therapy. Thus, incorporating patient A's values and unique risks of benefit and harm, she is twice as likely to be helped as harmed by medical therapy. On the other hand, patient B stated that death was 5 times worse than a stroke and incorporating this into his LHH you calculate: $LHH_B = [(1/20) \times 1.25] : [(1/63) \times 2 \times 5] = 3$ to 1 in favor of medical therapy.

These 2 cases illustrate how to incorporate your patient's values into the decision-making process. At present, this process is time-consuming and inexact, and we don't know how much difference it makes to patients or their clinical outcomes. Thus, this approach is best considered as a logical and feasible, but untested, model. Computerized versions of this approach should make it more clinically useful. If you are unsure of your patient's $f$ or if there is some uncertainty around your patient's estimate of values, you could do a sensitivity analysis (inserting different values for these variables into the above equation to see how this is reflected in the LHH). We've described a simple formulation for the LHH (ignoring other outcomes from carotid endarterectomy and the risks of the diagnostic workup),[7,6] but this could be modified for more complex situations.

## RESOLUTION OF THE SCENARIO

Before making a final decision with your patient, you need to determine what the perioperative complication rate is in your own practice setting. If we as-

sume that local surgical expertise is sufficient to apply the study results and use our patients' individual risks of benefit and harm from surgery, adjusted for their unique values, medical therapy appears to be the favored management strategy for both patients.

**Funding/Support:** Dr McAlister is a Population Health Investigator of the Alberta Heritage Foundation for Medical Research.

**Acknowledgment:** We thank Peter Rothwell, MD, Department of Clinical Neurology, Radcliffe Infirmary, Oxford, England, for providing unpublished information on reference 42.

## REFERENCES

1. Sauve JS, Laupacis A, Ostbye T, Feagan B, Sackett DL. Does this patient have a clinically important carotid bruit? *JAMA*. 1993;270:2843-2845.
2. Eliasziw M, Rankin RN, Fox AJ, Haynes RB, Barnett HJ. Accuracy and prognostic consequences of ultrasonography in identifying severe carotid artery stenosis. *Stroke*. 1995;26:1747-1752.
3. Eliasziw M, Streifler JY, Fox AJ, et al. Significance of plaque ulceration in symptomatic patients with high-grade carotid stenosis. *Stroke*. 1994;25:304-308.
4. Barnett HJ, Taylor DW, Eliasziw M, et al. Benefit of carotid endarterectomy in patients with symptomatic moderate or severe stenosis. *N Engl J Med*. 1998; 339:1415-1425.
5. Yusuf S, Held P, Teo KK. Selection of patients for randomized controlled trials: implications of wide or narrow eligibility criteria. *Stat Med*. 1990;9: 73-86.
6. Guyatt GH, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, II: how to use an article about therapy or prevention, A: are the results of the study valid? *JAMA*. 1993;270:2598-2601.
7. Guyatt GH, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, II: how to use an article about therapy or prevention, B: what were the results and will they help me in caring for my patients? *JAMA*. 1994;271:59-63.
8. Dans AL, Dans LF, Guyatt GH, Richardson S, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, XIV: how to decide on the applicability of clinical trial results to your patient. *JAMA*. 1998;279:545-549.
9. Guyatt GH, Sinclair J, Cook DJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, XVI: how to use a treatment recommendation. *JAMA*. 1999;281:1836-1843.
10. Glasziou P, Guyatt GH, Dans AL, Dans LF, Straus S, Sackett DL. Applying the results of trials and systematic reviews to individual patients (editorial). *ACP J Club*. 1998;129:A15-A16.
11. Garg R, Yusuf S, for the Collaborative Group on ACE Inhibitor Trials. Overview of randomized trials of angiotensin-converting enzyme inhibitors on mortality and morbidity in patients with heart failure. *JAMA*. 1995;273:1450-1456.
12. Gottlieb SS, McCarter RJ, Vogel RA. Effect of beta-blockade on mortality among high-risk and low-risk patients after myocardial infarction. *N Engl J Med*. 1998;339:489-497.
13. McAlister FA, Taylor L, Teo KK et al. The treatment and prevention of coronary heart disease in Canada: do older patients receive efficacious therapies? *J Am Geriatr Soc*. 1999;47:811-818.
14. Chassin MR. Appropriate use of carotid endarterectomy. *N Engl J Med*. 1998;339:1468-1471.
15. Tu JV, Hannan EL, Anderson GM, et al. The fall and rise of carotid endarterectomy in the United States and Canada. *N Engl J Med*. 1998;339:1441-1447.
16. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *N Engl J Med*. 1988;318:1728-1733.
17. Sackett DL, Haynes RB. Summarising the effects of therapy: a new table and some more terms (editorial). *ACP J Club*. 1997;127:A15-A16.
18. McQuay HJ, Moore RA. Using numerical results from systematic reviews in clinical practice. *Ann Intern Med*. 1997;126:712-720.
19. Rothwell PM, Slattery J, Warlow CP. A systematic review of the risks of stroke and death due to endarterectomy for symptomatic carotid stenosis. *Stroke*. 1996;27:260-265.
20. Sackett DL, Deeks JJ, Altman DG. Down with odds ratios! *Evidence-Based Med*. 1996;1:164-166.
21. Zhang J, Yu KF. What's the relative risk? *JAMA*. 1998;280:1690-1691.
22. Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. *Evidence-Based Medicine: How to Practice and Teach EBM*. London, England: Churchill-Livingstone; 2000.
23. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA*. 1991;266:93-98.
24. Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. *Ann Intern Med*. 1992;116: 78-84.
25. Atrial Fibrillation Investigators. Risk factors for stroke and efficacy of antithrombotic therapy in atrial fibrillation: analysis of pooled data from five randomized controlled trials. *Arch Intern Med*. 1994;154: 1449-1457.
26. Scandinavian Simvastatin Survival Study Group. Baseline serum cholesterol and treatment effect in the Scandinavian Simvastatin Survival Study (4S). *Lancet*. 1995;345:1274-1275.
27. ACE Inibitor Myocardial Infarction Collaborative Group. Indications for ACE inhibitors in the early treatment of acute myocardial infarction: systematic overview of individual data from 100000 patients in randomized trials. *Circulation*. 1998;97:2202-2212.
28. Fibrinolytic Therapy Trialists' Collaborative Group. Indications for fibrinolytic therapy in suspected acute myocardial infarction: collaborative overview of early mortality and major morbidity results from all randomised trials of more than 1000 patients. *Lancet*. 1994;343:311-322.
29. Yusuf S, Zucker D, Peduzzi P, et al. Effect of coronary artery bypass graft surgery on survival: overview of 10-year results from randomised trials by the Coronary Artery Bypass Graft Surgery Trialists Collaboration. *Lancet*. 1994;344:563-570.
30. Collins R, Peto R, MacMahon S, et al. Blood pressure, stroke, and coronary heart disease, 2: short-term reductions in blood pressure: overview of randomised drug trials in their epidemiological context. *Lancet*. 1990;335:827-838.
31. Davey Smith G, Egger M. Who benefits from medical interventions? *BMJ*. 1994;308:72-74.
32. Sharp SJ, Thompson SG, Altman DG. The relation between treatment benefit and underlying risk in meta-analysis. *BMJ*. 1996;313:735-738.
33. Rothwell PM. Can overall results of clinical trials be applied to all patients? *Lancet*. 1995;345:1616-1619.
34. Schmid CH, Lau J, McIntosh MW, Cappelleri JC. An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. *Stat Med*. 1998;17:1923-1942.
35. Ioannidis JP, Lau J. Heterogeneity of the baseline risk within patient populations of clinical trials: a proposed evaluation algorithm. *Am J Epidemiol*. 1998; 148:1117-1126.
36. North American Symptomatic Carotid Endarterectomy Trial Collaborators. Beneficial effect of carotid endarterectomy in symptomatic patients with high-grade carotid stenosis. *N Engl J Med*. 1991;325: 445-453.
37. The Executive Committee for the Asymptomatic Carotid Atherosclerosis Study. Endarterectomy for asymptomatic carotid artery stenosis. *JAMA*. 1995; 273:1421-1428.
38. Stewart LA, Clarke MJ. Practical methodology of meta-analyses (overviews) using updated individual patient data. *Stat Med*. 1995;14:2057-2079.
39. Laupacis A, Sekar N, Stiell IG. Clinical prediction rules: a review and suggested modifications of methodological standards. *JAMA*. 1997;277:488-494.
40. Glasziou PP, Irwig LM. An evidence-based approach to individualising treatment. *BMJ*. 1995;311: 1356-1359.
41. The European Carotid Surgery Trialists' Collaborative Group. Randomised trial of endarterectomy for recently symptomatic carotid stenosis: final results of the MRC European Carotid Surgery Trial (ECST). *Lancet*. 1998;351:1379-1387.
42. Rothwell PM, Warlow CP. Prediction of benefit from carotid endarterectomy in individual patients: a risk-modelling study. *Lancet*. 1999;353:2105-2110.
43. Laupacis A, Wells G, Richardson WS, Tugwell P, for the Evidence-Based Medicine Working Group. User's guides to the medical literature, V: how to use an article about prognosis. *JAMA*. 1994;272:234-237.
44. Elneihoum AM, Goransson M, Falke P, Janzon L. Three-year survival and recurrence after stroke in Malmo, Sweden: an analysis of stroke registry data. *Stroke*. 1998;29:2114-2117.
45. Rothwell PM, Slattery J, Warlow CP. Clinical and angiographic predictors of stroke and death from carotid endarterectomy: systematic review. *BMJ*. 1997; 315:1571-1577.
46. Chatellier G, Zapletal E, Lemaitre D, Menard J, Degoulet P. The number needed to treat: a clinically useful nomogram in its proper context. *BMJ*. 1996;312: 426-429.
47. Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. *BMJ*. 1995;310:452-454.
48. Grover SA, Lowensteyn I, Esrey KL, et al. Do doctors accurately assess coronary risk in their patients? preliminary results of the coronary health assessment study. *BMJ*. 1995;310:975-978.
49. Guyatt GH, Sackett DL, Sinclair JC, Hayward R, Cook DJ, Cook RJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, IX: a method for grading health care recommendations. *JAMA*. 1995;274:1800-1804.
50. Solomon NA, Glick HA, Russo CJ, Lee J, Schulman KA. Patient preferences for stroke outcomes. *Stroke*. 1994;25:1721-1725.
51. Nease RF, Kneeland T, O'Connor GT, et al. Variation in patient utilities for outcome of the management of chronic stable angina: implications for clinical practice guidelines. *JAMA*. 1995;273:1185-1190.
52. Samsa GP, Matchar DB, Goldstein L, et al. Utilities for major stroke: results from a survey of preferences among persons at increased risk for stroke. *Am Heart J*. 1998;136:703-713.
53. Sculpher M. The cost-effectiveness of preference-based treatment allocation: the case of hysterectomy versus endometrial resection in the treatment of menorrhagia. *Health Econ*. 1998;7:129-142.
54. Llewellyn-Thomas HA. Investigating patients' preferences for different treatment options. *Can J Nurs Res*. 1997;29:45-64.
55. Richardson WS, Detsky AS, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VII: how to use a clinical decision analysis, B: what are the results and will they help me in caring for my patients? *JAMA*. 1995;273:1610-1613.

**56.** Szabo E, Moody H, Hamilton T, Ang C, Kovithavongs C, Kjellstrand C. Choice of treatment improves quality of life: a study of patients undergoing dialysis. *Arch Intern Med*. 1997;157:1352-1356.
**57.** Greenfield S, Kaplan SH, Ware JE Jr, Yano EM, Frank HJ. Patients' participation in medical care: effects on blood sugar control and quality of life in diabetes. *J Gen Intern Med*. 1988;3:448-457.
**58.** Kaplan SH, Greenfield S, Ware JE Jr. Assessing the effects of physician-patient interactions on the outcomes of chronic disease. *Med Care*. 1989;27(suppl 3):S110-S127.
**59.** Schulman BA. Active patient orientation and outcomes in hypertension treatment: application of a socio-organizational perspective. *Med Care*. 1979;17:267-280.
**60.** Stewart MA. Effective physician-patient communication and health outcomes: a review. *CMAJ*. 1995; 152:1423-1433.
**61.** Vickery DM, Golaszewski TJ, Wright EC, Kalmer H. The effect of self-care interventions on the use of medical service within a Medicare population. *Med Care*. 1988;26:580-588.
**62.** Gage BF, Cardinalli AB, Owens DK. Cost-effectiveness of preference-based antithrombotic therapy for patients with nonvalvular atrial fibrillation. *Stroke*. 1998;29:1083-1091.
**63.** Strull WM, Lo B, Charles G. Do patients want to participate in medical decision making? *JAMA*. 1984; 252:2990-2994.
**64.** Degner LF, Kristjanson LJ, Bowman D, et al. Information needs and decisional preferences in women with breast cancer. *JAMA*. 1997;277:1485-1492.
**65.** Stiggelbout A, Kiebert GM. A role for the sick role. *CMAJ*. 1997;157:383-389.
**66.** Rothenbacher D, Lutz MP, Porzsolt F. Treatment decisions in palliative cancer care: patients' preferences for involvement and doctors' knowledge about it. *Eur J Cancer*. 1997;33:1184-1189.
**67.** Mazur DJ, Hickam DH. Patients' preferences for risk disclosure and role in decision making for invasive medical procedures. *J Gen Intern Med*. 1997;12: 114-117.
**68.** Margalith I, Shapiro A. Anxiety and patient participation in clinical decision-making: the case of patients with ureteral calculi. *Soc Sci Med*. 1997;45:419-427.
**69.** Torrance GW. Measurement of health state utilities for economic appraisal: a review. *J Health Econ*. 1986;5:1-30.
**70.** Lilford RJ, Pauker SG, Braunholtz A, Chard J. Decision analysis and the implementation of research findings. *BMJ*. 1998;317:405-409.
**71.** Riegelman R, Schroth WS. Adjusting the number needed to treat: incorporating adjustments for the utility and timing of benefits and harms. *Med Decis Making*. 1993;13:247-252.
**72.** Llewellyn-Thomas HA, McGreal MJ, Thiel EC, et al. Patients' willingness to enter clinical trials: measuring the association with perceived benefit and decision making preference. *Soc Sci Med*. 1991;32:35-42.
**73.** Levine MN, Gafni A, Markham B, MacFarlane D. A bedside decision instrument to elicit a patient's preference concerning adjuvant chemotherapy for breast cancer. *Ann Intern Med*. 1992;117:53-58.
**74.** O'Connor AM. Consumer/patient decision support in the new millennium: where should our research take us? *Can J Nurs Res*. 1997;29:7-12.
**75.** O'Connor AM, Rostom A, Fiset V, et al. Decision aids for patients facing health treatment or screening decisions: systematic review. *BMJ*. 1999;319:731-734.
**76.** Bain DJ, Fergie N, Quin RO, Greene M. Role of arteriography in the selection of patients for carotid endarterectomy. *Br J Surg*. 1998;85:768-770.

> Sit down before fact as a little child, be prepared to give up every preconceived notion, follow humbly wherever and to whatsoever abysses Nature leads, or you shall learn nothing.
> —Thomas Huxley (1825-1895)

# Users' Guides to the Medical Literature: XX. Integrating Research Evidence With the Care of the Individual Patient

Finlay A. McAlister; Sharon E. Straus; Gordon H. Guyatt; et al.

*JAMA*. 2000;283(21):2829-2836 (doi:10.1001/jama.283.21.2829)

http://jama.ama-assn.org/cgi/content/full/283/21/2829

# Users' Guides to the Medical Literature
## XXI. Using Electronic Health Information Resources in Evidence-Based Practice

Dereck L. Hunt, MD, MSc
Roman Jaeschke, MD, MSc
K. Ann McKibbon, MLS
for the Evidence-Based Medicine Working Group

### CLINICAL SCENARIO

You are a general internist reviewing the condition of a 55-year-old woman with type 2 diabetes mellitus and hypertension. Her glycemic control is excellent with metformin, and she has no history of complications. To manage her hypertension, she takes a small daily dose of a thiazide diuretic. During the examination, you note that her weight is stable, she has no evidence of peripheral neuropathy, and her blood pressure is 155/88 mm Hg. After arranging for glycosylated hemoglobin, cholesterol, and microalbumin assessments, you reassure your patient that she is doing well and ask her to return in 3 months. After she has left, you notice that her blood pressure over the past 6 months has been about the same as it was today. You wonder if she would benefit from more aggressive blood pressure control. Specifically, in this patient with diabetes mellitus, would tighter blood pressure control improve survival or delay the onset of complications? You decide to find if the medical literature can help resolve the issue.

Practicing evidence-based medicine involves integrating individual clinical expertise with the best available evidence from systematic research.[1] The necessary skills include formulating a concise question that addresses uncertainties in patient management and quickly identifying the highest-quality relevant information from the medical literature. The previous articles in this series have provided guides for the steps that follow identification of the best evidence—systematically assessing its validity and applicability. In this Users' Guide, we present an approach to choosing and subsequently searching the most efficient electronic resource for finding the best evidence. We have focused primarily on electronic resources as these are generally easier to search and more current than many print sources.[2] However, with the relatively recent appearance of many of the resources we recommend, little research specifically addresses their relative merits. The approaches we describe reflect our experiences and those of our colleagues working individually or with medical trainees and encompass a wide range of learning levels.

### THE CLINICAL QUESTION

The first step in the search for evidence is to identify uncertainties in patient care and formulate these into questions. Specific questions can arise when we are not sure about the benefits and risks associated with different therapeutic approaches for a well-defined group of patients or are unaware of the value of a diagnostic test or prognosis of a disease condition.[3] More general questions deal with broader topics. What therapeutic approaches are available for a given condition? What complications can develop in people who have a certain disease? While a properly defined clinical study could answer a focused clinical question, general clinical questions are too broad to be answered by a single study or meta-analysis.

#### Matching Your Question to the Best Medical Information Resource

The optimal medical information resource depends, to a large extent, on the type of question that you have and time you have available.[4] To answer focused clinical questions, the most efficient approach is to begin with a "prefiltered" evidence-based medicine resource such as *Best Evidence*, the Cochrane Library, or Clinical

**Author Affiliations:** Department of Medicine (Drs Hunt and Jaeschke and Ms McKibbon) and Health Information Research Unit (Dr Hunt and Ms McKibbon), McMaster University, Hamilton, Ontario.

The original list of members of the Evidence-Based Medicine Working Group (with affiliations) appears in the first article of this series (*JAMA*. 1993;270:2093-2095). A list of new members appears in the 10th article of the series (*JAMA*. 1996;275:1435-1439). The following members contributed to this article: Gordon Guyatt, MD, MSc, Brian Haynes, MD, PhD, Anne Holbrook, MD, PharmD, Les Irwig, MBBCh, PhD, Hui Lee, MD, MSc, Virginia Moyer, MD, MPH, and David Sackett, MD, MSc.

**Financial Disclosures:** Dr Jaeschke and Ms McKibbon are associated with the production of *Best Evidence* and *ACP Journal Club*. Dr Hunt has produced a chapter in *Clinical Evidence*.

**Corresponding Author:** Dereck L. Hunt, MD, MSc, Henderson Campus, Hamilton Health Sciences Corporation, 711 Concession St, Hamilton, Ontario, Canada L8V 1C3 (e-mail: huntdl@fhs.csu.mcmaster.ca).

**Reprints:** Gordon Guyatt, MD, MSc, Room 2C12, McMaster University, Health Sciences Centre, 1200 Main St W, Hamilton, Ontario, Canada L8N 3Z5 (e-mail: guyatt@fhs.mcmaster.ca).

**Users' Guides to the Medical Literature Section Editor:** Drummond Rennie, MD, Deputy Editor (West).

Evidence that are updated with methodologically sound and clinically important studies on a regular basis and have been designed to make searching easy. To find answers to more general medical questions, electronic versions of medical textbooks are often more helpful. UpToDate and *Scientific American Medicine* provide background information on many topics, in addition to answers to more specific questions. MEDLINE, the bibliographic database maintained by the US National Library of Medicine, can be used to find answers to both focused and background medical questions. The size and complexity of this database, however, makes searching somewhat more difficult and time consuming. We review the databases suitable for answering a specific clinical question and illustrate their use with the example of the optimal blood pressure target level in diabetic patients (TABLE).

### Using Prefiltered Evidence-Based Medicine Resources to Answer Focused Clinical Questions

*Best Evidence.* A good place to start looking for answers to focused clinical questions is *Best Evidence*. Available in CD-ROM format, this is the electronic version of 2 paper-based abstract journals: *ACP Journal Club* and *Evidence-Based Medicine*. (These journals were combined into 1 journal, *ACP Journal Club*, in North and South America in January 2000. *Evidence-Based Medicine* is still available outside the Americas.) For these publications, 150 medical journals are systematically searched on a regular basis to identify studies that are both methodologically sound and clinically relevant. By "methodologically sound" we mean that studies meet validity criteria familiar to readers of this Users' Guides series: for example, the treatment section includes only randomized trials with 80% follow-up and the diagnosis section only studies that make an independent, blind comparison of a test with a gold diagnostic standard.

*ACP Journal Club* and *Evidence-Based Medicine* present structured abstracts of studies that meet these criteria, along with an accompanying commentary by an expert who puts the study findings into clinical perspective. Clinicians can find other studies that meet methodological criteria, but have been judged less relevant, in a section of *Best Evidence* entitled "Other Articles Noted." *Best Evidence* is updated annually and now includes more than 1600 abstracted articles related to general internal medicine dating back to 1991. After 5 years, the editors review each article to make sure that it has not become outdated in light of more recent evidence. In addition to general internal medicine, *Best Evidence* includes a broader range of articles published since 1995 encompassing obstetrics and gynecology, family medicine, pediatrics, psychiatry, and surgery.

Because *Best Evidence* contains only methodologically sound articles, it is substantially smaller than many other medical literature databases and thus easier to search. To locate information on blood pressure control in people with type 2 diabetes, we used the *search* option in *Best Evidence 3*. We entered the terms *hypertension*, *diabetes*, and *mortality*, resulting in a list of 90 articles. Many of these citations, however, dealt with the prognosis of patients with diabetes and were not directly relevant for our question. We therefore returned to the search option, entered the same terms, but clicked on the *Therapeutics and Preven-*

**Table.** Medical Information Resource Contact Information

| Resource | Internet Address | Annual Cost, $ |
|---|---|---|
| *Best Evidence* | http://www.acponline.org/catalog/electronic/best_evidence.htm | 110 (CD-ROM) |
| Cochrane Library | http://www.updateusa.com/cochrane.htm | 225 |
| UpToDate | http://www.uptodate.com | 495 (CD-ROM) |
| MEDLINE | | |
| PubMED | http://www.ncbi.nlm.nih.gov/PubMed | Free |
| Internet Grateful Med | http://igm.nlm.nih.gov | Free |
| Other sources | http://www.medmatrix.org/info/medlinetable.asp | Free |
| *Scientific American Medicine* | http://www.samed.com | 245 (print and online versions)<br>(159 for online access only) |
| *Clinical Evidence* | http://www.evidence.org/index-welcome.htm | To be announced<br>(115 in print) |
| *Harrison's* Online | http://www.harrisonsonline.com | 89 |
| eMedicine | http://www.emedicine.com | Free |
| Medical Matrix | http://www.medmatrix.org/reg/login.asp | Free |
| ScHARR Netting the Evidence | http://www.shef.ac.uk/uni/academic/R-Z/scharr/ir/netting.html | Free |
| Medical World Search | http://www.mwsearch.com | Free |
| Journal listings | http://www.nthames-health.tpmde.ac.uk/connect/journals.htm<br>http://www.pslgroup.com/dg/medjournals.htm | Free<br>Free |
| Clinical practice guidelines | http://www.guidelines.gov<br>http://www.cma.ca/cpgs | Free<br>Free |
| MD Consult | http://www.mdconsult.com | 199.95 |
| EBMR Reviews (OVID) | http://www.ovid.com/products/cip/ebmr.cfm | 1275 (institutional price<br>for 1 user) |

*tion* option before asking *Best Evidence* to complete the search. This yielded a shorter list of 19 articles, all pertaining to therapy. An article entitled "Diuretics Reduced Cardiovascular Disease Events in Diabetic and Nondiabetic Patients"[5] looked promising. Double-clicking on this title produced a structured abstract indicating that diabetic participants in the Systolic Hypertension in the Elderly Program trial had a significant reduction in cardiovascular events with diuretic therapy. This interesting study did not, however, answer the question of the optimal blood pressure goal for people with diabetes.

As in this case, searching *Best Evidence* will not always be successful. This may be because high-quality evidence is not available. Alternatively, a relevant trial may have been published after the most recent edition of *Best Evidence* was released or before 1991. Well-done studies published since 1991 also may not appear in *Best Evidence* if the topic was felt to pertain more to subspecialty care than to general internal medicine. Despite these limitations, searching *Best Evidence* will often be rewarding.

**Cochrane Library.** The Cochrane Collaboration, an international organization that prepares, maintains, and disseminates systematic reviews of health care interventions, offers another electronic resource for locating high-quality information quickly. The Cochrane Library focuses primarily on systematic reviews of controlled trials of therapeutic interventions and thus provides little help in addressing other aspects of medical care, such as the value of a new diagnostic test or a patient's prognosis.

Updated quarterly, the Cochrane Library is available in CD-ROM format or over the Internet and contains 3 main sections. The first of these, the Cochrane Database of Systematic Reviews (CDSR), includes the complete reports for all of the systematic reviews that have been prepared by members of the Cochrane Collaboration (663 reviews in the fourth issue for 1999) and the protocols for Cochrane systematic reviews that are under way. A second part of the Cochrane Library, the Database of Reviews of Effectiveness (DARE) includes systematic reviews that have been published outside the collaboration: the fourth issue for 1999 included 2470 such reviews. The third section of the library, the Cochrane Controlled Trials Registry (CCTR), contains a growing list of more than 250 000 references to trials that Cochrane investigators have found by searching a wide range of sources. The sources include the MEDLINE and EMBASE (*Excerpta Medica*) bibliographic databases, hand searches, and the reference lists of potentially relevant original studies and reviews. While most citations refer to randomized trials, the database also includes a small number of observational studies. In addition to the 3 main sections, the Cochrane Library also includes information about the Cochrane Collaboration and information on how to conduct a systematic review.

To search the Cochrane Library, you can simply enter terms in the first screen that appears after selecting *search*. Alternatively, if you have access to the CD-ROM version, you can create more complex search strategies that include Medical Subject Heading (MeSH) terms and logical operators (see the section on MEDLINE, for an introduction to MeSH terms and logical operators). To find information about blood pressure control in people with diabetes, we entered the search terms *diabetes*, *hypertension*, and *mortality* using the 1999 version of the Cochrane Library (issue 4). This yielded 35 reports in the CDSR, 3 citations in the DARE, and 112 citations in the CCTR. A Cochrane review entitled "Antihypertensive Therapy in Diabetes Mellitus"[6] appeared promising. Double-clicking on this item, we found an entire Cochrane Collaboration systematic review, including information on the methods, inclusion and exclusion criteria, results, and a discussion. The results presented the findings in both textual and graphical forms. As was the case with the article found in *Best Evidence*, however, this review did not help resolve the issue of the optimal blood pressure goal for people with diabetes mellitus.

Turning to the CCTR (we double clicked on the *CCTR* option to make the citation titles appear), we found an article entitled "Effects of Intensive Blood-Pressure Lowering and Low-Dose Aspirin in Patients With Hypertension: Principal Results of the Hypertension Optimal Treatment (HOT) Randomised Trial"[7] and another entitled "Tight Blood Pressure Control and Risk of Macrovascular and Microvascular Complications in Type 2 Diabetes: UKPDS 38."[8] These were both within the first 20 citations listed in the CCTR for our search. Selecting the first of these yielded an abstract of the Hypertension Optimal Treatment (HOT) study,[7] a randomized controlled trial that compared 3 different blood pressure management strategies in persons with hypertension. Selecting the second citation produced an abstract for the UKPDS 38 study, a randomized trial enrolling persons with type 2 diabetes and hypertension and evaluating the effect of aiming for a blood pressure of less than 150/85 or 180/105 mm Hg. After an average of 8.4 years of follow-up, the tight blood pressure control arm had a 32% reduction in the risk of death related to diabetes (95% confidence interval, 6%-51%; $P = .02$).

**UpToDate.** One electronic textbook, UpToDate, is carefully updated every 4 months and is very well referenced. While UpToDate, unlike *Best Evidence* and the CDSR, does not have a set of explicit methodological quality criteria that must be met for articles to be included, it does reference many high-quality studies. To locate information on blood pressure control in people with type 2 diabetes, we entered the term *diabetes* in the search window. We found a list of 20 options and selected *diabetes mellitus, type 2*. This yielded 49 titles, including 1 entitled "Treatment of Hypertension in Diabetes." The chapter reviewed the pathogenesis and treatment of hypertension in people with diabetes. It also had a section on the "goal of blood pressure reduction"; including a detailed description of the 2 large randomized trials[7,8] that we found in the Cochrane

Library specifically addressing the clinical outcomes associated with more aggressive compared with less aggressive blood pressure management strategies. The text summarized the design and findings of these 2 studies, and we could retrieve the study abstracts by simply clicking on the references. Currently, UpToDate is available only on CD-ROM, but an Internet version is planned for late 2000.

**MEDLINE.** If a search of UpToDate, *Best Evidence*, and the Cochrane Library does not provide a satisfactory answer to a focused clinical question, it may be time to turn to MEDLINE. The US National Library of Medicine maintains this impressive bibliographic database that includes more than 9 million citations to both clinical and preclinical studies. A complementary database known as PreMEDLINE includes citations and abstracts for studies that have been published recently and have not yet been indexed. MEDLINE is an attractive database for finding medical information because of its relatively comprehensive coverage of medical journals and ready accessibility. Anyone with Internet access can search MEDLINE free of charge using PubMed or Internet Grateful Med, and most health sciences or hospital libraries provide access to MEDLINE.

These positive features are balanced with a disadvantage that relates to MEDLINE's size and the range of publications it encompasses. Searching MEDLINE effectively requires careful thought and a thorough knowledge of how the database is structured and publications are indexed. Understanding how to use MeSH terms, textword searching and exploding, and the logical operators AND and OR to combine different search results is essential. If you are unfamiliar with MEDLINE searching techniques, an article by Greenhalgh[9] presents a good introduction. Readers who suspect that they may have gaps in their searching skills should also strongly consider spending some time with an experienced medical librarian or taking a course on MEDLINE searching. Another potential source of information on searching techniques is to visit an Internet Web site designed to introduce the topic. A listing of tutorials designed to assist users of different MEDLINE systems and at different experience levels is available at http://www.docnet.org.uk/drfelix/medtut.html. More detailed information on searching MEDLINE and a number of other large bibliographic databases, including EMBASE (*Excerpta Medica*), is also available in a recently released reference book.[10] In this article, we present only the most crucial and basic MEDLINE searching advice.

MEDLINE indexers choose MeSH terms for each article. These headings provide one strategy for searching. It is important to note, however, that indexers reference articles under the most specific subject heading available (for example, *ventricular dysfunction, left*, rather than the more general term *ventricular dysfunction*). The implication of this for searching is that using a more general heading (*ventricular dysfunction*) risks missing many articles of interest. A command known as *explode* can be used to address this. Using the *explode* command identifies all articles that have been indexed using a given MeSH term as well as articles indexed using more specific terms.

Another fundamental search strategy substitutes reliance on the decisions made by MEDLINE indexers with the choices of study authors regarding terminology. Using *text word* searching makes it possible to identify all articles in which either the study title or abstract includes a certain term. Experience with MEDLINE allows clinicians to develop their preferred search strategies. Comprehensive searches will usually use both MeSH terms and text words.

To search for information pertaining to blood pressure control targets in people with type 2 diabetes, we used the National Library of Medicine's new PubMed MEDLINE searching system. We began by entering the term *diabetes mellitus* and clicking the *Go* button. This yielded a total of 139 223 citations dating back to 1966. Notice that before searching MEDLINE and PreMEDLINE, the PubMed system processed our request. Rather than simply completing a textword search, PubMed developed a more comprehensive strategy that also included the most appropriate MeSH term. To further increase the yield of citations, PubMed also automatically exploded the MeSH term. PubMed searched MEDLINE and PreMEDLINE using the strategy: *diabetes mellitus* (textword) OR *explode diabetes mellitus* (MeSH term)

The *OR* in the strategy is called a *logical operator*. It asks MEDLINE to combine the publications found using either the first search term or the second search term to make a more comprehensive list of publications in which diabetes is a topic of discussion.

We then searched using the term *hypertension* (175 063 references) and the term *mortality* (305 978 references). To combine these 3 searches, we initially clicked on the *History* button, which showed us a summary. By entering the term *#1 AND #2 AND #3* in the search window, we were able to ask PubMed to locate those citations in which diabetes mellitus, hypertension, and mortality were all addressed.

Unfortunately, the list of publications that MEDLINE identified included 1838 references, prompting us to take advantage of another searching technique designed to help identify particular types of clinical studies. *Search hedges* are systematically tested search strategies that help identify methodologically sound studies pertaining to questions of therapy, diagnosis, prognosis, or harm. A complete listing of the strategies is available, along with the sensitivities and specificities for each different approach.[11,12] While the strategies tend to be complex, many MEDLINE searching systems now have them automatically available for use. The PubMed system even has a special section with these strategies entitled *Clinical Queries*. As an alternative to the hedges, clinicians can use *single best terms* for finding higher quality studies. These terms include *clinical trial* (publication type) for treatment; *sensitivity* (text word) for diagnosis; *explode cohort studies* (MESH term) for prognosis; and *risk* (text word) for harm.

Combining our previous strategy with the term *clinical trial* (publication type) yielded a list of 108 publi-

cations. Once again, we found references to the UKPDS trial[8] and the HOT trial[7] in the citation list.

### Finding Answers to More General Questions: Textbooks and the Internet

Clinicians sometimes have general questions that are unlikely to have been answered by a single study or meta-analysis. This often occurs if they encounter a patient problem they have not seen recently and need to review the differential diagnosis, complications, or the range of therapeutic options. In these situations, prefiltered evidence-based medicine resources such as *Best Evidence* and the Cochrane Library are unlikely to be helpful. Referring to a textbook that is well referenced and updated frequently is likely to be faster and more rewarding. We have already referred to UpToDate. *Scientific American Medicine* is also updated regularly and supplies references for many statements so that you can assess how current the material is and even read the original articles. Other textbooks available in electronic formats, such as *Harrison's Principles of Internal Medicine*, can also provide valuable general background information. Additionally, new textbooks that are entirely Internet-based, such as eMedicine, are now appearing.

This brings us to the World Wide Web, which is rapidly becoming an important source of medical information. A vast number of resources can now be accessed using the Internet—some for a fee, some free-of-charge. To make these resources more accessible, certain Web sites have been specifically designed to provide links to medical information locations or to facilitate searching for medical information on the Internet. Examples of such Web sites include Medical Matrix, ScHARR, and Medical World Search (Table). The Internet can also be used to access medical journals as well as clinical practice guidelines. We must, however, issue a "user beware" caveat: some of these guidelines may fail to meet Users' Guides criteria for evidence-based guidelines.[13,14] An example of a site that provides access to many resources, including journals, textbooks, and guidelines, albeit for a fee, is MD Consult. Lastly, Web sites produced and maintained by reputable organizations such as the American Cancer Society (http://www.cancer.org) or the American Diabetes Association (http://www.diabetes.org) provide another approach for finding information.

## RESOLUTION OF THE SCENARIO

Finding the articles that addressed your clinical question required 5 to 30 minutes, depending on the resource used.† A full assessment of the validity and applicability required an additional half hour. The UKPDS study[8] is the closest match to your patient and her clinical situation. The study shows a clear reduction of diabetes-related mortality with tight blood pressure control in persons with type 2 diabetes mellitus and hypertension. You decide to initiate treatment with an angiotensin-converting enzyme inhibitor at her next appointment with the goal of lowering her blood pressure.

## CONCLUSION

The health sciences literature is enormous and continues to expand rapidly. To the extent that this reflects ongoing research and identification of potential improvements for patient care, this expansion is very promising. At the same time, however, it makes the task of locating the best and most current therapy or diagnostic test more challenging. The emergence of new information products specifically designed to provide ready access to high-quality, clinically relevant, and up-to-date information is thus timely and encouraging. An additional electronic product we are looking forward to in 2000 is Clinical Evidence, produced by the BMJ Publishing Group and American College of Physicians–American Society of Internal Medicine. It is a growing compendium of evidence pertaining to treatments of specific conditions. Also, electronic resources that facilitate simultaneous searching of MEDLINE, *Best Evidence*, and the Cochrane Library are now available through services such as OVID Technology's Evidence-Based Medicine Reviews. Many health sciences libraries subscribe to this service and individual subscriptions can be started. Active research and development continues for integrated products. Among the challenges for staying up-to-date, clinicians can therefore add the task of keeping current their knowledge of optimal search strategies and resources.

**Acknowledgment:** Basit Chaudray, MD, and Sharon Strauss, MD, provided helpful comments on an earlier draft of the manuscript. Deborah Maddock coordinated the activities of the EBM Working Group that led to the production of this article.

### REFERENCES

1. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. *BMJ*. 1996;312:71-72.
2. McKibbon KA, Richardson WS, Walker Dilks C. Finding answers to well-built clinical questions. *Evidence-Based Med*. 1999;6:164-167.
3. Richardson WS, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: a key to evidence-based decisions [editorial]. *ACP J Club*. 1995; 123:A12-A13.
4. Sackett DL, Straus SE. Finding and applying evidence during clinical rounds: the "evidence cart." *JAMA*. 1998;280:1336-1338.
5. Diuretics reduced cardiovascular disease events in diabetic and nondiabetic patients [abstract]. *ACP J Club*. 1997;126:57.
6. Fuller J, Stevens LK, Chaturvedi N, Holloway JF. Antihypertensive therapy in diabetes mellitus (Cochrane Review). The Cochrane Library [serial on CD-ROM]. 1999;4.
7. Hansson L, Zanchetti A, Carruthers SG, et al, for the HOT Study Group. Effects of intensive blood-pressure lowering and low-dose aspirin in patients with hypertension: principal results of the Hypertension Optimal Treatment (HOT) randomised trial. *Lancet*. 1998; 351:1755-1762.
8. UK Prospective Diabetes Study Group. Tight blood pressure control and risk of macrovascular and microvascular complications in type 2 diabetes: UKPDS 38. *BMJ*. 1998;317:703-713.
9. Greenhalgh T. How to read a paper: the Medline database. *BMJ*. 1997;315:180-183.
10. McKibbon A, Eady A, Marks S. *PDQ: Evidence-Based Principles and Practice*. Hamilton, Ontario: BC Decker; 1999.
11. Haynes RB, Wilczynski N, McKibbon KA, et al. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. *J Am Med Inform Assoc*. 1994;1:447-458.
12. Wilczynski NL, Walker CJ, McKibbon KA, Haynes RB. Assessment of methodological search filters in MEDLINE. *Proc Annu Symp Comput Appl Med Care*. 1994;17:601-605.
13. Hayward R, Wilson MC, Tunis SR, Bass EB, Guyatt GH, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VIII: how to use clinical practice guidelines, A: are the recommendations valid? *JAMA*. 1995;274:70-74.
14. Wilson MC, Hayward R, Tunis SR, Bass EB, Guyatt GH, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VIII: how to use clinical practice guidelines, B: what are the recommendations and will they help me in caring for my patients? *JAMA*.1995;274:1630-1632.

# Users' Guides to the Medical Literature: XXI. Using Electronic Health Information Resources in Evidence-Based Practice

Dereck L. Hunt; Roman Jaeschke; K. Ann McKibbon; et al.

*JAMA*. 2000;283(14):1875-1879 (doi:10.1001/jama.283.14.1875)

http://jama.ama-assn.org/cgi/content/full/283/14/1875

# Users' Guides to the Medical Literature

## XXII: How to Use Articles About Clinical Decision Rules

Thomas G. McGinn, MD
Gordon H. Guyatt, MD
Peter C. Wyer, MD
C. David Naylor, MD
Ian G. Stiell, MD
W. Scott Richardson, MD
for the Evidence-Based Medicine Working Group

**Clinical experience provides clinicians with an intuitive sense of which findings on history, physical examination, and investigation are critical in making an accurate diagnosis, or an accurate assessment of a patient's fate. A clinical decision rule (CDR) is a clinical tool that quantifies the individual contributions that various components of the history, physical examination, and basic laboratory results make toward the diagnosis, prognosis, or likely response to treatment in a patient. Clinical decision rules attempt to formally test, simplify, and increase the accuracy of clinicians' diagnostic and prognostic assessments. Existing CDRs guide clinicians, establish pretest probability, provide screening tests for common problems, and estimate risk. Three steps are involved in the development and testing of a CDR: creation of the rule, testing or validating the rule, and assessing the impact of the rule on clinical behavior. Clinicians evaluating CDRs for possible clinical use should assess the following components: the method of derivation; the validation of the CDR to ensure that its repeated use leads to the same results; and its predictive power. We consider CDRs that have been validated in a new clinical setting to be level 1 CDRs and most appropriate for implementation. Level 1 CDRs have the potential to inform clinical judgment, to change clinical behavior, and to reduce unnecessary costs, while maintaining quality of care and patient satisfaction.**

*JAMA. 2000;284:79-84* www.jama.com

### CLINICAL SCENARIO

You are the medical director of a busy inner-city emergency department. Faced with a limited budget and pressure to improve efficiency, you have conducted an audit of radiological procedures ordered for minor trauma and found a high rate of x-rays ordered for ankle and knee trauma. You are aware of the Ottawa ankle rules (FIGURE 1) that identify patients for whom ankle radiographs can be omitted without adverse consequences. In addition, you are aware that a small number of faculty and residents currently rely on these models to make quick frontline decisions in the emergency department.

You are interested in knowing the accuracy of the rules, whether they are applicable to your patient population, and whether you should be implementing the rules in your own practice. Furthermore, you wonder if implementing the rules can change clinical behavior and reduce costs without compromising quality care. You decide to consult the original medical literature and to assess the evidence for yourself.

### THE SEARCH

Currently, *decision rules* have no separate medical subject heading (MeSH) in the National Library of Medicine MEDLINE database. You therefore search PubMed under the MeSH heading *ankle fractures* and add the text words *rules* and *decision rules*. This search yields 5 citations, of which 3 deal directly with the Ottawa clinical decision rules for ankle fractures.[1-3]

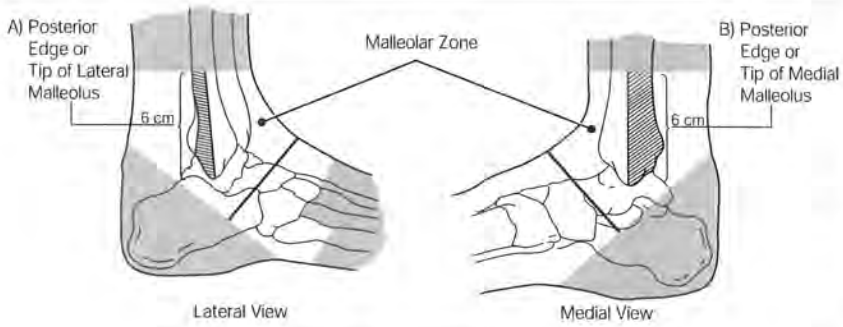
In reviewing these articles and deciding whether to implement changes in your emergency department, you require criteria for determining the strength of the inference you can make about the accuracy and impact of the Ottawa ankle rules. This article will provide you with the tools to answer those questions.

### CLINICAL DECISION RULES

Establishing patients' diagnosis and prognosis are closely linked activities central to every physician's practice. The diagnoses we make and our assessment of patients' prognosis often determine the recommendations we make to our patients. Clinical experience provides us with an intuitive sense of which findings on history, physical examination, and investigation are critical in making an accurate diagnosis or an accurate assessment of our patients' con-

**Author Affiliations:** The original list of members (with affiliations) appears in the first article of this series (*JAMA*. 1993;270:2093-2095). A list of new members appears in the 10th article of the series (*JAMA*. 1996;275:1435-1439). A full list of the EBM Working Group members, including institutional affiliations and career awards, was presented in the Introduction to this series and in Users' Guide X. The following members contributed to this article: Deborah Cook, MD, Roman Jaeschke, MD, Thomas Newman, MD, Jim Nishikawa, MD, Mark Wilson, MD.
**Corresponding Author:** Thomas G. McGinn, MD, Adult Primary Care, Mount Sinai Medical Center, One Gustave Levy Place, New York, NY 10029-6574 (e-mail: thomas.mcginn@mountsinai.org).
**Reprints:** Gordon Guyatt, MD, Room 2C12, McMaster University, Health Sciences Centre, 1200 Main St W, Hamilton, Ontario, Canada, L8N 3Z5.
**Users' Guides to the Medical Literature Section Editor:** Drummond Rennie, MD, Deputy Editor.

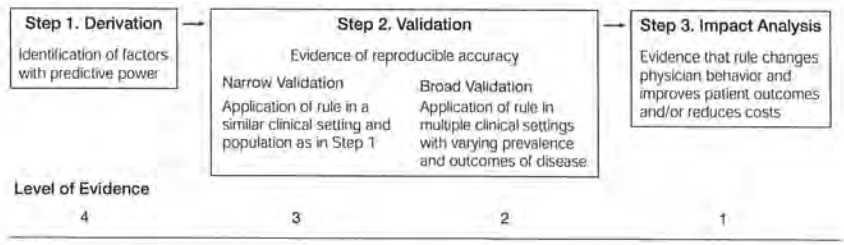
**Figure 1.** Ottawa Ankle Rules




An ankle x-ray series is required only if there is any pain in the malleolar zone and any of these findings:

1. Bone tenderness at A
or
2. Bone tenderness at B
or
3. Inability to bear weight both immediately and in emergency department

**Figure 2.** Development of a Clinical Decision Rule




dition. While often extraordinarily accurate, this intuition may sometimes be misleading.

A clinical decision rule (CDR) can be defined as a clinical tool that quantifies the individual contributions that various components of the history, physical examination, and basic laboratory results make toward the diagnosis, prognosis, or likely response to treatment in an individual patient.[4] Clinical decision rules attempt to formally test, simplify, and increase the accuracy of clinicians' diagnostic and prognostic assessments and are most likely to be useful in situations where decision making is complex, the clinical stakes are high, or there are opportunities to achieve cost savings without compromising patient care. Available CDRs include guides for whether to treat sore throats[5] and for establishing a pretest probability of pulmonary embolus.[6] Other CDRs provide screening tests for common problems that frequently go undetected, including alcoholism[7] and depression.[8] Another category of CDRs help estimate risk, such as the risk of developing delirium in hospitalized patients[9] or the risk of bleeding while receiving anticoagulation therapy.[10]

Developing and testing a CDR involves 3 steps: creating or deriving the rule, testing or validating the rule, and assessing the impact of the rule on clinical behavior (impact analysis). The validation process may require several studies to fully test the accuracy of the rule at different clinical sites (FIGURE 2). Each step in the development of a CDR may be published separately by different authors, or all 3 steps may be included in a single article. TABLE 1 presents a hierarchy that can guide clinicians in assessing the full range of evidence supporting use of a CDR in their practice.

We note that our hierarchy applies only to CDRs intended for application in clinical practice. Investigators may use identical methodology to generate equations that stratify patients into different risk groups for nonclinical purposes. For example, investigators can use such equations for statistical adjustment in studies involving large databases. These rules, which are not so clinical, do not involve application by front-line practitioners, and thus require a somewhat different hierarchy of strength of evidence.

We will now review the steps in the development and testing of a CDR. We will relate each stage of the process to the hierarchy presented in Table 1. Although we will address issues of interest to investigators engaged in developing CDRs, we do so only for the purpose of equipping our clinician readers with the knowledge and tools they need to evaluate existing CDRs for application to clinical practice.

## Developing a Clinical Decision Rule

Our search found 3 articles related to the Ottawa ankle rules, the first of which described the CDR derivation.[1] Investigators who develop a CDR begin by constructing a list of potential predictors of the outcome of interest, in this case, radiological ankle fractures. The list typically includes items from the history, physical examination, and basic laboratory tests. The investigators then examine a group of patients and determine if the candidate clinical predictors are present and the patient's status on the outcome of interest, in this case, the result of the ankle radiograph. Statistical analysis reveals which predictors are most powerful and which predictors can be omitted from the rule without loss of predictive power. Typically, the statistical techniques used in this process are based on logistic regression; readers can find

a clinician-friendly description of these methods in another article.[11] Other techniques that investigators sometimes use include discriminant analysis,[12] which produces equations similar to regression analysis; recursive partitioning analysis, which builds a tree in which the patient populations are split into smaller and smaller categories based on risk factors[13]; and neural networks.[14]

Clinical decision rules that investigators have derived, but not validated, should not be considered ready for clinical application (Table 1). Investigators interested in performing the validation of a CDR, however, need criteria to judge whether investigators have conducted a rigorous derivation process and, thus, whether the rule is promising enough to move forward to the validation phase. A list of important criteria for derivation is provided in TABLE 2. Interested readers can find a complete discussion on the derivation process and these criteria in an article by Laupacis et al.[4]

**Table 1.** Hierarchy of Evidence for Clinical Decision Rules

| |
|---|
| Level 1: Rules that can be used in a wide variety of settings with confidence that they can change clinician behavior and improve patient outcomes |
| At least 1 prospective validation in a different population and 1 impact analysis, demonstrating change in clinician behavior with beneficial consequences |
| Level 2: Rules that can be used in various settings with confidence in their accuracy |
| Demonstrated accuracy in either 1 large prospective study including a broad spectrum of patients and clinicians or validated in several smaller settings that differ from one another |
| Level 3: Rules that clinicians may consider using with caution and only if patients in the study are similar to those in the clinician's clinical setting |
| Validated in only 1 narrow prospective sample |
| Level 4: Rules that need further evaluation before they can be applied clinically |
| Derived but not validated or validated only in split samples, large retrospective databases, or by statistical techniques |

*Adapted, with permission, from Mount Sinai Department of Medicine Evidence-Based Medicine Homepage (http://med.mssm.edu/ebm/).

**Table 2.** Methodological Standards for Derivation of a Clinical Decision Rule

1. Were all important predictors included in the derivation process?
2. Were all important predictors present in a significant proportion of the study population?
3. Were all the outcome events and predictors clearly defined?
4. Were those assessing the outcome event blinded to the presence of the predictors and those assessing the presence of predictors blinded to the outcome event?
5. Was the sample size adequate (including adequate number of outcome events)?
6. Does the rule make clinical sense?

### Validation

There are 3 reasons why even rigorously derived CDRs are not ready for application in clinical practice without further validation. First, CDRs may reflect associations between given predictors and outcomes that are due primarily to chance. If that is so, a different set of predictors will emerge in a different group of patients, even if the patients come from the same setting. Second, predictors may be idiosyncratic to the population, to the clinicians using the rule, or to other aspects of the design of individual studies. If that is so, the rule may fail in a new setting. Perhaps most important, clinicians may, because of problems in the feasibility of rule application in the clinical setting, fail to implement a rule comprehensively or accurately. The result would be that a rule succeeds in theory but fails in practice.

Statistical methods can deal with the first of these problems. For instance, investigators may split their population into 2 groups and use one to develop the rule and the other to test it. Alternatively, they may use more sophisticated statistical methods built on the same logic. Conceptually, these approaches involve removing 1 patient from the sample, generating the rule using the remainder of the patients, and testing it on the patient who was removed from the sample. This procedure, sometimes referred to as a bootstrap technique, is repeated in sequence for every patient being studied.

While statistical validation within the same setting or group of subjects reduces the likelihood that the rule reflects the play of chance rather than true associations, it fails to address the other 2 threats to validity. The success of the CDR may be peculiar to the particular populations of patients and clinicians involved in the derivation study. Even if this is not so, clinicians may have difficulties using the rule in practice, difficulties that compromise its predictive power. Thus, to graduate from level 4, studies must involve clinicians actually using the rule in practice.

A CDR developed to predict serious outcomes (eg, heart failure and ventricular arrhythmia) in syncope patients highlights the importance of validation.[15] Investigators derived the rule using data from 252 patients who presented to the emergency department and then attempted to prospectively validate it in a sample of 374 patients. The CDR gave individuals a score from 0 to 4, depending on the number of clinical predictors present. The probability of poor outcomes corresponding to almost every score in the derivation set was approximately twice that of the validation. For example, in the derivation set the risk of a poor outcome in a patient with a score on the CDR of 3 was estimated to be 52%; a patient with the same score in the validation set had a probability of a poor outcome of only 27%. This variation in results may have been caused by a difference in the severity of the syncope cases entered into the 2 studies or to different criteria for generating a score of 3. Because of the risk that it will provide misleading information when applied in a real-world clinical setting, we situate a CDR that has undergone development without validation as level 4 on our hierarchy (Table 1).

Despite this major limitation, clinicians can still extract clinically relevant messages from an article describing the development of a CDR. They may wish to note the most important predictors and consider them more carefully in their own practice. They may also consider giving less importance to variables that failed to show predictive power. For instance, in developing a CDR to predict mortality from pneumonia, the investigators found that white

**Table 3.** Methodological Standards for Validation of a Clinical Decision Rule

1. Were the patients chosen in an unbiased fashion and do they represent a wide spectrum of severity of disease?
2. Was there a blinded assessment of the criterion standard for all patients?
3. Was there an explicit and accurate interpretation of the predictor variables and the actual rule without knowledge of the outcome?
4. Was there 100% follow up of those enrolled?

blood cell count had no bearing on subsequent mortality.[16] This being the case, clinicians maywish to put less weight on white blood cell count when making decisions about admitting pneumonia patients to the hospital.

To move up the hierarchy, CDRs must provide additional evidence of validity. The second article found in our search described the refinement and prospective validation of the Ottawa ankle rules.[2] Validation of a CDR involves demonstrating that its repeated application as part of the process of clinical care leads to the same results. Ideally, a validation entails the investigators applying the rule prospectively in a new population with a different prevalence and spectrum of disease from that of the patients in whom the rule was derived. One key issue is to be sure that the CDR performs similarly in a variety of populations and in the hands of a variety of clinicians working in a variety of institutions. A second issue is to be sure that the CDR works well when clinicians are applying it consciously as a rule, as opposed to a purely statistical validation.

If the setting in which the CDR was originally developed was limited and its validation has been confined to this setting, application by clinicians working in other settings is less secure. Validation in a similar setting can take a number of forms. Most simply, after developing the CDR, the investigators return to their population, draw a new sample of patients, and test the rule's performance. Thus, we classify rules that have been validated in the same, or very similar limited or narrow populations, to the sample used in the development as level 3 on our hierarchy and recommend clinicians use the results cautiously (Table 1).

If investigators draw patients in the derivation phase from a sufficiently heterogeneous population across a variety of institutions, testing the rule in the same population provides strong validation. Validation in a new population provides the clinician with strong inferences about the usefulness of the rule, corresponding to level 2 in our hierarchy (Table 1).

The Ottawa ankle rules were first derived in 2 large university-based emergency departments in Ottawa[1] and were then prospectively validated in a large sample of patients from the same emergency departments.[2] At this stage, the rules would be classified as level 2 in our hierarchy because of the large number and diversity of patients and physicians involved in the study. Since that initial validation, the rules have been validated in several different clinical sites with relatively consistent results.[17-20] This evidence even further strengthens our inference about their predictive power.

Many CDRs are derived and then validated in a small, narrowly selected group of patients (level 3). One such rule was derived to predict preserved left ventricular function after a myocardial infarction.[21] The initial derivation relied on data from 314 patients admitted to 1 tertiary care center. The investigators derived the rule using data from 162 patients and then performed a validation in 152 patients in the same setting. Of those whom the CDR identified as having preserved ejection fraction, 99% indeed had preserved left ventricular function. At this stage, we would consider the rule had met criteria for level 3, and its use should be restricted to settings similar to the validation study, ie, similar coronary care unit settings.

Investigators further validated the CDR for preserved left ventricular function, in 2 larger trials, one that enrolled 213 patients[22] from a single site and a larger trial that enrolled 1891 patients from several different institutions.[23] In both studies, of those patients predicted to have preserved ventricular function (ejection fraction >40%), 86% actually had preserved ventricular function. This drop in predictive value changes the implications of applying the rule in clinical practice. At this point in development, the rule would be considered level 2, meaning that the rule can used in clinical settings with a high degree of confidence but with the adjusted values. The development of this rule highlights the importance of the validation of a rule in a diverse patient population before broadly applying it in clinical settings.

Whether or not investigators have conducted their validation study in a similar, narrow (level 3) population or a broad, heterogeneous (level 2) population, their results allow stronger inferences if they have adhered to the methodological standards listed in TABLE 3. First, were the patients chosen in an unbiased fashion, and do they represent a wide spectrum of severity of disease? Second, was there a blinded assessment of the criterion standard for all patients? Third, was there an explicit and accurate interpretation of the predictor variables and actual rule without knowledge of the outcome? If those evaluating predictor status of study patients are aware of the outcome or if those assessing the outcome are aware of patients' status with respect to the predictors, their assessments may be biased. For instance, in a CDR developed to predict the presence of pneumonia in patients presenting with cough,[24] the authors make no mention of blinding during either the derivation or the validation process. Knowledge of history or physical examination findings may have influenced the judgements of the unblinded radiologists. Lastly, investigators should achieve close to 100% follow-up of those they enrolled. Interested readers can find a complete discussion of the validation process and these criteria in an article by Laupacis et al.[4]

The investigators testing the Ottawa ankle rules enrolled consecutive patients, obtained radiographs for all of them, and ensured that not only were the clinicians assessing the clinical pre-

dictors unaware of the radiographic results but that the radiologists had no knowledge of the clinical data.

### Interpreting the Results

Whatever the level of evidence associated with a CDR, its usefulness will depend on its predictive power. Investigators may report their results in a variety of ways. The ankle component of the Ottawa ankle rules states that an ankle x-ray series is only indicated for patients with pain near the malleoli and either inability to bear weight or localized bone tenderness at the posterior edge or tip of either malleolus (Figure 1). The developers calculated the sensitivity and specificity of their rule as a diagnostic test using this criterion. In the development process, all patients with fracture had a positive result (sensitivity of 100%), but only 40% of those without fractures had a negative result (specificity of 40%). These results suggest that if clinicians order radiographs only in those patients with a positive result they will not miss any fractures and will avoid the test in 40% of those without a fracture.

The validation study confirmed these results; in particular, the test maintained a sensitivity of 100%. This is reassuring, and more so because the sample size was sufficiently large to result in a relatively narrow confidence interval (CI) (95% CIs, 93%-100%). Thus, clinicians adopting the rule would miss very few, if any, fractures.

Another way of reporting CDR results is in terms of probability of the target condition being present given a particular CDR result. For example, a recent CDR for pulmonary embolus derived by Wells and colleagues[6] placed patients into low (3.4%; 95% CI, 2.2%-5%), intermediate (28%; 95% CI, 23.4%-32.2%), or high probability (78%; 95% CI, 69.2%-86.0%) categories. When investigators report CDR results in this fashion, they are implicitly incorporating all clinical information. In doing so, they remove any need for clinicians to consider independent information in deciding on the likelihood of the diagnosis or a patient's prognosis.

Finally, CDRs may also report their results as likelihood ratios (LRs) or as absolute or relative risks. For example the CAGE, a CDR for detecting alcoholism, has been reported as LRs (eg, for CAGE scores of 0/4, LR=0.14; for 1/4, LR=1.5; for 2/4, LR=4.5; for 3/4, LR=13; and for 4/4, LR=100). In this example, the probability of disease, alcoholism, depends on the combination of the prevalence of disease in the community and the score on the CAGE CDR.[7] When investigators report their results as LRs, they are implicitly suggesting that clinicians should use other, independent information to generate a pretest (or prerule) probability. They can then use the LRs generated by the rule to establish a posttest probability. Clinicians can find approaches to using LRs in clinical practice in a previous Users' Guide.[25]

### Impact Analysis

Use of a CDR involves remembering predictor variables and often entails making calculations to determine a patient's probability of having the CDR's target outcome. Pocket cards and computer algorithms can facilitate the task of using complex CDRs. Nonetheless, CDRs demand clinician time and energy, and their use is warranted only if they change physician behavior and if that behavior change results in improved patient outcomes or reduced costs while maintaining quality of care. If these conditions are not met, whatever the accuracy of a CDR, attempts to use it systematically will be a waste of time.

There are a number of reasons why an accurate CDR may not produce a change in behavior or an improvement in outcomes. First, clinicians' intuitive estimation of probabilities may be as good as, if not better than, the CDR. If this is so, CDR information will not improve their practice. Second, the calculations involved may be cumbersome, and clinicians may, as a result, not use the rule. Finally, there may be practical barriers to acting on the results of the CDR. For instance, in the case of the Ottawa ankle rules, clinicians may be sufficiently concerned about protecting themselves against litigation that they order radiographs despite a CDR result suggesting a negligible probability of fracture.

These are the considerations that lead us to classify a CDR with evidence of reproducible accuracy in diverse populations as level 2 and insist on a positive result from a study of impact before a CDR graduates to level 1.

Ideally, an impact study would randomize patients, or larger administrative units, to the application or nonapplication of the CDR and follow up patients for all relevant outcomes (including quality of life, morbidity, and resource utilization). Randomization of individual patients is unlikely to be appropriate because one would expect the participating clinicians to incorporate the rule into the care of all their patients. A suitable alternative is to randomize institutions or practice settings and conduct analyses appropriate to these larger units of randomization. Another potential design is to look at a group before and after clinicians began to use the CDR and compare that with a control group in which there has been no intervention.

Investigators examining the impact of the Ottawa ankle rules randomized 6 emergency departments to use or not use their CDR.[3] Prior to initiating the study, 1 center dropped out, leaving a total of 5 emergency departments, 2 in the intervention group and 3 in the usual care group. The intervention consisted of introducing the CDR at a general meeting, distributing pocket cards summarizing the rules, posting the rule throughout the emergency department, and applying preprinted data collection forms to each chart. In the control group, the only intervention was the introduction of preprinted data collection forms without the Ottawa ankle rules attached to each chart.

A total of 1911 eligible patients entered the study: 1005 in the control group and 906 in the intervention group. There were 691 radiographs requested in the intervention group and 996 in the control group. In an analysis that focused on the ordering physician, the in-

vestigators found that the mean proportion of patients referred for radiography was 99.6% in the control group and 78.9% in the intervention group ($P = .03$). The investigators noted 3 missed fractures in the intervention group, none of which led to adverse outcomes. Thus, the investigators demonstrated a positive resource utilization impact of the Ottawa ankle rules (decreased test ordering) without increase in adverse outcomes, moving the CDR to level 1 in the hierarchy (Table 1).

## RESOLUTION OF THE SCENARIO

You have found level 1 evidence supporting the use of the Ottawa ankle rules in reducing unnecessary ankle radiographs in patients presenting to the emergency department with ankle injuries. You therefore feel confident that you can productively use the rule in your own practice. However, another recent study makes you aware that changing the behavior of your colleagues to realize the possible reductions in cost may be a challenge: Cameron and Naylor[26] reported on an initiative in which clinicians expert in the use of the Ottawa ankle rules trained 16 other individuals to teach the use of the rules. These individuals returned to their emergency departments armed with slides, overheads, a 13-minute instructional video, and a mandate to train their colleagues locally and regionally in the use of the rules.

Unfortunately this program led to no change in the use of ankle radiography. The results demonstrate that even the availability of a level 1 CDR may require local implementation strategies with known effectiveness in changing provider behavior to ensure implementation.[27-29] Among the possible strategies, which are most likely to be effective if used as part of a package of interventions, include computer reminders, mobilization of local opinion leaders, one-to-one conversations with a respected information source (academic detailing), and audit and feedback.

## CONCLUSION

Clinical decision rules inform our clinical judgment and have the potential to change clinical behavior and reduce unnecessary costs while maintaining quality of care and patient satisfaction. The challenge for clinicians is to evaluate the strength of the rule and its likely impact and to find ways of efficiently incorporating level 1 rules into their daily practice.

A summary of some frequently used CDRs, evaluated in an evidence-based fashion (ie, highlighting the level of evidence), is currently available on the Internet for clinician use (http://med.mssm.edu/ebm).

**Acknowledgment:** We thank Deborah Maddock, McMaster University, for her superb coordination of the Users' Guide project. Dr McGinn would like to thank Gerald Paccione, MD, and the internal medicine residents at Montefiore Medical Center for their input in the area of CDRs over the years and Roseanne Leipzig, MD, Mount Sinai Medical Center, for her input to the manuscript.

### REFERENCES

1. Stiell IG, Greenberg GH, McKnight RD, Nair RC, McDowall I, Worthington JR. A study to develop clinical decision rules for the use of radiography in acute ankle injuries. *Ann Emerg Med*. 1992;21:384-390.
2. Stiell IG, Greenberg GH, McKnight RD, et al. Decision rules for the use of radiography in acute ankle injuries: refinement and prospective validation. *JAMA*. 1993;269:1127-1132.
3. Auleley G, Ravaud P, Giraudeau B, et al. Implementation of the Ottawa ankle rules in France: a multicenter randomized controlled trial. *JAMA*. 1997;277:1935-1939.
4. Laupacis A, Sekar N, Stiell I. Clinical prediction rules: a review and suggested modifications of methodological standards. *JAMA*. 1997;277:488-494.
5. Walsh BT, Bookheim WW, Johnson RC, Tompkins RK. Recognition of streptococcal pharyngitis in adults. *Arch Intern Med*. 1975;135:1493-1497.
6. Wells PS, Ginsberg JS, Anderson DR, et al. Use of a clinical model for safe management of patients with suspected pulmonary embolism. *Ann Intern Med*. 1998;129:997-1005.
7. Buchsbaum DG, Buchanan RG, Centor RM, Schnoll SH, Lawton MJ. Screening for alcohol abuse using CAGE scores and likelihood ratios. *Ann Intern Med*. 1991;115:774-777.
8. Whooley MA, Avins AL, Miranda J, Browner WS. Case-finding instruments for depression: two questions are as good as many. *J Gen Intern Med*. 1997;12:439-445.
9. Inouye SK, Viscoli CM, Horwitz RI, Hurst LD, Tinetti ME. A predictive model for delirium in hospitalized elderly medical patients based on admission characteristics. *Ann Intern Med*. 1993;119:474-481.
10. Landefeld CS, Goldman L. Major bleeding in outpatients treated with warfarin: incidence and prediction by factors known at the start of outpatient therapy. *Am J Med*. 1989;87:144-152.
11. Guyatt GH, Walter S, Shannon H, Cook D, Jaeschke R, Heddle H. Basic statistics for clinicians, 4: correlation and regression. *CMAJ*. 1995;152:497-504.
12. Rudy TE, Kubinski JA, Boston JR. Multivariate analysis and repeated measurements: a primer. *J Crit Care*. 1992;7:30-41.
13. Cook EF, Goldman L. Empiric comparison of multivariate analytic techniques: advantages and disadvantages of recursive partitioning analysis. *J Chronic Dis*. 1984;39:721-731.
14. Baxt WG. Application of artificial neural networks to clinical medicine. *Lancet*. 1995;346:1135-1138.
15. Martin TP, Hanusa BH, Kapoor WN. Risk stratification of patients with syncope. *Ann Emerg Med*. 1997;29:459-466.
16. Fine MJ, Auble TE, Yealy DE, et al. A prediction rule to identify low-risk patients with community-acquired pneumonia. *N Engl J Med*. 1997;336:243-250.
17. Lucchesi GM, Jackson RE, Peacock WF, Cerasani C, Swor RA. Sensitivity of the Ottawa ankle rules. *Ann Emerg Med*. 1995;26:1-5.
18. Kerr L, Kelly A, Grant J, et al. Failed validation of a clinical decision rule for the use of radiography in acute ankle injury. *N Z Med J*. 1994;107:294-295.
19. Stiell I, Wells G, Laupacis A, et al. Multicenter trial to introduce the Ottawa ankle rules for use of radiography in acute ankle injuries. *BMJ*. 1995;311:594-597.
20. Auleley G, Kerboull L, Durieux P, Cosquer M, Courpied J, Ravaud P. Validation of the Ottawa ankle rules in France: a study in the surgical emergency department of a teaching hospital. *Ann Emerg Med*. 1998;32:14-18.
21. Silver MT, Rose GA, Paul SD, O'Donnell CJ, O'Gara PT, Eagle KA. A clinical rule to predict preserved left ventricular ejection fraction in patients after myocardial infarction. *Ann Intern Med*. 1994;12:750-756.
22. Tobin K, Stomel R, Harber D, Karavite D, Sievers J, Eagle K. Validation in a community hospital setting of a clinical rule to predict preserved left ventricular ejection fraction in patients after myocardial infarction. *Arch Intern Med*. 1999;159:353-357.
23. Krumholz HM, Howes CJ, Murillo JE, Vaccarino LV, Radford MJ, Ellerbeck EF. Validation of a clinical prediction rule for left ventricular ejection fraction after myocardial infarction in patients > or = to 65 years old. *Am J Cardiol*. 1997;80:11-15.
24. Heckerling PS, Tape TG, Wigton RS, et al. Clinical prediction rule for pulmonary infiltrates. *Ann Intern Med*. 1990;113:664-670.
25. Jaeschke R, Guyatt GH, Sackett DL, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, III: how to use an article about a diagnostic test, B: what are the results and will they help me in caring for my patients? *JAMA*. 1994;271:703-707.
26. Cameron C, Naylor CD. No impact from active dissemination of the Ottawa Ankle Rules: further evidence of the need for local implementation of practice guidelines. *CMAJ*. 1999;160:1165-1168.
27. Davis DA, Thomson MA, Oxman AD, Haynes RB. Changing physician performance: a systematic review of the effect of continuing medical education strategies. *JAMA*. 1995;274:700-705.
28. Cabana MD, Rand CS, Powe NR, et al. Why don't physicians follow clinical practice guidelines? a framework for improvement. *JAMA*. 1999;282:1458-1465.
29. Davis D, O'Brien MA, Freemantle N, Wolf FM, Mazmanian P, Taylor-Vaisey A. Impact of formal continuing medical education: do conferences, workshops, rounds, and other traditional continuing education activities change physician behavior or health care outcomes? *JAMA*. 1999;282:867-874.

# Users' Guides to the Medical Literature: XXII: How to Use Articles About Clinical Decision Rules

Thomas G. McGinn; Gordon H. Guyatt; Peter C. Wyer; et al.

*JAMA*. 2000;284(1):79-84 (doi:10.1001/jama.284.1.79)

http://jama.ama-assn.org/cgi/content/full/284/1/79

# Users' Guides to the Medical Literature

## XXIII. Qualitative Research in Health Care

## A. Are the Results of the Study Valid?

Mita K. Giacomini, PhD

Deborah J. Cook, MD, MSc

for the Evidence-Based Medicine Working Group

**Quantitative research is designed to test well-specified hypotheses, determine whether an intervention did more harm than good, and find out how much a risk factor predisposes persons to disease. Equally important, qualitative research offers insight into emotional and experiential phenomena in health care to determine what, how, and why. There are 4 essential aspects of qualitative analysis. First, the participant selection must be well reasoned and their inclusion must be relevant to the research question. Second, the data collection methods must be appropriate for the research objectives and setting. Third, the data collection process, which includes field observation, interviews, and document analysis, must be comprehensive enough to support rich and robust descriptions of the observed events. Fourth, the data must be appropriately analyzed and the findings adequately corroborated by using multiple sources of information, more than 1 investigator to collect and analyze the raw data, member checking to establish whether the participants' viewpoints were adequately interpreted, or by comparison with existing social science theories. Qualitative studies offer an alternative when insight into the research is not well established or when conventional theories seem inadequate.**

*JAMA. 2000;284:357-362* www.jama.com

### CLINICAL SCENARIO

At a Monday morning meeting of your hospital's continuous quality improvement committee, the last agenda item is an initiative to enhance patient-clinician communication. The chair proposes that all medical charts include a form to record patient wishes about cardiopulmonary resuscitation and end-of-life care. The committee members agree in principle on the goals of enhanced communication and more accurate documentation of patient preferences. However, you raise potential concerns about how these forms might change the nature of end-of-life decision making and even impair communication. As the meeting draws to a close, you pose a fundamental question to the group for discussion the following week: Could life support preference forms unduly routinize and constrain dialogue between clinicians and patients or family members?

### THE SEARCH

Emerging from the meeting, you resolve to learn more about the influence of institutional record keeping on "do not resuscitate" communication during acute illness. Back in your office, you do a quick search of MEDLINE using key words *resuscitation orders* (508 hits) and *patient-physician relations* (5040 hits), and *patient participation* (1680 hits). Of 11 citations, 1 publication is a cultural analysis that you pick up en route to clinic.[1] The objectives of this study were to examine the influence of a Limitations of Medical Care form on discussions about cardiopulmonary resuscitation and the meaning that underlies this communication.

### INTRODUCTION

Clinicians are trained to think mechanistically and to draw conclusions using pathophysiologic rationale and deductive reasoning. The biomedical literature reflects this orientation, and clinicians are therefore most familiar with deductive quantitative research. Quantitative studies (such as epidemio-

**Author Affiliations:** Department of Clinical Epidemiology and Biostatistics (Drs Giacomini and Cook), Centre for Health Economics and Policy Analysis (Dr Giacomini), Department of Medicine, Divisions of General Medicine and Critical Care for the Evidence-Based Medicine Working Group (Dr Cook), McMaster University, Faculty of Health Sciences, Hamilton, Ontario.

The original list of members (with affiliations) appears in the first article of the series (*JAMA*. 1993; 270:2093-2095). A list of new members appears in the 10th article of the series (*JAMA*. 1996;275:1435-1439). The following members of the Evidence-Based Working Group contributed to this article: Gordon H. Guyatt, MD, MSc, Daren Heyland, MD, Anne Holbrook, MD, MSc, Virginia Moyer, MD, MPH, Andrew D. Oxman, MD, MSc, and W. Scott Richardson, MD. Dr Cook is a Career Scientist of the Ontario Ministry of Health. Dr Giacomini is a National Health Research Scholar of Health Canada.

**Reprints:** Gordon H. Guyatt, MD, MSc, Department of Clinical Epidemiology and Biostatistics, Room 2C12, 1200 Main St W, McMaster University Faculty of Health Sciences, Hamilton, Ontario, Canada L8N 3Z5.

**Users' Guides to the Medical Literature Section Editor:** Drummond Rennie, MD, Deputy Editor.

logic investigations and clinical trials) aim to test well-specified hypotheses concerning some predetermined variables. These studies suitably answer questions such as whether (eg, whether an intervention did more good than harm), or how much (eg, how strongly a risk factor predisposes patients to a disease). However, medicine is not only a mechanistic and quantitative science but also an interpretive art.[2]

Interpretive research asks questions about social interactions that can be addressed systematically through qualitative methods.[3] Qualitative research offers insight into social, emotional, and experiential phenomena in health care. Examples include inquiry about the meaning of illness to patients, their loved ones, and their families or about the attitudes and behavior of patients and clinicians. Qualitative research questions tend not to ask whether or how much but rather to explore what, how, and why. Qualitative studies may pursue a variety of theory-generating aims, including to explore and describe social phenomena faithfully (including surveying diverse perspectives or by giving voice to those not usually heard[4]), to identify potentially important variables or concepts, to recognize patterns and relationships, and to generate coherent theories and hypotheses. Qualitative reports do not typically generate answers but rather generate narrative accounts, explanations, typologies of phenomena, conceptual frameworks, and the like. For example, Ventres et al[1] explore what patient-physician communication occurred during discussions about resuscitation and how the use of a standard form influences communication between physicians and families about do-not-resuscitate orders. Another qualitative study probes why family members select certain processes for discontinuing life support.[5]

Just as clinicians use complementary types of information to draw clinical conclusions, complementary research methods are often useful in examining different aspects of a health problem.[6-9] Qualitative studies offer a rigorous alternative to armchair hypothesizing in areas for which insight may not be well established or for which conventional theories seem inadequate. Qualitative and quantitative studies each make useful contributions to knowledge in themselves. They may also be used in tandem—qualitative investigation to generate theories and identify relevant variables and quantitative investigation to test the implied hypotheses about relationships between those variables. Alternatively, qualitative and quantitative approaches can unfold concurrently within a research program, informing each other during the analysis and interpretation phases, yielding findings that are broader in scope and richer in meaning than if only 1 approach were used. Details about how to conduct qualitative research,[10-13] as well as the attributes and limitations of qualitative vs quantitative research approaches have been published elsewhere.[14-20]

## THE GUIDES

In this 2-part Users' Guide, we suggest guides for understanding and critically appraising qualitative research articles using the previously established Users' Guides framework: (1) Are the results of this study valid (or *credible*)? (2) What are the results? and (3) How can they help me care for my patients? In the first article of this pair, we focus on assessing the *validity* of qualitative research reports.

### Are the Results of the Study Valid?

Clinical readers traditionally think of research validity as the truthful correspondence of results with an objective reality. Qualitative research offers empirically based insights about social or personal experiences, which necessarily have a strongly subjective—but no less real—nature than biomedical phenomena. To avoid confusion, qualitative researchers typically avoid the term *valid* in favor of alternatives such as *credible*.[9,12(pp289-331)] Even so, qualitative insights must emerge from systematic observations and competent interpretation, correspond well to the social reality experienced by the participants and also have meaning for those who will read and learn from the report. Clinical readers in particular need to judge the relevance of qualitative research reports to their own practice, interests, or patient care questions.

To judge the methodologic rigor of qualitative research reports, readers need to appraise critically the study design and analysis. This appraisal should examine whether the study was designed to address its research question and objectives appropriately and whether it was conducted rigorously enough to achieve its empirical aims. Ventres et al[1(p134)] clearly describe their objective: "to examine the use of the Limitations of Medical Care form in the context of actual hospital practice, . . . to evaluate interactive elements of the resuscitation decision, . . . [and] to explore what is said when discussing code status, how information is communicated among parties involved, and the meaning that underlies this communication." Consistent with typical aims of qualitative inquiry, the study focuses on social interactions and their meaning. The objectives describe the social phenomena to be explored and described, rather than specific hypotheses to be tested.

The Methods section of a qualitative study should describe several aspects of the research design, including (1) how study participants were selected, (2) the methods used to generate data, (3) the comprehensiveness of data collection, and, (4) procedures for analyzing the data and corroborating the findings. As with any research, qualitative research involving human subjects must undergo ethics review and approval and this approval should be noted in the report. Special ethical dilemmas in qualitative research[21] should be addressed in the ethics and peer review of the study protocol, but usually cannot be appraised from the published report alone. Following are some general guidelines to help readers determine whether qualitative research design and execution is appropriate for the research objectives.

### Were Participants Relevant to the Research Question and Was Their Selection Well Reasoned?

Qualitative studies discover and describe important variables, particularly in terms of the social dynamics and the subjective realities of those involved a given situation.[3,12(pp70-91)] The units of analysis in a given qualitative study therefore may include myriad social phenomena, such as individuals, groups, documents, artifacts, interactions, dialogues, incidents, or settings.

The exploratory nature of qualitative research typically requires investigators not to prespecify a study population in strict terms, lest an important person, variable, or unit of analysis be overlooked. In some studies (eg, content analyses of documents), the scope of data collection can be prespecified, but if so, the rationale should be sensible to the reader. The consecutive or random selection of participants that is common in quantitative research is replaced by purposive sampling in qualitative research. Sampling aims to cover a range of potentially relevant social phenomena and perspectives from an appropriate array of data sources. Selection criteria often evolve over the course of analysis, and investigators return repeatedly to the data to explore new cases or new angles. Purposive sampling might aim to represent any of the following: typical cases, unusual cases, critical cases, politically important cases, or cases with connections to other cases (ie, *snowball sampling*).* Least compelling is the pursuit of merely convenient cases that are most easily accessed. Nevertheless, many qualitative studies do rely on convenience sampling to some extent (eg, for pragmatic reasons, study participants may only be those who speak the same language as the investigators, or only individuals who are willing to be interviewed). Readers of qualitative studies should look for sound reasoning for describing and justifying the participant selection strategies.

In the report by Ventres et al,[1] the unit of analysis was not the patient but rather the social interaction among several parties: the patient, family members, nurses, social workers, clergy, and residents involved in resuscitation discussions about a particular patient. From a potential sample of 8 patients, 3 cases were selected for in-depth study. The criteria for selecting these particular 3 cases were not specified, leaving readers unable to judge their appropriateness and how comprehensively they illustrate communication issues involving resuscitation directives in the hospital.

### Were the Data Collection Methods Appropriate for the Research Objectives and Setting?

The most common qualitative data collection methods involve field observations, interviews, or document analysis, separately or in combination. The collected data allow the researchers to observe, as clearly as possible, the social interactions or behavior that they seek to describe.

**Field Observation.** The purpose of field observation is to record social phenomena directly and prospectively. There are 2 basic approaches: direct observation by investigators themselves and indirect observation through audiotape or videotape recording. In direct observation, investigators spend time in the social milieu that they are studying and record observations in the form of detailed field notes or journals. Observational techniques are categorized according to the role of the investigator in the setting (ie, nonparticipant or participant) observation. Field analysis techniques require investigators to consider explicitly how their presence might influence their findings.

In nonparticipant observation, the researcher stays relatively uninvolved in the social interactions he/she observes. The crucial question for critical appraisal is whether a "fly on the wall" observer of a particular social setting will effectively be ignored by study participants or might instead inadvertently influence participants' behavior. For example, a researcher in crowded waiting room may go unnoticed and hence observe the natural unfolding of events. In contrast, in a clinic examining room, he/she may be conspicuous, and significantly change the social interactions he/she is there to observe. Audiotape or videotape recordings are sometimes used as less intrusive methods of capturing data. However, they also have drawbacks. First, recorders can occupy a social role and be experienced by participants as partaking in surveillance, thus influencing participants' behavior. Second, recorders' observational powers are limited by their range of operation: if the action is moving around or if visual cues are missing, important information may be lost.

In participant-observation investigations, the researcher is acknowledged as a part of the social setting, either as a researcher per se or as a more directly involved actor (eg, social worker, ethicist, committee member, etc). Again, the question for critical appraisal is whether the dual observer-participant role allows access to natural candid social interactions among other participants in the setting.

In both participant and nonparticipant field observation, the effect of the researcher on the social setting can never be controlled for (a common goal of experimental study designs). Interactions between researchers and those they study are somewhat paradoxically but necessarily regarded as both a useful source of data and a potential source of bias.[12(pp92-109)] More than 1 observational technique (eg, personal observations and audiotape recording dialogue) can sometimes be used to capture more detailed data and to help analyze observer effects.

**Interviews.** Qualitative studies may use several types of interviews. The most popular are semistructured, in-depth, individual interviews and focus groups. Structured approaches, such as standardized questionnaires, are usually inappropriate for qualitative research, because they presuppose too much of what respondents might say and do not allow respondents to express themselves in their own terms. These problems limit the opportunity to gain insight into personal and social phenomena and can im-

*References 12 (pp187-220), 13 (pp145-198).

pose the investigators' preconceived notions onto the data.

The appropriate interview method depends on the topic. Individual interviews tend to be more useful for evoking personal experiences and perspectives, particularly on sensitive topics. Group interviews tend to be more useful for capturing interpersonal dynamics, language, and culture. Focus groups can be appropriate for discussing emotionally sensitive topics if participants feel empowered to speak in the presence of peers; however, the public forum of a focus group can also inhibit candid disclosure.[22,23] Critical readers should look for the rationale for choosing a particular approach and its appropriateness for the topics addressed. Using more than 1 interview method may be helpful for capturing a wider range of information.

**Document Analysis.** Finally, documents such as charts, journals, correspondence, and other material artifacts can provide qualitative data.[24] These are especially useful in policy, historical, or organizational studies of health care. There are different approaches to the analysis of documents. One involves counting specific content elements (eg, frequencies of particular words being used) while the other involves interpreting text as one would interpret any other form of communication (eg, seeking nuances of meaning and considering context). The former approach, especially if used alone, rarely provides adequate information for a qualitative, interpretive analysis.

Ventres et al[1] used 3 types of data collection: participant observation, audiotapes of discussions, and semistructured interviews. Details of the interview strategy appear in an appendix and provide additional information about the content of the interviews and techniques used to elicit responses. Three types of questions were asked: open-ended, semistructured, and contrast questions, to elicit opinions on contrasting hypothetical patient situations. The use of multiple data collection methods and sources adds rigor to this study, because it allows investigators to examine discussions of the Limitations of Medical Care from several angles and to capture information with one method that may be overlooked for another.

### Was the Data Collection Comprehensive Enough to Support Rich and Robust Descriptions of the Observed Events?

Another critical appraisal question is whether the social setting or experience was observed thoroughly enough to support rich and robust descriptions of the observed events. The analytic issue here is not one of sample size in the statistical sense. Rather than aim for a specific number of participants (or other units of analysis), researchers should strive for adequately in-depth observations. A qualitative study involving many participants but only cursory interactions with each 1 may be less rigorous than a study involving few participants but extensive observation of each. Data collection needs to be comprehensive enough in both breadth (types of observations) and depth (extent of observation of each type) to generate and support the interpretations. This criterion has a circular quality, that is, whether data are adequate depends to some extent on the nature of the findings and vice versa. For this reason, qualitative data collection and analysis steps usually iterate: data collection is followed by analysis, which in turn gives direction for new data collection, and so forth.

Several aspects of a qualitative report indicate how extensively the investigators collected data: the number of observations, interviews, or documents; the duration of the observations; the duration of the study period; the diversity of units of analysis and data collection techniques; the number of investigators involved in collecting and analyzing data; and, the degree of investigators' involvement in data collection and analysis.

Interpretive research is characterized by voluminous data, consisting of paper files (eg, field notes, transcripts, journals, analytic memos, photocopied documents, etc) and electronic media (eg, word-processed transcripts, audiotapes, videotapes, etc). How these data are recorded and accessed affects the depth and quality of the findings. The goal of data collection is to produce detailed data as representative of the experience as possible and to leave a trail of data and analysis that another investigator could potentially follow. While qualitative research cannot be replicated, it can be audited. Of course, outsiders to a study cannot observe exactly what the investigators observed, and because auditors bring their own unique perspectives, they can legitimately develop somewhat different interpretations of the same data. Such alternative interpretations would not necessarily reveal an analysis as faulty, since there are multiple truthful ways to depict social behavior. Nevertheless, in principle, qualitative researchers should organize and interpret their data in such a way that another investigator could follow what was done and could see a clear correspondence between the empirical data and the interpreted findings.

There are several conventions for taking field observations and interview notes.* Most emphasize thoroughness, the classification of observations, and self-consciousness of personal experiences and biases. Taping and transcribing interviews (or other dialogue) is desirable. Qualitative research transcription is different from that used for medical dictation. For typical medical records, breathing, pauses, and changes in volume are ignored by the transcriptionist. For a qualitative research transcript, these behaviors can provide valuable data that help elaborate the meaning of the spoken words; in fact, transcripts are seldom corrected for grammar or word choices. Qualitative investigators also often keep records of their personal thoughts and experiences to distinguish them carefully from other observations. This helps to isolate personal biases, as well as to use personal experiences as analytically useful information.†

*References 12 (pp250-288), 13 (pp199-276), 25, 26.
†References 12 (pp250-288), 13 (pp199-276, 371-459), 25, 26.

Ventres et al[1] conducted their study over 4 months, during which family practice residents identified 8 hospitalized patients about whom they had discussions regarding resuscitation. Of these, investigators observed 3 discussions among patients, their families, and their physicians; 2 of these 3 cases are reported in detail. Both before and after the discussions, interviews were conducted with the patients, family members, nurses, social workers, clergy, and physicians regarding the decision-making process. Ventres et al audiotaped and transcribed interviews as well as discussions among physicians, patients, and families. The transcription process is detailed in an appendix to the article. An observer also made written records of nonverbal communications, which are not well captured by audiotape. Finally, the investigators also recorded secondary interpretive data (ie, their personal interpretations of the discussions they observed). By collecting data using several methods, these investigators enhanced their ability to capture important nuances in communication and to develop robust accounts of the discussions.

The inclusion of patients, family members, and several members of the health care team as participants in this study increases the number of perspectives from which the issue of resuscitation was considered. No key participant's perspectives seem to have been overlooked in the data collection. However, whether data collection was comprehensive for each participant is difficult to assess, given the different roles that each have in such decisions and the complexities of end-of-life dialogue. Examining only 3 cases in which resuscitation is discussed is unlikely to capture the diversity of perspectives, content, and styles found in such conversations and could produce a limited description. The authors themselves note that this small number of cases is a potential study limitation and that more variability may have yielded further insight into other possible structures of resuscitation discussions.

### Were the Data Appropriately Analyzed and the Findings Adequately Corroborated?

Qualitative researchers begin with a general exploratory question and preliminary concepts. They then collect relevant data, observe patterns in the data, organize these into a conceptual framework, and resume data collection to explore and challenge this conceptual framework. This cycle may be repeated several times. The iteration between data collection, analysis, and theory development continues until a conceptual framework is well-developed and further observations yield minimal or no new information to further challenge or elaborate the framework (a point variously referred to as *theoretical saturation*[27] or *informational redundancy*[12(pp221-249)]). This analysis-stopping criterion is so basic to qualitative analysis that authors seldom declare that they reached this point and assume this to be understood by the reader.

In the course of analysis, key findings are also triangulated, meaning that they are corroborated using multiple sources of information (the term *triangulation* is a metaphor and does not mean literally that 3 or more sources are required). The appropriate number of sources will depend on the importance of the findings, their implications for theory and the investigators' confidence in their validity. Because no 2 qualitative data sources will generate exactly the same interpretation, much of the art of qualitative interpretation involves exploring why and how different information sources yield slightly different results.[9,28]

Readers may encounter several useful triangulation techniques for validating qualitative data and their interpretation in analysis.[9,12(289-331),28] Investigator triangulation requires more than 1 investigator to collect and analyze the raw data, such that the findings emerge through consensus between investigators. This is best accomplished by an investigative team. Use of external investigators is controversial because their involvement in the case could be too superficial to yield deep understanding.[12(pp289-331),28] Team members representing different disciplines helps to prevent the personal or disciplinary biases of a single researcher from excessively influencing the findings. Member checking involves sharing draft study findings with the participants to inquire whether their viewpoints were faithfully interpreted, whether there are gross errors of fact, and whether the account makes sense to participants with different perspectives. Theory triangulation,[29] is a process whereby emergent findings are corroborated with existing social science theories.[21] It is conventional for authors to report how their qualitative findings relate to prevailing social theory, though it is controversial whether such theories should be used to guide the research design or analysis.

Some qualitative research reports describe the use of qualitative analysis software packages. Readers should not equate the use of computers with analytic rigor. Such software is a data management tool offering efficient methods for storing, organizing, and retrieving qualitative data. These programs do not perform analysis. Investigators themselves conduct the analysis as they create the keywords, categories, and logical relationships used to organize and interpret electronic data. The credibility of qualitative study findings depend on these investigator judgments that cannot be programmed into software packages. More generally, credible qualitative interpretation requires well-trained and well-prepared investigators who approach their work with both discipline and creativity.[9]

We indicated earlier that qualitative data collection must be comprehensive—adequate in its breadth and depth to yield a meaningful description. The closely related criterion for judging whether the data were analyzed appropriately is whether this comprehensiveness was determined in part by research results themselves, with the aims of challenging, elaborating, and corroborating the findings. This is most apparent when researchers state that they alternated between data collection and

analysis, collected data with the purpose of elucidating the analysis-in-progress, collected data until analytic saturation or redundancy was reached, or triangulated findings using any of the methods mentioned.

Ventres et al[1(p141)] approached data coding using 3 broad preliminary concepts in patient-clinician communication: (1) control, (2) giving or withholding information, and (3) attentiveness. Researchers commonly use sensible, broad conceptual categories such as these to begin making sense of their data, but the categories also are commonly revised in the course of analysis. These investigators noted that data collection and analysis proceeded iteratively, by reporting that, "data collected and analyzed on the first members of the sample influenced the collection of information on subsequent members." Several triangulation techniques were used, including methodologic triangulation (using several data collection methods of participant observation, audiotaping, and semistructured interviews), investigator triangulation (duplicate interpretation of audiotapes), disciplinary triangulation (clinical, anthropological, psychiatric, and sociologic perspectives), and member checking (by professional and lay participants in the study).

The authors report that the principal author and a sociolinguist reviewed the audiotapes blinded to "all but necessary case information," however it is unclear which data were and were not available to these investigators prior to analysis. Readers should not assume that blinding necessarily improved the rigor of the analysis, since limiting access to data also limits investigators' ability to make well-informed interpretations of possibly complex social interactions.

We note that Ventres et al's final findings quite appropriately do not strictly follow their 3 provisional analytic categories (control, information giving, attentiveness), but instead reveal more specific and concrete dynamics focusing on (1) the Limitations of Medical Care form's tendency to frame discussions to exclude patient values and beliefs, (2) family-physician differences in reasoning style, and (3) consequential confusion between instrumental treatment decisions and more general goals of care. This progression suggests that the conceptual findings did develop as a result of the empirical observations. The authors relate their findings back to general social health policy and ethical concerns about who is and who should be in control of limitations-of-care decision processes.

Having determined that the validity of the study by Ventres et al[1] is sufficient to gain some understanding of the impact of a Limitations of Medical Care form on patient-clinician communication, we turn to the second part of this Users' Guide. In it, we will address, What are the results, and How do they help me care for my patients?


**REFERENCES**

1. Ventres W, Nichter M, Reed R, Frankel R. Limitation of medical care: an ethnographic analysis. *J Clin Ethics*. 1993;4:134-145.
2. Battista RN, Hodge MJ, Vineis P. Medicine, practice and guidelines: the uneasy juncture of science and art. *J Clin Epidemiol*. 1995;48:875-880.
3. Hughes J. The interpretive alternative. In: Hughes J, ed. *The Philosophy of Social Research*. New York, NY: Longman; 1990:89-112.
4. Sofaer S. Qualitatve methods: what are they and why use them? *J Health Serv Res*. 1999;34:1101-1118.
5. Tilden VP, Tolle SW, Garland MJ, Nelson CA. Decisions about life-sustaining treatment: impact of physicians' behaviors on the family. *Arch Intern Med*. 1995; 155:633-638.
6. Rosenfield PL. The potential of transdisciplinary research for sustaining and extending linkages between the health and social sciences. *Soc Sci Med*. 1992;35:1343-1347.
7. Stange KC, Zyzanski SJ. Integrating qualitative and quantitative research methods. *Fam Med*. 1989;21: 448-451.
8. Goering P, Streiner DL. Reconcilable differences: the marriage of qualitative and quantitative methods. *Can J Psychiatr*. 1996;41:491-497.
9. Patton MQ. Enhancing the quality and credibility of qualitative analysis. *Health Serv Res*. 1999;34:1189-1208.
10. Denzin NK, Lincoln YS. *Handbook of Qualitative Research*. Thousand Oaks, Calif: SAGE Publications; 1994.
11. Corbin J, Strauss A. Grounded theory research: procedures, canons, and evaluative criteria. *Qualitative Sociology*. 1990;13:3-23.
12. Lincoln YS, Guba EG. *Naturalistic Inquiry*. London, England: Sage Publications; 1985.
13. Patton MQ. *Qualitative Evaluation and Research Methods*. London, England: Sage Publications; 1990.
14. Poses RM, Isen AM. Qualitative research in medicine and health care: questions and controversy. *J Gen Intern Med*. 1998;13:32-38.
15. Robling MR, Owen PA, Allery LA, et al. In defense of qualitative research: responses to the Poses and Isen perspectives article. *J Gen Intern Med*. 1998; 13:64-72.
16. Guba EG, Lincoln YS. Competing paradigms in qualitative research. In: Denzin NK, Lincoln YS, eds. *Handbook of Qualitative Research*. Thousand Oaks, Calif: Sage Publications; 1994:105-117.
17. Morse J. Is qualitative research complete? *Qual Health Res*. 1996;6:3-5.
18. Smith JK. Quantitative vs qualitative research: an attempt to clarify the issue. *Educ Res*. 1983;12:6-13.
19. Neumann WL. The meanings of methodology. In: Neumann WL, ed. *Social Research Methods*. Boston, Mass: Allyn & Bacon; 1991:43-66.
20. Waitzkin H. On studying the discourse of medical encounters: a critique of quantitative and qualitative methods and a proposal for reasonable compromise. *Med Care*. 1990;28:473-488.
21. Holloway I, Wheeler S. Ethical issues in qualitative nursing research. *Nurs Ethics*. 1995;2:223-232.
22. Kitzinger J. Introducing focus groups. *BMJ*. 1995; 311:299-302.
23. Steward DW, Shamdasani PN. *Group Dynamics and Focus Group Research: Focus Groups: Theory and Practice*. London, England: Sage Publications; 1990: 33-50.
24. Hodder I. The interpretation of documents and material culture. In: Denzin NK, Lincoln YS, eds. *Handbook of Qualitative Research*. London, England: Sage Publications; 1994:393-402.
25. Kirk J, Miller ML. *Reliability and Validity in Qualitative Research*. London, England: Sage Publications; 1986.
26. Schatzman L, Strauss AL. Strategy for recording. In: Schatzman L, Strauss AL, eds. *Field Research: Strategies for a Natural Sociology*. Englewood Cliffs, NJ: Prentice-Hall; 1973:94-107.
27. Glaser B, Strauss AL. *The Constant Comparative Methods of Qualitative Analysis: Discovery of Grounded Theory*. New York, NY: Aldine de Gruyter; 1967:101-116.
28. Stake R. Triangulation. In: Stake R, ed. *The Art of Case Study Research*. London, Engalnd: Sage Publications; 1995:107-120.
29. Denzin NK. *Sociological Methods*. New York, NY: McGraw Hill; 1978.

# Users' Guides to the Medical Literature: XXIII. Qualitative Research in Health Care A. Are the Results of the Study Valid?

Mita K. Giacomini; Deborah J. Cook; for the Evidence-Based Medicine Working Group

*JAMA*. 2000;284(3):357-362 (doi:10.1001/jama.284.3.357)

http://jama.ama-assn.org/cgi/content/full/284/3/357

# Users' Guides to the Medical Literature

## XXIII. Qualitative Research in Health Care
## B. What Are the Results and How Do They Help Me Care for My Patients?

Mita K. Giacomini, PhD
Deborah J. Cook, MD
for the Evidence-Based Medicine Working Group

**The second part of this 2-part series on how to interpret qualitative research addresses, "what are the results," and, "how do they help me care for my patients?" Qualitative analysis is a process of summarizing and interpreting data to develop theoretical insights that describe and explain social phenomena such as interactions, experiences, roles, perspectives, symbols, and organizations. Key results are often illustrated with excerpts from interview transcripts, field notes, or documents. The results of a qualitative research report are best understood as an empirically based contribution to ongoing dialogue and exploration. Empirically based theory evolves from a process of exploration, discovery, analysis, and synthesis. Each concept should be defined carefully in a way that is meaningful to the reader. Concepts should be adequately developed and illustrated when theoretical conclusions are drawn. Arguments should be explained and justified. The qualitative research report ideally should address how the findings relate to other theories in the field. The qualitative study can provide a useful road map for understanding and navigating similar social settings interactions, or relationships.**

*JAMA. 2000;284:478-482* www.jama.com

IN THE FIRST OF THIS 2-PART ARTICLE on using qualitative research[1] we described a hospital's continuous quality improvement committee initiative to introduce a medical form designed to enhance patient-clinician communication about cardiopulmonary resuscitation. The clinician in this scenario wondered whether the impact of introducing such a document had been evaluated with respect to its influence on patient-clinician communication. She found the study by Ventres et al[2] and critically appraised its validity.

The objective of the study was to examine how a limitation of medical care form affects resuscitation dialogue among patients, their families, and resident physicians. The investigators collected data through participant observation, audiotapes of life support discussions, and semistructured interview. Participants included patients, family members, nurses, social workers, clergy, and resident physicians. The article analyzes thoroughly the decision-making discussions concerning 3 of 8 patient cases studied. Analytic rigor is demonstrated by the corroboration (triangulation) of findings among different sources of data, multidisciplinary investigators, and critiques of the analysis by study participants. Although many perspectives were incorporated in this study and 3 cases were considered comprehensive, the breadth was probably too narrow to capture the diversity of communication and decision-making styles concerning end-of-life treatment. In the second part of this Users' Guide on how to interpret qualitative research, we will address the questions: What are the results of this study? and, how do the results help me care for my patients?

### WHAT ARE THE RESULTS OF THE STUDY?

In summary, Ventres and colleagues[2] found that use of the limitation of medical care form, which is intended to facilitate decision making, can routinize the clinician-patient dialogue to meet bureaucratic needs, narrowing rather than enhancing communication about resuscitation. After outlining the foundation of the results of qualitative re-

**Author Affiliations:** Department of Clinical Epidemiology and Biostatistics (Drs Giacomini and Cook), Centre for Health Economics and Policy Analysis (Dr Giacomini), Department of Medicine, Divisions of General Medicine and Critical Care for the Evidence-Based Medicine Working Group (Dr Cook), McMaster University, Faculty of Health Sciences, Hamilton, Ontario.

The original list of members (with affiliations) appears in the first article of the series (*JAMA*. 1993;270:2093-2095). A list of new members appears in the 10th article of the series (*JAMA*. 1996;275:1435-1439). The following members of the Evidence-Based Medicine Working Group contributed to the article: Gordon H. Guyatt, MD, MSc, Daren Heyland, MD, Anne Holbrook, MD, MSc, Virginia Moyer, MD, MPH, Andrew D. Oxman, MD, MSc, and W. Scott Richardson, MD. Dr Cook is a Career Scientist of the Ontario Ministry of Health. Dr Giacomini is a National Health Research Scholar of Health Canada.

**Reprints:** Gordon H. Guyatt, MD, MSc, Department of Clinical Epidemiology and Biostatistics, Room 2C12, 1200 Main St W, McMaster University Faculty of Health Sciences, Hamilton, Ontario, Canada L8N 3Z5.

**Users' Guides to the Medical Literature Section Editor:** Drummond Rennie, MD, Deputy Editor.

search reports below, we describe the results of that study in more detail.

The goal of qualitative research is to develop theoretical insights that describe and explain social phenomena such as interactions, experiences, roles, perspectives, symbols, and organizations. Qualitative analysis is foremost a process of summarizing and interpreting data, "based on the value of trying to represent faithfully and accurately the social worlds or phenomena studied."[3] A good qualitative report will be received as robust and truthful across multiple perspectives (ie, those of study participants, authors, readers, colleagues). Broad endorsement does not make the findings infallible but helps to establish that the analysis offers a meaningful approximation to the truth of a social phenomenon.

Qualitative results contain description and theory. Reports typically present these in an integrated fashion, by describing key theoretical insights and illustrating them with descriptions from the data. Readers can judge the importance and usefulness of the findings by asking how evocative and thorough the descriptions are, as well as how comprehensive and relevant the theoretical insights are.

### How Evocative and Thorough Is the Description?

The product of a qualitative study is a narrative. It describes a social phenomenon and draws theoretical insights (and sometimes practical lessons) in conclusion. The writing style should be clear, accessible, and "tell the story" well. A good qualitative report provides enough descriptive detail to evoke a vivid picture of the social setting or interactions studied. To do this, authors usually illustrate key findings with data excerpts from field notes, interview transcripts, or documents. These data should clearly support the main points and offer contextual detail. The use of examples and reference to sources gives the reader insight into the nature of the social phenomenon as well as the sensibility of how investigators interpreted it. Because of the importance of detail in qualitative reports, some health research journals allow substantially longer page limits for qualitative studies. However, longer articles are not necessarily superior. Unfocused analyses, weighted too heavily with description, can obscure the study's main focus At the other extreme, theoretical treatises that do not include adequate support by providing illustrative data and empirical description may raise questions about the extent to which the findings were derived from the evidence.

In their results section, Ventres et al tell the story by recounting the case histories of 2 patients and those involved in their care. These 2 scenarios are organized chronologically (rather than conceptually), which helps draw the reader into the events and discussions as they unfold. The narratives are liberally illustrated with excerpts from interviews and taped discussions, which give readers more intimate insight into the situations studied. The excerpts also support the authors' interpretations of the structure of these life support discussions (ie, as involving characteristic content, dyadic conversation, and pervasive ambiguity). Although the exposition is restricted to 2 cases and selected excerpts, the information is rich and coherently organized.

### How Comprehensive and Relevant Are the Theoretical Conclusions?

Qualitative inquiry aims to develop theoretical conclusions. Some systematic approaches to theory development are described.[4-7] However, there is no correct approach. Whatever the system, the investigators' training, perceptiveness, creativity, and intellectual discipline will also play a role.[8,9] The critical analysis of social theory commands extensive attention in the humanities and social sciences, much of which is beyond the scope of this Users' Guide. Basically, to be meaningful and useful, a theory should be adequately comprehensive and relevant.

**Comprehensiveness.** Theoretical findings must be well reasoned and coherent. Elder and Miller[10] suggest that coherent theory possesses the qualities of parsimony (invokes a minimal number of assumptions), consistency (accords with what is already known and inconsistencies are well explored and explained), clarity (expresses ideas evocatively and sensibly), and fertility (suggests promising directions for further investigation). On a concrete level, narrative arguments should be logical and plausible, metaphors should provide useful analogies, and illustrative frameworks such as diagrams should meaningfully label the elements and relationships depicted.

Readers could think of theory as having a kind of anatomy and should examine each of its parts to understand its contribution to knowledge. Theory consists of concepts and their relationships. Furthermore, empirically based theory evolves from a process of exploration, discovery, analysis, and synthesis. In its final form, empirically based theory relates clearly to the data and makes a contribution to theoretical knowledge in the field. Readers can examine these 5 aspects of theory by asking the following corresponding questions.

*What Major and Minor Concepts Does the Theory Entail, and How Well Defined Are They?* Concepts are the basic building blocks of theory. Sometimes (but not necessarily) concepts will be organized hierarchically, with 1 overriding concept (perhaps a useful metaphor), a few broad categories within it, and a series of subcategories within those. It is possible for qualitative concepts to overlap or to be related in a nonhierarchical structure such as a web of interrelationships. Taxonomies and domain descriptions are conceptual frameworks that commonly appear in the biomedical literature. Whatever their number and form, each concept should be defined carefully and in a way that is meaningful to the reader.

*What Are the Relationships Between the Conceptual Categories, Are These Dynamics Clearly Described, and Do They Make Sense?* These questions focus on relationships between concepts. Such

dynamics may take a form similar to quantitative relationships between variables (eg, changes in one variable causing an increase or decrease in another). Alternatively, categories may have qualitative effects on each other (eg, one phenomenon may frame the form that another may take).

*Are the Concepts Adequately Developed and Illustrated?* Several devices may be used to explain how the theoretical conclusions were drawn. For example, a report may describe chronologically the experience of entering the field and from there lead the reader through the key discovery experiences that form the backbone of the author's findings (however this approach is not appropriate for all studies, such as document analysis or the study of familiar settings). Theory can also be explained and justified using other rhetorical devices, such as argument. Conceptual frameworks are strongest when their categories or variables embrace a full range of empirical phenomena observed. Illustrative data excerpts offer glimpses into the analytic process, but these glimpses help demonstrate how the investigators interpreted the data. If the illustrative examples do not seem to fit well with the interpretive explanation, the validity of the rest of the analysis comes into question.

*Where Does the Empirically Generated Theory Fit in Relation to Existing Theory and Beliefs in the Field?* Readers should look for whether the results of a qualitative research report address how the findings relate to other theory in the field. Empirically developed insights need not agree with existing beliefs. Whether they agree or not, the findings' relationship to prevailing theories and beliefs should be addressed in a critical manner. Qualitative approaches vary with regard to the role that theoretical literature plays: some methods use existing literature to guide empirical work, whereas others do not address the literature until after empirical findings are established.[5,11] In either case, the report should indicate how the findings relate to scholarship in the field.

Ventres et al[2] offer relatively pragmatic theoretical conclusions about how an administrative form can reflect and reinforce mechanistic objective-oriented dialogue to the neglect of patient needs, values, and beliefs. In this study, the hospital's Limitation of Medical Care form was used as both the foundation for dialogue and the vehicle for expression of patient wishes. Ventres et al describe how the form, together with conventional physician communication styles, can have the adverse effect of structuring conversations to obstruct candid conversation and obscure patient and family wishes. To help the clinician best, the study might have developed a more comprehensive model of communication about life support or of how administrative forms express (or suppress) meaningful health directives. Ventres et al[2] do not develop their theoretical conclusions to this degree. Rich description with relatively light theorizing is typical of many ethnographic or naturalistic studies, and this appraisal does not by any means indicate a scientific failing of the research. However, it may limit the usefulness of the research for the clinician's purposes. We should also note that this type of qualitative study does not feed directly into a hypothesis-testing research program, because it does not put forth specific variables or causal relationships to be tested. This limits neither the research's usefulness nor its scientific contribution, and this study demonstrates well the value of qualitative studies for the purposes of enlightenment. Although the report offers modest formal theory, it does offer credible, evocative evidence of the sorts of dynamics that can occur during life support discussions. The illustrative excerpts and interpretive descriptions offer the clinical readers a vicarious experience and a unique vantage on interactions among patients, families, physicians, and medical forms.

The study's findings allow the practicing clinician to stand back from the clinical encounter and view some common communication dynamics from a more critical distance. Normally, clinicians are directly involved in their discussions with patients and families, and cannot both participate actively in a conversation and analyze it objectively. Clinicians reading the study by Ventres et al may recognize in the scenarios something of themselves, the people they care for, and the administrative forms they use. It may be surprising and affirming to see graphic evidence that inanimate medical forms can "participate" in discussions and control what can be said and heard. The theoretical insight that such medical forms can play an active role in communication may help clinicians recognize this dynamic in other settings. This qualitative evidence provides a cautionary tale of how medical forms can do more than promote administrative efficiency.

**Relevance.** The results of a qualitative research report are understood best as an empirically based contribution to ongoing dialogue and exploration, rather than as documentation of an invariant fact. The dialogue affects the meanings of social experiences, and the results of a dialogue translate these experiences for persons who might not otherwise understand each other's perspectives well. The relevance of the results of a qualitative article depends partly on its ability to communicate how well the investigators and the study participants communicated and how well the results of their communication is conveyed to the readers of the report. Each of these parties should be involved actively in making sense of the research results.[10]

The results of the study by Ventres et al[2] translate the perspectives of participants (patients, families, resident physicians, and clinicians involved in end-of-life decisions) and the readers of the research. For clinicians who are not routinely engaged in end-of-life decisions, these results offer a window-like view that provides insight into a clinical world many clinicians do not enter. For clinicians more involved in end-of-life decisions, this study offers a view more analogous to a mirror that reflects familiar interactions in a way

that allows clinicians to examine their own role, other participants' roles, and even the role of a medical form in determining how end-of-life decision making unfolds. Operating either as window or mirror, valuable perspective can be gained from qualitative evidence. The study highlights the potential tyranny of administrative forms when they are used to structure sensitive personal discussions.

## HOW DO THE RESULTS OF THIS STUDY HELP ME CARE FOR PATIENTS?

In their descriptive role, qualitative research findings can enhance awareness of social dynamics in the clinical setting. As illustrated by Ventres et al,[2] social dynamics can influence powerfully the process of care and consequently the outcomes. The more clinicians and patients are conscious of social factors at work in health care, the more constructively they can use them or change them in the pursuit of health and healing. In their theory-generating role, qualitative findings provide models for understanding. These models can be used to analyze similar situations and, similar to all models, help to simplify clinicians' understanding of complex phenomena. Qualitative studies may give clinicians insight into the experiences of patients and their families.

### Does This Study Help Me to Understand the Context of My Practice?

One criterion for the generalizability of a qualitative study is whether it provides a useful road map for readers to understand and navigate similar social settings themselves. The North American cultural value of autonomy was encoded in 1991 by Congress in the Patient Self-Determination Act. Since then many health care systems have created documents such as advance directives and other decision-making tools to systematize conversations about end-of-life care.

The article by Ventres et al[2] invites us to contemplate this policy trend critically. Readers may reflect on how business metaphors have infiltrated clinical practice, and how these types of resuscitation documents symbolically contractualize health care at the end of life, especially when patients are referred to as "clients," and health care workers as "providers." In this study, discussions about resuscitation were intervention specific, focusing on a series of basic and advanced life support technologies, in part due to the task-oriented prompts of the limitation of medical care form. One family member of a patient who was unable to speak for himself explained that "resuscitation was not appropriate in Indian culture."[2(p139)] The resident continued to describe the technical details of resuscitation even after the family had made it clear that it was not desired, which made this family member feel as though the physician did not really trust the family's decision (or implicitly, their portrayal of his wishes, were he able to speak for himself).

### Does This Study Help Me Understand My Relationships With My Patients and Their Families?

Interpretive research offers clinicians an understanding of roles and relationships. Many qualitative studies of interest to clinicians focus on communication among patients, families and caregivers. Other studies describe behaviors of these groups, either in isolation or during interactions with others.

In the study by Ventres et al,[2] the acuity and severity of the patients' illness meant that dialogue typically occurred between resident physicians and family members instead of patients themselves. The small number of patients and resident physicians studied in a university hospital limits the range of discussion styles that were identified. Some clinicians may be more likely to have prior long-term relationships with patients than those developed among family practice residents involved in this study, allowing for such conversations to occur in the relative comfort of the outpatient setting rather than during an acute illness episode. Regardless of whether readers work with resident physicians (or are resident physicians themselves), a report such as this affords an opportunity for all readers to ask themselves frankly how they broach end-of-life discussions with hospitalized patients, whether they can relate to the communication styles described in the study, and if they can, what implications this has for their practice.

Some clinicians may tend to focus on the overall goals of care in ways that are culturally meaningful for patients, rather than consider discrete interventions, as were reported in this study. Some clinicians may revisit goals of health care periodically and not necessarily coincidentally with hospital admissions. The study by Ventres et al[2] can increase our self-consciousness about how well we listen to patients and families, what language we use when explaining resuscitation to them, how well we try to understand their values and preferences (especially when patients and surrogate decision makers give discordant messages),[12,13] and how clinicians may unwittingly influence patient wishes even as they try to discern those wishes.

## SCENARIO RESOLUTION

Reflecting on the article by Ventres et al,[2] you cast your mind back to the continuous quality improvement committee meeting you attended this morning about patient-clinician communication. Thinking about your hospital's proposal for a similar Limitations of Medical Care form you are concerned. You wonder to what extent introduction of this form might shift your own discussions with patients away from eliciting illness experiences and understanding values to a more stilted dialogue with patients or next of kin about technological aspects of basic and advanced life support.

You decide that at the next meeting you will share the evidence you found about routinizing conversations between clinicians and patients, should such a Limitation of Medical Care form be introduced. You plan to circulate the Ventres et al[2] article before the next

meeting and recommend that the committee use it to help outline the potential advantages and disadvantages of introducing such a document in your hospital. Meanwhile, if this form is adopted, you plan to request that the committee evaluate its influence on end-of-life discussions, using multidisciplinary qualitative research methods.


**REFERENCES**

1. Giacomini MK, Cook DJ, for the Evidence-Based Medicine Working Group. Users' Guides to the medical literature, XXIII: qualitative research in health care A. are the results of the study valid? *JAMA*. 2000; 284:357-362.
2. Ventres W, Nichter M, Reed R, Frankel R. Limitation of medical care: an ethnographic analysis. *J Clin Ethics*. 1993;4:134-145.
3. Altheide DL, Johnson JM. Criteria for assessing interpretive validity in qualitative research. In: Denzin NK, Lincoln YS, eds. *Handbook of Qualitative Research*. London, England: Sage Publications; 1994: 485-499.
4. Schatzman L, Strauss AL. Strategy for analyzing. In: *Field Research: Strategies for a Natural Sociology*. Englewood Cliffs, NJ: Prentice-Hall; 1973:108-127.
5. Strauss A, Corbin J. *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*. London, England: Sage Publications; 1990.
6. Glaser B, Strauss AL. *Discovery of Grounded Theory*. New York, NY: Aldine de Gruyter; 1967:101-116.
7. Miles M, Huberman M. *Qualitative Data Analysis*. London, England: Sage Publications; 1994:245-262.
8. Patton MQ. Enhancing the quality and credibility of qualitative analysis. *Health Serv Res*. 1999;34: 1189-1208.
9. Lincoln YS, Guba EG. *Naturalistic Inquiry*. London, England: Sage Publications; 1985:92-109.
10. Elder NC, Miller WL. Reading and evaluating qualitative research studies. *J Fam Pract*. 1995;41:279-285.
11. Hamberg K, Johansson E, Lindgren G, Westman G. Scientific rigour in qualitative research: examples from a study of women's health in family practice. *Fam Pract*. 1994;11:176-181.
12. Secker AB, Meier DE, Mulvihill M, et al. Substituted judgment: how accurate are proxy predictions? *Ann Intern Med*. 1991;115:92-98.
13. Uhlmann RF, Pearlman RA, Cain KC. Physicians' and spouses' predictions of elderly patients' resuscitation preferences. *J Gerontol*. 1988;43:M115-M121.


> A scientist is one who, when he does not know the answer, is rigorously disciplined to speak up and say so unashamedly; which is the essential feature by which modern science is distinguished from primitive superstition, which knew all the answers except how to say, "I do not know."
>
> —Homer W. Smith (1895-1962)

# Users' Guides to the Medical Literature: XXIII. Qualitative Research in Health Care B. What Are the Results and How Do They Help Me Care for My Patients?

Mita K. Giacomini; Deborah J. Cook; for the Evidence-Based Medicine Working Group

*JAMA*. 2000;284(4):478-482 (doi:10.1001/jama.284.4.478)

http://jama.ama-assn.org/cgi/content/full/284/4/478

# Users' Guides to the Medical Literature

## XXIV. How to Use an Article on the Clinical Manifestations of Disease

W. Scott Richardson, MD
Mark C. Wilson, MD, MPH
John W. Williams, Jr, MD, MHS
Virginia A. Moyer, MD, MPH
C. David Naylor, MD, DPhil
for the Evidence-Based Medicine Working Group

**Clinicians rely on knowledge about the clinical manifestations of disease to make clinical diagnoses. Before using research on the frequency of clinical features found in patients with a disease, clinicians should appraise the evidence for its validity, results, and applicability. For validity, 4 issues are important—how the diagnoses were verified, how the study sample relates to all patients with the disease, how the clinical findings were sought, and how the clinical findings were characterized. Ideally, investigators will verify the presence of disease in study patients using credible criteria that are independent of the clinical manifestations under study. Also, ideally the study patients will represent the full spectrum of the disease, undergo a thorough and consistent search for clinical findings, and these findings will be well characterized in nature and timing.**

**The main results of these studies are expressed as the number and percentages of patients with each manifestation. Confidence intervals can describe the precision of these frequencies. Most clinical findings occur with only intermediate frequency, and since these frequencies are equivalent to diagnostic sensitivities, this means that the absence of a single finding is rarely powerful enough to exclude the disease. Before acting on the evidence, clinicians should consider whether it applies to their own patients and whether it has been superseded by new developments. Detailed knowledge of the clinical manifestations of disease should increase clinicians' ability to raise diagnostic hypotheses, select differential diagnoses, and verify final diagnoses.**

*JAMA. 2000;284:869-875* www.jama.com

### CLINICAL SCENARIO

You are a general internist working in a teaching hospital paged to the emergency department to evaluate a 58-year-old man with new-onset pain in his chest and back. On the way to the emergency department, you think of myocardial ischemia as your leading hypothesis and you wonder whether aortic dissection should be actively considered in this patient.

In the emergency department, the patient describes to you the sudden onset of severe pain in the center of his chest radiating to his neck and mid back. He has long-standing hypertension, for which he takes a diuretic. You find a normal thoracic wall, clear lungs, equal pulses, a diastolic murmur of aortic regurgitation, and diastolic hypotension with blood pressure of 162/56 mm Hg. The electrocardiogram shows left ventricular hypertrophy but no signs of ischemia or infarction. The first set of cardiac enzyme levels is normal. The portable chest radiograph shows widening of the mediastinum. An arterial blood gas evaluation shows mild respiratory alkalosis and normal oxygenation. By now, your suspicion of acute aortic dissection has grown, so you arrange definitive testing for this diagnosis and consult with the cardiothoracic surgical team, after explaining the situation to the patient and family.

While you wait for the test results, the resident in the emergency department

**Author Affiliations:** Departments of Ambulatory Care and Research, South Texas Veterans Health Care System and Medicine, University of Texas Health Sciences Center at San Antonio, San Antonio (Drs Richardson and Williams); Department of Medicine, Wake-Forest University School of Medicine, Winston-Salem, NC (Dr Wilson); Departments of Pediatrics and Internal Medicine and the Center for Evidence-Based Medicine and Population Health, the University of Texas Health Sciences Center at Houston (Dr Moyer); and Department of Medicine and Office of the Dean, Faculty of Medicine, University of Toronto, Ontario (Dr Naylor). The original list of members (with affiliations) appears in the first article of the series (*JAMA*. 1993; 270:2093-2095). A list of new members appears in the 10th article of the series (*JAMA*. 1996;275:1435-1439). The following members of the Evidence-Based Working Group contributed to this article: Eric Bass, MD, MPH, Gordon H. Guyatt, MD, MSc, Les Irwig, MBBCh, PhD, and Hui Lee, MD, MSc.

**Reprints:** Gordon H. Guyatt, MD, MSc, Department of Clinical Epidemiology and Biostatistics, Room 2C12, 1200 Main St W, McMaster University Faculty of Health Sciences, Hamilton, Ontario, Canada L8N 3Z5.

**Users' Guides to the Medical Literature Section Editor:** Drummond Rennie, MD, Deputy Editor.

asks you about this patient and whether aortic dissection really needs to be actively considered. Together, you review the findings found useful in determining whether a patient is having a myocardial infarction[1] and then discuss the clinical findings seen with aortic dissection. The resident asks whether the normal pulses and equal blood pressures in the arms can rule out dissection without further testing. You reply, "I don't know. If we knew the frequencies of the clinical findings in aortic dissection, we could better interpret our examination and select his differential diagnosis. Rather than guess, why don't we look this up while we wait for his test results?"

## THE SEARCH

You begin by articulating your knowledge gap as a question: "In patients with confirmed acute aortic dissection, how frequently would a detailed and careful evaluation yield each of several clinical findings, such as pain radiating to the back, pulse asymmetry, diastolic hypotension, or diastolic murmur?" You turn to a networked computer in the emergency department that gives you full access to MEDLINE from the hospital's library, which you search using strategies reviewed elsewhere.[2,3] In the MEDLINE file since 1966, you combine medical subject headings *aneurysm, dissecting* (5027 citations) and *aortic aneurysm, thoracic* (1699 citations) with *aortic dissection* as a text word (2330 citations) to yield a set of 6410 citations. Next, you use the floating subheadings *di* for diagnosis (applied to articles that include clinical findings from patient examination) and *co* for complications (indicates conditions that coexist or follow the specified disease process). Combining these sets yields 86 citations, which drops to 33 when you limit to adult patients and to the English language. Scrolling through these titles, you find a relevant citation by Spittell et al[4] that is linked to the full text online in your library.

## UNDERSTANDING CLINICAL MANIFESTATIONS

In busy clinical practice, diagnosis is our daily bread. As we see sick persons, we classify their illnesses as instances or cases of disease,[5-11] to serve them by using the available knowledge about what is wrong, what it may mean, and what might be done to maximize their well-being.[9-11] To categorize illnesses, we use a classification system, or taxonomy of disease, with diseases representing the classes into which illnesses are grouped.[5-7] These taxonomic categories are generally defined by similarities in the illnesses of afflicted persons, including similarities of clinical features, anatomic abnormalities, physiologic derangements, causative microorganisms, or genetic and molecular lesions.

If we are to classify our patients' illnesses into diseases, we need to know the features by which different diseases are recognized and discriminated. In other words, we need to know the clinical manifestations of each disease that we expect to diagnose. We use the terms *clinical findings* and *clinical manifestations* interchangeably to mean findings that the clinician can gather directly from the patient, during the medical interview or the physical examination (we find less useful a rigid distinction between symptoms and signs).[6]

How specifically can we use knowledge of the clinical manifestations of disease for clinical diagnosis? First, when initially evaluating a patient's illness, single findings or clusters of findings can cue us to raise diagnostic hypotheses. In the clinical scenario, the sudden (rather than crescendo) onset of pain and the radiation of the pain to the back triggered the hypothesis of aortic dissection. Thus, when we recognize that a patient's illness includes features seen in a given disease, we "activate" that diagnostic possibility for further inquiry. Without such knowledge, the clinical features will not cue hypotheses, so we may fail to consider the correct diagnosis.

Second, knowing the clinical manifestations of disease can help us when selecting a patient-specific differential diagnosis and when deciding whether to use further testing to actively exclude a disorder. In the clinical scenario, while some of the patient's features (chest pain and risk factors for coronary atherosclerosis) suggest myocardial ischemia, other features (pain onset and radiation) suggest aortic dissection, so you plan to pursue testing for both. Thus, while aortic dissection is less common than myocardial ischemia, it is serious and treatable, so the presence of some of its features in this patient has led you to place dissection on your short list of active alternatives to be excluded.[12] In general, when considering an uncommon disease, experienced clinicians use the presence of 1 or more of its clinical manifestations, combined with knowledge of disease probability, prognosis, and responsiveness to treatment, to help them decide whether to actively consider this condition along with more common diseases. With incomplete or inaccurate knowledge of the clinical manifestations of diseases, we risk selecting flawed differential diagnoses.

Third, after diagnostic testing is completed and interpreted, we can use the clinical manifestations of disease in verifying a patient's final diagnosis.[13] Before concluding that a diagnosis is correct, we (often implicitly) test how well it explains the patient's illness, compared with the alternative possibilities. As shown more explicitly in TABLE 1, verifying a patient's final diagnosis depends heavily on detailed knowledge of the clinical manifestations of disease. While ideally a final diagnosis should explain that all the patient's findings should be coherent with the patient's observed pathophysiologic state, the best fit among the alternatives, the simplest explanation overall, the only possibility not yet disproved, and the 1 hypothesis that best predicts the patient's course, in actual practice, we often accept diagnoses that meet only some of these considerations. If our knowledge of the clinical manifestations of disease is inaccurate, we risk prematurely accepting an incorrect diagnosis or pursuing further testing despite good verification of the correct diagnosis.

What lessons can we learn from the frequencies of clinical manifestations of disease? First, textbook descriptions of disease may emphasize the presence of

classic findings that are hallmarks of the diagnosis. Yet when studied systematically, such manifestations may be uncommon, and if we were to rely on their presence to diagnose the disorder, we would miss many cases. For example, hemoptysis has been described as a hallmark of acute pulmonary embolism, yet when 327 patients with angiographically proven pulmonary emboli were examined, only 30% were found to have hemoptysis.[14] Second, the reverse lesson can be learned, because some manifestations may be more common than usually believed. For instance, the murmur of aortic regurgitation was found in 40 of 124 patients with confirmed aortic dissection, suggesting that clinicians should purposefully seek this finding in suspected cases.[15] Similar to these examples, most findings occur with intermediate frequencies. Since these frequencies are equivalent to diagnostic sensitivities, these intermediate values mean that individually, most findings cannot rule out disease. Since specificities or likelihood ratios cannot be obtained from studies of the clinical manifestations of disease, we are unable to revise our estimates of disease probability using these findings alone. The third lesson represents the exception to this general rule. A few manifestations of disease might be so common that they occur in virtually all diseased patients. As the proportion of diseased patients with a similar finding nears 100%, the absence of this finding becomes powerful for excluding the disease. This is because as the sensitivity goes to 100%, the false-negative rate approaches 0, effectively ruling out the disorder.[16-18]

How does the knowledge about clinical manifestations of diseases fit with other knowledge for use in diagnostic thinking? Expert diagnosticians that we have known or have read about appear to have detailed knowledge of 4 kinds: (1) remembered cases of real patients they have cared for; (2) knowledge of clinical problems, including which diseases cause them and how likely those are; (3) knowledge of the accuracy and precision of test results; and (4) knowledge of the clinical manifestations of diseases.[19,20] They can draw on this extensive knowledge as they proceed through the diagnostic steps of raising diagnostic possibilities, selecting a patient-specific differential diagnosis, choosing and interpreting diagnostic tests, and verifying a patient's final diagnosis. These 4 forms of knowledge complement each other, and no single form can replace the others for their intended uses. Knowledge of the probability of diseases that cause a clinical problem is particularly useful for selecting a patient's differential diagnosis and estimating pretest probability.[12,18] Knowledge of the likelihood ratios of test results is most useful for choosing and interpreting diagnostic tests and estimating posttest probability.[16-18] Knowledge of the clinical manifestations of disease is useful for raising diagnostic possibilities, selecting differential diagnoses, and verifying a patient's final diagnosis. In an archery analogy, if pretest probability is how we aim our arrows and the power of diagnostic tests is the strength of our bow, our disease taxonomy (based on clinical manifestations) contains the targets we shoot toward.

Where can we find knowledge about the frequencies of the clinical manifestations of disease? One source is from clinical experience, either our own or of others.[19-21] Here, we focus on the other major source of this knowledge, the medical literature, eg, the article about aortic dissection retrieved by the search.[4] This Users' Guide will help you understand articles about the clinical manifestations of disease, judge their validity, and decide whether to use them in refining your disease taxonomy for clinical diagnoses (TABLE 2).

Before doing that, it is important to be clear about what these articles cannot do. First, studies of the clinical manifestations of a disorder generally include patients only if they are known to have that specific disorder and exclude patients with other diseases. This means that such studies cannot provide evidence about how well the clinical findings discriminate between diseases, such as through likelihood ratios for these findings.[16-18] Second, since the study sample includes patients with only 1 disorder, studies of the clinical manifestations of disease cannot provide evidence about the probability of different diseases in patients with a given clinical problem.[12] Third, studies of the clinical manifestations of disease generally do not provide informa-

**Table 1.** Explicit Tests for Verifying a Patient's Diagnosis

| |
|---|
| Adequacy |
| • Does this diagnostic hypothesis adequately explain all the patient's clinical findings? |
| • If not, does it explain the patient's important findings? |
| Coherence |
| • Does this diagnostic hypothesis fit the pathophysiologic state observed and/or inferred in this patient? |
| • Thus, is this hypothesis pathophysiologically coherent? |
| Primacy |
| • Does this diagnostic hypothesis provide the best fit to the pattern of the patient's illness? |
| • Is there no hypothesis that fits the patient's illness better? |
| Parsimony |
| • Is this diagnostic hypothesis the simplest explanation of this patient's illness? |
| • Is there no hypothesis that is simpler? |
| Robustness |
| • Is this diagnostic hypothesis robust to attempts to falsify it? |
| • Has it escaped disproof? |
| Prediction |
| • Does this diagnostic hypothesis best predict the subsequent course of the patient's illness? |
| • Is there no hypothesis that predicts the patient's course better? |

**Table 2.** Users' Guides for Articles on the Clinical Manifestations of Disease

| |
|---|
| Are the results of the study valid? |
| Primary guides: |
| • Was the presence of disease verified using credible criteria that are independent of the clinical manifestations under study? |
| • Did the patient sample represent the full spectrum of those with this disorder? |
| Additional guides: |
| • Were clinical manifestations sought thoroughly, carefully, and consistently? |
| • Were the clinical manifestations classified by when and how they occurred? |
| What were the results? |
| • How frequent were the clinical manifestations of disease? |
| • How precise were the estimates of frequency? |
| • When and how did these clinical manifestations occur in the course of disease? |
| Will these results help me in caring for my patients? |
| • Are the study patients similar to my own? |
| • Is it unlikely that the disease manifestations have changed since this evidence was gathered? |

tion about how reliably clinicians gather these findings.[22,23]

## THE GUIDES

### Are the Results Valid?

**Was the Presence of Disease Verified Using Credible Criteria That Are Independent of the Clinical Manifestations Under Study?** This question addresses 2 closely linked issues. First, how sure are investigators that the study patients really did have this particular disease to explain their illnesses and not other diseases? While clinicians often encounter tentative diagnoses in practice, in a research study such diagnostic uncertainty could introduce bias, because the patient sample might include not only patients with this disease but also other diseases. To minimize this bias, investigators can use a set of explicit diagnostic criteria and include in the study sample only patients who meet these criteria. Ideally, for every disease there would be a set of widely accepted diagnostic criteria, including 1 or more well-established reference standard tests that can be applied reproducibly in a blinded fashion. Reference standards can be anatomic, physiologic, radiographic, or genetic, to name a few. To judge how the presence of disease was verified, look for which standards were used, how they were used, and whether the standards are clinically credible.

Second, are the diagnostic criteria independent of the clinical manifestations under study? When no reference standards exist, investigators' degree of diagnostic certainty is much lower. In these situations, known sometimes as *syndrome diagnosis*,[5] diagnostic criteria still can be made and used. They usually comprise a list of clinical features that must be present for the diagnosis to be made. For instance, the definition of chronic fatigue syndrome uses an explicit set of clinical features as diagnostic criteria.[24] Such explicit criteria often represent an advance over an implicit haphazard approach and for a time may be the best available method for clinical diagnosis.

However, trouble can arise when investigators use clinical manifestations to make the syndrome diagnosis, select the patient sample, and then examine the frequency of these same clinical findings in the study patients. This testing of manifestations that are incorporated into the definition creates circular reasoning that can bias upward the frequencies of these findings in the study sample, known as *incorporation bias*. For example, in a study of manifestations among 36 patients with relapsing polychondritis, the investigators used diagnostic criteria based on several characteristic clinical findings.[25] Although this study may be the best available method for clinical diagnosis, incorporation bias is inevitable and it limits the inferences we can draw about the frequency of manifestations. In judging the independence of verifying criteria, compare the list of these criteria with the list of clinical manifestations studied to examine for overlap.

Spittell et al[4] studied 235 patients whose aortic dissections were confirmed by surgical intervention (n=162), autopsy (n=27), or radiographic studies (n=47). Thus, the diagnoses of study patients appear to have been verified using clinically credible means that are independent of the clinical manifestations.

**Did the Patient Sample Represent the Full Spectrum of Those With This Disorder?** By selecting a specific disease for research, the investigators determine the population from which the study patients should be selected. Ideally, the study sample mirrors the whole population of those with the disease, so that the frequency of clinical manifestations in the sample approximates that of the population. Such a patient sample is termed *representative*, and the more representative the sample is, the more accurate the resulting frequencies of clinical findings. Conversely, the less representative the study sample, the less confident we can be that the frequencies of clinical manifestations found are accurate.[26]

To judge the representativeness of the study sample, we suggest 3 tactics. First, examine the setting from which study patients come. Patients seen in referral care settings might have higher proportions of unusual findings or illnesses difficult to diagnose, yielding different frequencies of clinical manifestations than patients in community practice.[27] Second, examine the methods the investigators used to identify and include the study patients and exclude others. Were all the important demographic groups (age, sex, race, etc) included? Were any important subgroups excluded that would threaten the validity of the results? Third, examine the description of the study patients' illnesses. Are patients with mild, moderate, and severe symptoms present? If different clinical patterns of disease are known, does the sample include patients with each pattern?

Combining these 3 considerations, you can judge whether the spectrum of included patients is full enough that the study can yield valid results about clinical manifestations of this disease. For instance, in a study of patients with thyrotoxic periodic paralysis, the investigators included in the sample only the 19 patients who were hospitalized during an episode of paralysis, excluding 11 patients who were diagnosed during the study period but who were not admitted.[28] To the extent that hospitalized patients may have worse or different clinical manifestations than those not admitted, such a restriction might introduce bias into the study.

Investigators may deliberately choose the task of describing the manifestations of a disease in a purposefully narrowed target population, whether demographic (eg, a study of the findings of myocardial infarction in the aged[29]), prognostic (eg, a study of the clinical findings in patients with fatal pulmonary embolism[30]), or by site of care (eg, a study of the findings in patients with ruptured abdominal aortic aneurysm who present to internists, not emergency departments[31]). In such situations, you can look to see whether the study sample is representative of the limited target population.

Spittell et al[4] reported a study of patients treated at the Mayo Clinic, which provides both community hospital care

and tertiary referral care. The study sample had patients with aortic dissection that was both acute (<2 weeks) in 158 patients (67%) and chronic (≥2 weeks) in 78 patients (33%). In 60 patients, the initial clinical impression was a diagnosis other than aortic dissection. The sample included patients with sudden death, including 10 out-of-hospital cardiac arrests and 5 in-hospital cardiac arrests. It also included 11 patients without pain but with other symptoms, along with 33 patients without pain or other symptoms who had abnormal chest radiograph findings. Thus, the study patients had a wide array of clinical presentations and may be sufficiently representative of the full spectrum of this disorder.

**Were Clinical Manifestations Sought Thoroughly, Carefully, and Consistently?** This criterion addresses 3 closely related issues. First, were study patients evaluated thoroughly enough to detect clinical findings if they were present? Within reason, the more comprehensive the workup, the lower the chance of missing findings and drawing invalid conclusions about their frequency. Second, how did the investigators ensure that the information they gathered was correct and free of distortion? Were symptoms inquired about in neutral nonjudgmental ways? Were patients examined by skilled examiners? The more carefully the data were gathered, the more credible the resulting frequencies will be. Third, how consistently was the evaluation carried out? Inconsistent assessments might yield erroneous frequencies of disease manifestations.

You may find it relatively easy to judge the thoroughness, care, and consistency of the search for manifestations when the patients were evaluated prospectively using a standardized diagnostic approach. It becomes harder to judge when patients were studied retrospectively after their investigation was complete or when the evaluation was not standardized. For example, in a retrospective analysis of disease manifestations in 68 patients with lumbar spinal stenosis, the investigators do not describe the search for clinical findings in enough detail for us to judge how well they protected against biased ascertainment.[32] Ordinarily, a prospective study of clinical manifestations of disease will provide more credible results than a retrospective study.

Spittell et al[4] retrospectively reviewed the charts of their patients after the clinical evaluations were completed. The diagnostic workup of these patients is not described explicitly. The tables of results include much detail about the clinical examination, suggesting a careful approach, but uncertainty remains about whether the investigators avoided bias during workup.

**Were the Clinical Manifestations Classified by When and How They Occurred?** Clinical manifestations of disease can range from the permanent to the fleeting. They can occur early, late, or throughout the course of the disease. The most complete information about the timing of disease manifestations might be obtained if the investigators began collecting data the instant the disease starts in each patient and continued collecting through the end of the illness. Since knowing this "zero time" with certainty is impossible for most diseases, investigators can use the next strongest approach, that of targeting all findings that occur from the onset of patients' first symptoms of this illness episode. Studies that do not start collecting at the beginning of the episode, or that do not report the timing of evaluation relative to symptom onset, may have inadvertently missed findings, and our confidence in their validity decreases. For instance, in a study of the clinical manifestations in 92 patients with fatal pulmonary embolism, investigators recorded findings for just the 24 hours before death, so they may have missed transient but important clues to the diagnosis that occurred before then.[30]

Studies of this type also can describe qualitative findings that are useful in clinical diagnosis, particularly when triggering initial diagnostic hypotheses. For instance, the pain of aortic dissection is often described as a tearing or ripping sensation that is located in the center of the torso and reaches maximal intensity quite quickly.[15] Just as with the temporal aspects, these qualitative descriptions are more credible if they were gathered deliberately and carefully.

Spittell et al[4] describe the clinical manifestations of dissection at presentation for patients with both acute and chronic aortic dissection. They also describe the location of pain in relation to the site of dissection, the various clusters of pain with other findings, along with unusual findings such as hoarseness and dysphagia. Thus, despite the retrospective design, the investigators appear to have classified the temporal and qualitative features accurately enough to provide valid results for patients with acute dissection. We may be less confident in the results for chronic dissection, since early findings might have been missed.

## What Were the Results?

**How Frequent Were the Clinical Manifestations of Disease?** Studies of clinical manifestations of disease often display the main results in a table listing the clinical findings, along with the number and percentages of patients with each of those manifestations. Since patients usually have more than 1 finding, these proportions are not mutually exclusive. Some studies also report the number of patients with any of the findings, either in total or by particular group.

Spittell et al[4] report that 168 patients (74%) initially had acute onset of severe pain, 35 (15%) were asymptomatic but had abnormal chest radiograph findings, and 15 (6.3%) experienced cardiac arrest or sudden death. Of the 235 patients, 217 (92.3%) had a cardiac examination recorded; 22 (11%) had murmurs of aortic regurgitation detected. Pulse deficits were uncommon, occurring in 14 (6%) patients. Thus, the diagnostic sensitivity of pulse deficit is only 6%, so that using pulse deficits to exclude dissection would lead to missing 94% of cases.

**How Precise Were These Estimates of Frequency?** Even when valid, these measured frequencies of findings are only estimates of the true frequen-

cies. You can examine the precision of these estimates using their confidence intervals (CIs). If the authors do not provide the CIs for you, you can calculate 95% CIs with the following formula:

$$95\%\ \text{CI} = p \pm 1.96 \times \sqrt{(p\,[1-p])/n}$$

Here $p$ is the proportion of patients with the finding of interest, and $n$ is the number of patients in the sample.[33] This formula becomes inaccurate when the number of cases is 5 or fewer, so approximations have been developed for this situation.[34,35]

For instance, consider the clinical finding of pulse deficit, found in 14 of the 217 patients in whom it was sought by Spittell et al.[4] Using the above formula, we would start with $p = 0.06$, $(1 - p) = 0.94$, and $n = 217$; this yields a CI of 0.06 +/- 0.03. Thus, the most likely frequency of pulse deficit is 6%, and it may range between 3% and 9%.

Whether you consider the CIs sufficiently precise depends on how you expect to use the information. For example, for a finding that occurs in 50% of cases, you might examine for it but not plan to use its absence to exclude the diagnosis. If the CI for this estimate ranged from 30% to 70%, it would not change your expected use of the information, so the result may be precise enough. On the other hand, for a finding that occurs in 98% of patients, you might hope to use its absence to help you rule out the diagnosis. If the CI for this estimate ranged from 80% to 100% (half of the prior 40-point range), it could mean that using this finding to exclude the diagnosis might lead you to miss up to 20% of patients. Such a result would be too imprecise to rule out this disorder.

**When and How Did These Clinical Manifestations Occur in the Course of Disease?** Research on the clinical manifestations of disease can yield additional insights beyond the frequency of findings. Some studies will report on the temporal sequence of symptoms, characterizing symptoms as *presenting*, prompted patients to seek care; *concurring*, did not prompt care but were present initially; or *eventual*, not present initially, but found subsequently. For instance, in 100 patients with pancreatic cancer, investigators described weight loss and abdominal pain as presenting manifestations in 75 and 72 patients, respectively, while jaundice, commonly taught as a key presenting sign, was found in only 24 patients.[36] In addition to chronology, such studies can also describe the location, quality, intensity, aggravating and alleviating factors, situational context, and associated findings for important manifestations.

Spittell et al[4] describe in detail the symptoms at initial assessment, both as individual findings and in clusters (their Tables 3, 6, and 7). The authors also describe the location of pain and its association with the site of dissection (their Tables 4 and 5). The delayed manifestations are not described in much detail.

### Will the Results Help Me in Caring My Patients?

**Are the Study Patients Similar to My Own?** This question is about whether the clinical setting and patient characteristics are similar enough to yours to allow you to extrapolate the results to your practice. The closer the match, the more confident you can be in applying the results. Ask yourself whether the setting or the patients are so different from yours that you cannot use the results.[37] Do your patients come from a geographic, demographic, cultural, or clinical group that you would expect to differ importantly in the ways in which this particular disorder is expressed? For instance, the presenting symptoms of acute myocardial infarction were found to differ with advancing patient age, when studied in 777 elderly hospitalized patients; syncope, stroke, and acute confusion were more common and were sometimes the sole presenting symptom.[29]

Spittell et al[4] studied patients who were seen at the Mayo Clinic with aortic dissection. The referral filters through which patients arrived are not described, although you know that Mayo provides community hospital care for Olmsted County residents along with referred care for others. Of the 235 patients, 158 (67%) were men, like your patient. The study patients ranged in age from 17 to 94 years, with a mean age very close to your patient. The patients are not described with respect to comorbid conditions, socioeconomic status, race, or cultural background. Thus, while some uncertainty remains, these patients are sufficiently similar to the patient in the scenario that the results could be extrapolated.

**Is It Unlikely That the Disease Manifestations Have Changed Since This Evidence Was Gathered?** As time passes, evidence about the clinical manifestations of disease can become obsolete. New diseases can arise and old diseases can present in new ways. New disease taxonomies can be built, changing the borders between disease states. Such events can so alter the clinical manifestations of disease that previously valid studies may no longer be applicable to current practice. For example, consider how much the arrival of human immunodeficiency virus disease has changed our concept of pneumonia caused by *Pneumocystis carinii*.[38,39]

Similar changes can occur as the result of progress in health science or medical practice. For instance, early descriptions of *Clostridium difficile* infection emphasized severe cases of life-threatening colitis. As diagnostic testing improved and awareness of the infection widened, milder cases were documented and a broader variety of presenting manifestations was recognized.[40] Treatment advances can change the course of disease so that previously common clinical manifestations might become less frequent. Also, new treatments bring the chance of new iatrogenic disease, which may combine with underlying diseases in new ways.

The study by Spittell et al[4] was published in 1993 and reports on patients seen from 1980 to 1990. You know of no new diseases arising since then that would change the clinical features of dissection. Both testing for suspected dissection and treatment for hypertension (major risk factor for dissection)

have changed during this period, but you expect they would not change the presenting clinical features of acute dissection.

## RESOLUTION OF THE SCENARIO

Based on the evidence from Spittell et al,[4] you and the resident agree not to use the absence of pulse deficit to rule out aortic dissection. Given the presence of the aortic regurgitation murmur and the diastolic hypotension, along with the patient's known risk and the absence of findings for myocardial infarction, the resident now agrees with your suspicion of dissection. When completed, this patient's aortogram confirms aortic dissection of the ascending aorta and arch, complicated by aortic regurgitation.

We recommend applying these Users' Guides to identify good evidence about the clinical manifestations of disease. As you do so, this detailed knowledge of the clinical findings of disease should increase your ability to raise diagnostic hypotheses, select differential diagnoses, and verify your final diagnoses.

While this article was in press, another study of the clinical manifestations of this disease was published, based on 464 patients with acute aortic dissection collected from 12 international referral centers.[41] Overall, the frequencies of clinical findings were similar; for instance, pulse deficit was found in 15.1% and diastolic murmur in 31.6%.

**Funding/Support:** Dr Williams is a Veterans Affairs Health Services Research & Development Career Development Awardee.

**Disclaimer:** The views expressed in this article are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs.

### REFERENCES


1. Panju AA, Hemmelgarn BR, Guyatt GH, Simel DL. Is this patient having a myocardial infarction? *JAMA*. 1998;280:1256-1263.
2. Hunt DL, Jaeschke R, McKibbon KA, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, XXI: using electronic health information resources in evidence-based practice. *JAMA*. 2000;283:1875-1879.
3. McKibbon KA, ed. *PDQ Evidence-Based Principles and Practice*. Hamilton, Ontario: BC Decker; 1999.
4. Spittell PC, Spittell JA, Joyce JW, et al. Clinical features and differential diagnosis of aortic dissection: experience with 236 cases (1980-1990). *Mayo Clin Proc*. 1993;68:642-651.
5. Wulff HR. *Rational Diagnosis and Treatment: An Introduction to Clinical Decision-Making*. 2nd ed. Oxford, England: Blackwell Scientific Publications; 1981.
6. King LS. *Medical Thinking: An Historical Preface*. Princeton, NJ: Princeton University Press; 1982.
7. Murphy EA. *The Logic of Medicine*. 2nd ed. Baltimore, Md: Johns Hopkins University Press; 1997.
8. Flegel KM. The case for "a case of . . ." [editorial]. *CMAJ*. 1997;157:286.
9. Glass RD. *Diagnosis: A Brief Introduction*. New York, NY: Oxford University Press; 1996.
10. Barondess JA, Carpenter CCJ, eds. *Differential Diagnosis*. Philadelphia, Pa: Lea & Febiger; 1994.
11. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical Epidemiology: A Basic Science for Clinical Medicine*. 2nd ed. Boston, Mass: Little Brown & Co; 1991:4-5.
12. Richardson WS, Wilson MC, Guyatt GH, Cook DJ, Nishikawa J, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, XV: how to use an article about disease probability for differential diagnosis. *JAMA*. 1999;281:1214-1219.
13. Kassirer JP, Kopelman RI. *Learning Clinical Reasoning*. Baltimore, Md: Williams & Wilkins; 1991:32-33.
14. Bell WR, Simon TL, DeMets DL. The clinical features of submassive and massive pulmonary emboli. *Am J Med*. 1977;62:355-360.
15. Slater EE, DeSanctis RW. The clinical recognition of dissecting aortic aneurysm. *Am J Med*. 1976;60:625-633.
16. Jaeschke R, Guyatt GH, Sackett DL, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, III: how to use an article about a diagnostic test, A: are the results valid? *JAMA*. 1994;271:389-391.
17. Jaeschke R, Guyatt GH, Sackett DL, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, III: how to use an article about a diagnostic test, B: what are the results and will they help me in patient care? *JAMA*. 1994;271:703-707.
18. Sackett DL, Straus SE, Richardson WS, Rosenberg WMC, Haynes RB, eds. *Evidence-Based Medicine: How To Practice and Teach EBM*. 2nd ed. Edinburgh, Scotland: Churchill Livingstone; 2000.
19. Schmidt HG, Norman GR, Boshuizen HPA. A cognitive perspective on medical expertise: theory and implications. *Acad Med*. 1990;65:611-621.
20. Bordage G. Elaborated knowledge: a key to successful diagnostic thinking. *Acad Med*. 1994;69:883-885.
21. Regehr G, Norman GR. Issues in cognitive psychology: implications for professional education. *Acad Med*. 1996;71:988-1001.
22. Department of Clinical Epidemiology and Biostatistics. Clinical disagreement, I: how often it occurs and why. *CMAJ*. 1980;123:499-504.
23. Department of Clinical Epidemiology and Biostatistics. Clinical disagreement, II: how to avoid it and learn from one's mistakes. *CMAJ*. 1980;123:613-617.
24. Fukuda K, Straus SE, Hickie I, Sharpe MC, Dobbins JG, Komaroff A, and the International Chronic Fatigue Syndrome Study Group. The chronic fatigue syndrome: a comprehensive approach to its definition and study. *Ann Intern Med*. 1994;121:953-959.
25. Trentham DE, Le CH. Relapsing polychondritis. *Ann Intern Med*. 1998;129:114-122.
26. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med*. 1978;299:926-930.
27. Fletcher RH, Fletcher SW, Wagner EH. *Clinical Epidemiology: The Essentials*. 3rd ed. Baltimore, Md: Williams & Wilkins; 1996.
28. Manoukian MA, Foote JA, Crapo LM. Clinical and metabolic features of thyrotoxic periodic paralysis in 24 episodes. *Arch Intern Med*. 1999;159:601-606.
29. Bayer AJ, Chadha JS, Farag RR, Pathy MS. Changing presentation of myocardial infarction with increasing age. *J Am Geriatr Soc*. 1986;34:263-266.
30. Morgenthaler TI, Ryu JH. Clinical characteristics of fatal pulmonary embolism in a referral hospital. *Mayo Clin Proc*. 1995;70:417-424.
31. Lederle FA, Parenti CM, Chute EP. Ruptured abdominal aortic aneurysm: the internist as diagnostician. *Am J Med*. 1994;96:163-167.
32. Hall S, Bartleson JD, Onofrio BM, Baker HL, Okazaki H, O'Duffy JD. Lumbar spinal stenosis: clinical features, diagnostic procedures, and results of surgical treatment in 68 patients. *Ann Intern Med*. 1985;103:271-275.
33. Altman DG. Confidence intervals [appendix]. In: Sackett DL, Straus SE, Richardson WS, Rosenberg WMC, Haynes RB, eds. *Evidence-Based Medicine: How to Practice and Teach EBM*. 2nd ed. Edinburgh, Scotland: Churchill Livingstone; 2000:233-243.
34. Hanley JA, Lippman-Hand A. If nothing goes wrong, is everything all right? interpreting zero numerators. *JAMA*. 1983;249:1743-1745.
35. Newman TB. If almost nothing goes wrong, is almost everything all right? interpreting small numerators. *JAMA*. 1995;274:1013.
36. Gudjonsson B, Livistone EM, Spiro HM. Cancer of the pancreas: diagnostic accuracy and survival statistics. *Cancer*. 1978;42:2494-2506.
37. Glasziou P, Guyatt GH, Dans AL, Dans LF, Straus SE, Sackett DL. Applying the results of trials and systematic reviews to individual patients [editorial]. *ACP J Club*. 1998;129:A15-A16.
38. Walzer PD, Perl DP, Krogstad DJ, Rawson PG, Schultz MG. *Pneumocystis carinii* pneumonia in the United States: epidemiologic, diagnostic and clinical features. *Ann Intern Med*. 1974;80:83-93.
39. Kovacs JA, Hiemenz JW, Macher AM, et al. *Pneumocystis carinii* pneumonia: a comparison between patients with AIDS and patients with other immunodeficiency states. *Ann Intern Med*. 1984;100:663-671.
40. Caputo GM, Weitekamp MR, Bacon AE, Whitener C. *Clostridium difficile* infection: a common clinical problem for the general internist. *J Gen Intern Med*. 1994;9:528-533.
41. Hagan PG, Nienaber CA, Isselbacher EM, et al. The international registry of acute aortic dissection: new insights into an old disease. *JAMA*. 2000;283:897-903.

# Users' Guides to the Medical Literature: XXIV. How to Use an Article on the Clinical Manifestations of Disease

W. Scott Richardson; Mark C. Wilson; John W. Williams, Jr; et al.

*JAMA*. 2000;284(7):869-875 (doi:10.1001/jama.284.7.869)

http://jama.ama-assn.org/cgi/content/full/284/7/869

# Users' Guides to the Medical Literature

## XXV. Evidence-Based Medicine: Principles for Applying the Users' Guides to Patient Care

Gordon H. Guyatt, MD, MSc
R. Brian Haynes, MD, PhD
Roman Z. Jaeschke, MD, MSc
Deborah J. Cook, MD, MSc
Lee Green, MD, MPH
C. David Naylor, MD, PhD
Mark C. Wilson, MD, MPH
W. Scott Richardson, MD
for the Evidence-Based Medicine Working Group

**This series provides clinicians with strategies and tools to interpret and integrate evidence from published research in their care of patients. The 2 key principles for applying all the articles in this series to patient care relate to the value-laden nature of clinical decisions and to the hierarchy of evidence postulated by evidence-based medicine. Clinicians need to be able to distinguish high from low quality in primary studies, systematic reviews, practice guidelines, and other integrative research focused on management recommendations. An evidence-based practitioner must also understand the patient's circumstances or predicament; identify knowledge gaps and frame questions to fill those gaps; conduct an efficient literature search; critically appraise the research evidence; and apply that evidence to patient care. However, treatment judgments often reflect clinician or societal values concerning whether intervention benefits are worth the cost. Many unanswered questions concerning how to elicit preferences and how to incorporate them in clinical encounters constitute an enormously challenging frontier for evidence-based medicine. Time limitation remains the biggest obstacle to evidence-based practice but clinicians should seek evidence from as high in the appropriate hierarchy of evidence as possible, and every clinical decision should be geared toward the particular circumstances of the patient.**

*JAMA. 2000;284:1290-1296* www.jama.com

### CLINICAL SCENARIO

A senior resident, a junior attending, a senior attending, and an emeritus professor were discussing evidence-based medicine (EBM) over lunch in the hospital cafeteria.

"EBM," announced the resident with some passion, "is a revolutionary development in medical practice." She went on to describe EBM's fundamental innovations in solving patient problems.

"A compelling exposition," remarked the emeritus professor.

"Wait a minute," the junior attending exclaimed, also with some heat, and presented an alternative position stating that EBM merely provided a set of additional tools for traditional approaches to patient care.

"You make a strong and convincing case," the emeritus professor commented.

"Wait a minute," the senior attending exclaimed to her older colleague, "their positions are diametrically opposed. They can't both be right."

The emeritus professor looked thoughtfully at the puzzled physician and, with the barest hint of a smile, replied, "Come to think of it, you're right too."

**See also Patient Page.**

### INTRODUCTION

Evidence-based medicine, the approach to clinical care that underlies the 24 Users' Guides to the Medical Literature, which *JAMA* has published during the last 8 years,[1] is about solving clinical problems. The Users' Guides

**Author Affiliations:** Departments of Clinical Epidemiology and Biostatistics (Drs Guyatt, Haynes, and Cook) and Medicine (Drs Haynes and Jaeschke), McMaster University, Hamilton, Ontario; Department of Medicine and Office of the Dean, Faculty of Medicine, University of Toronto, Ontario (Dr Naylor); Department of Family Medicine, University of Michigan, Ann Arbor (Dr Green); Department of Medicine, Wake-Forest University School of Medicine, Winston-Salem, NC (Dr Wilson); and Departments of Ambulatory Care and Research, South Texas Veterans Health Care System and Medicine, University of Texas Health Sciences Center, San Antonio (Dr Richardson).

The original list of members with affiliations appears in the first article of the series (*JAMA*. 1993; 270:2093-2095). A list of new members appears in the 10th article of the series (*JAMA*. 1996;275: 1435-1439). The following member of the Evidence-Based Medicine Working Group contributed to this article: Anne Holbrook, MD, MSc.

**Corresponding Author and Reprints:** Gordon H. Guyatt, MD, MSc, Department of Clinical Epidemiology and Biostatistics, Room 2C12, 1200 Main St W, McMaster University Faculty of Health Sciences, Hamilton, Ontario, Canada L8N 3Z5.

**Users' Guides to the Medical Literature Section Editor:** Drummond Rennie, MD, Deputy Editor.

provide clinicians with strategies and tools to interpret and integrate evidence from published research in their patient care. As we developed the Users' Guides, our understanding of EBM has evolved. In this article, since we are addressing physicians, we use the term *EBM* but what we report applies to all clinical care provisions and the rubric "evidence-based health care" is equally appropriate.

In 1992, in an article that provided a background to the Users' Guides, we described EBM as a shift in medical paradigms.[2] In contrast to the traditional paradigm, EBM acknowledges that intuition, unsystematic clinical experience, and pathophysiologic rationale are insufficient grounds for clinical decision making, and stresses the examination of evidence from clinical research. The philosophy underlying EBM suggests that a formal set of rules must complement medical training and common sense for clinicians to effectively interpret the results of clinical research. Finally, EBM places a lower value on authority than the traditional paradigm of medical practice.

While we continue to find the paradigm shift a valid way of conceptualizing EBM, as the scenario suggests, the world is often complex enough to invite more than 1 useful way of thinking about an idea or a phenomenon. In this article, we describe the 2 key principles that clinicians must grasp to be effective practitioners of EBM. One of these relates to the value-laden nature of clinical decisions; the other to the hierarchy of evidence postulated by EBM. We will also comment on additional skills necessary for optimal clinical practice and we conclude with a discussion of the challenges facing EBM in the new millennium.

## TWO FUNDAMENTAL PRINCIPLES OF EBM

An evidence-based practitioner must be able to understand the patient's circumstances or predicament (including issues such as social supports and financial resources); to identify knowledge gaps, and frame questions to fill those gaps; to conduct an efficient literature search; to critically appraise the research evidence; and to apply that evidence to patient care.[3] The Users' Guides have dealt with the framing of the question in the scenarios, with searching the literature,[4] with appraising the literature in the "Validity" section, and with applying the evidence in the "Results" and "Applicability" sections. Underlying these steps are 2 fundamental principles. One, relating primarily to the assessment of validity, posits a hierarchy of evidence to guide clinical decision making. Another, relating primarily to the application of evidence, suggests that decision makers must always trade off the benefits and risks, inconvenience, and costs associated with alternative management strategies, and in doing so consider the patient's values.[5] In the sections that follow, we will discuss these 2 principles in detail.

### Clinical Decision Making: Evidence Is Never Enough

Picture a patient with chronic pain due to terminal cancer who has come to terms with her condition, has resolved her affairs and said her goodbyes, and wishes only palliative therapy. The patient develops pneumococcal pneumonia. The evidence that antibiotic therapy reduces morbidity and mortality due to pneumococcal pneumonia is strong. Almost all clinicians would agree that this strong evidence does not dictate that this patient receive antibiotics. Despite the fact that antibiotics might reduce symptoms and prolong the patient's life, her values are such that she would prefer a rapid and natural passing.

Picture a second patient, an 85-year-old severely demented man, incontinent, contracted and mute, without family or friends, who spends his day in apparent discomfort. This man develops pneumococcal pneumonia. While many clinicians would argue that those responsible for this patient's care should not administer antibiotic therapy because of his circumstances, others would suggest they should. Once again, evidence of treatment effectiveness does not automatically imply that treatment be administered. The management decision requires a judgment about the trade-off between risks and benefits, and because values or preferences differ, the best course of action will vary between patients and between clinicians.

Picture a third patient, a healthy 30-year-old mother of 2 children who develops pneumococcal pneumonia. No clinician would have any doubt about the wisdom of administering antibiotic therapy to this patient. This does not mean that an underlying value judgment has been unnecessary. Rather, our values are sufficiently concordant, and the benefits so overwhelm the risks that the underlying value judgment is unapparent.

In current health care practice, judgments often reflect clinician or societal values concerning whether intervention benefits are worth the cost. Consider the decisions regarding administration of tissue-type plasminogen activator vs streptokinase to patients with acute myocardial infarction, or clopidogrel vs aspirin to patients with transient ischemic attack. In both cases, evidence from large randomized controlled trials (RCTs) suggests the more expensive agents are, for many patients, more effective. In both cases, many authoritative bodies recommend first-line treatment with the less effective drug, presumably because they believe society's resources would be better used in other ways. Implicitly, they are making a value or preference judgment about the trade-off between deaths and strokes prevented, and resources spent.

By values and preferences, we mean the underlying processes we bring to bear in weighing what our patients and our society will gain or lose when we make a management decision. A number of the Users' Guides focus on how clinicians can use research results to clearly understand the magnitude of potential benefits and risks associated with alternative management strategies.[6-10] Three Users' Guides focused on the pro-

**Table 1.** A Hierarchy of Strength of Evidence for Treatment Decisions

| |
|---|
| N of 1 randomized trial |
| Systematic reviews of randomized trials |
| Single randomized trial |
| Systematic review of observational studies addressing patient-important outcomes |
| Single observational study addressing patient-important outcomes |
| Physiologic studies |
| Unsystematic clinical observations |

cess of balancing those benefits and risks when using treatment recommendations[11,12] and in making individual treatment decisions.[13] The explicit enumeration and balancing of benefits and risks brings the underlying value judgments involved in making management decisions into bold relief.

Acknowledging that values play a role in every important patient care decision highlights our limited understanding of eliciting and incorporating societal and individual values. Health economists have played a major role in developing a science of measuring patient preferences.[14,15] Some decision aids are based on the assumption that if patients truly understand the potential risks and benefits, their decisions will reflect their preferences.[16] These developments constitute a promising start. Nevertheless, many unanswered questions concerning how to elicit preferences, and how to incorporate them in clinical encounters already subject to crushing time pressures, remain. Addressing these issues constitutes an enormously challenging frontier for EBM.

## A Hierarchy of Evidence

What is the nature of the evidence in EBM? We suggest a broad definition: any empirical observation about the apparent relationship between events constitutes potential evidence. Thus, the unsystematic observations of the individual clinician constitute one source of evidence, and physiologic experiments another. Unsystematic clinical observations are limited by small sample size and, more importantly, by limitations in human processes of making inferences.[17] Predictions about intervention effects on clinically important outcomes from physiologic experiments are usually right, but occasionally disastrously wrong. Recent examples include an increase in mortality with administration of growth hormone in critically ill patients[18]; of combined vasodilators and inotropes ibopamine[19] and epoprostonol[20] in patients with congestive heart failure (CHF); and of beta-carotene in patients with previous myocardial infarction,[21] as well as the mortality-reducing effect of β-blockers[22] despite long-held beliefs that their negative inotropic action would harm CHF patients. Observational studies are inevitably limited by the possibility that apparent differences in treatment effect are really due to differences in patients' prognosis in the treatment and control groups.

Given the limitations of unsystematic clinical observations and physiologic rationale, EBM suggests a hierarchy of evidence. TABLE 1 presents a hierarchy of study designs for issues of treatment. Different hierarchies are necessary for issues of diagnosis or prognosis. Clinical research goes beyond unsystematic clinical observation in providing strategies that avoid or attenuate the spurious results. Because few, if any, interventions are effective in all patients, we would ideally test a treatment in the patient to whom we would like to apply it. Numerous factors can lead clinicians astray as they try to interpret the results of conventional open trials of therapy, which include natural history, placebo effects, patient and health worker expectations, and the patient's desire to please.

The same strategies that minimize bias in conventional trials of therapy involving multiple patients can guard against misleading results in studies involving single patients.[23] In the N of 1 RCT, patients undertake pairs of treatment periods in which they receive a target treatment in 1 period of each pair, and a placebo or alternative in the other. Patients and clinicians are blind to allocation, the order of the target and control are randomized, and patients make quantitative ratings of their symptoms during each period. The N of 1 RCT continues until both the patient and clinician conclude that the patient is, or is not, obtaining benefit from the target intervention. N of 1 RCTs are unsuitable for short-term problems; for therapies that cure (such as surgical procedures); for therapies that act over long periods of time or prevent rare or unique events (such as stroke, myocardial infarction, or death); and are possible only when patients and clinicians have the interest and time required. However, when the conditions are right, N of 1 RCTs are feasible,[24,25] can provide definitive evidence of treatment effectiveness in individual patients, and may lead to long-term differences in treatment administration.[26]

When considering any source of evidence about treatment other than N of 1 RCTs, clinicians are generalizing from results in other people to their patients, inevitably weakening inferences about treatment impact and introducing complex issues of how trial results apply to individuals. Inferences may nevertheless be strong if results come from a systematic review of methodologically strong RCTs with consistent results and are generally somewhat weaker if we are dealing with only a single RCT unless it is large and has enrolled a diverse patient population (Table 1). Because observational studies may underestimate or more typically overestimate treatment effects in an unpredictable fashion,[27,28] their results are far less trustworthy than those of RCTs. Physiologic studies and unsystematic clinical observations provide the weakest inferences about treatment effects. The Users' Guides have summarized how clinicians can fully evaluate each of these types of studies.[29-31]

This hierarchy is not absolute. If treatment effects are sufficiently large and consistent, for instance, observational studies may provide more compelling evidence than most RCTs. Observational studies have allowed extremely strong inferences about the efficacy of insulin in diabetic ketoacidosis or hip replacement in patients with debilitating hip osteoarthritis. At the same time,

instances in which RCT results contradict consistent results from observational studies reinforce the need for caution. A recent striking example comes from a large, well-conducted RCT of hormone replacement therapy as secondary prevention of coronary artery disease in postmenopausal women. While the dramatically positive results of a number of observational studies had suggested the investigators would find a large reduction in risk of coronary events with hormone replacement therapy, the treated patients did no better than the control group.[32] Defining the extent to which clinicians should temper the strength of their inferences when only observational studies are available remains one of the important challenges for EBM. The challenge is particularly important given that much of the evidence regarding the harmful effects of our therapies comes from observational studies.

The hierarchy implies a clear course of action for physicians addressing patient problems—they should look for the highest available evidence from the hierarchy. The hierarchy makes it clear that any statement to the effect that there is no evidence addressing the effect of a particular treatment is a non sequitur. The evidence may be extremely weak—the unsystematic observation of a single clinician, or generalization from only indirectly related physiologic studies—but there is always evidence. Having described the fundamental principles of EBM, we will briefly comment on additional skills that clinicians must master for optimal patient care, and their relationship to EBM.

## CLINICAL SKILLS, HUMANISM, SOCIAL RESPONSIBILITY, AND EBM

The evidence-based process of resolving a clinical question will be fruitful only if the problem is appropriately formulated. One of us, a secondary care internist, developed a lesion on his lip shortly before an important presentation. He was quite concerned and, wondering if he should take acyclovir. He immediately spent 2 hours searching for the highest-quality evidence and reviewing the available RCTs. When he began to discuss his remaining uncertainty with his partner, an experienced dentist, she quickly cut short the discussion by exclaiming, "But, my dear, that isn't herpes!"

This story illustrates the necessity of obtaining the correct diagnosis before seeking and applying research evidence in practice, the value of extensive clinical experience, and the fallibility of clinical judgment. The essential skills of obtaining a history and conducting a physical examination and the astute formulation of the clinical problem come only with thorough background training and clinical experience. The clinician makes use of evidence-based reasoning by applying the likelihood ratios associated with positive or negative physical findings to interpret the results of the history and physical examination.[33] Clinical expertise is further required to define the relevant treatment options before examining the evidence regarding their expected benefits and risks.

Finally, clinicians rely on their expertise to define features that affect the generalizability of the results to the individual patient. We have noted that, except when clinicians have conducted N of 1 RCTs, they are attempting to generalize (or, one might say, particularize) results obtained in other patients to the individual before them. The clinician must judge the extent to which differences in the treatment (local surgical expertise, or the possibility of patient noncompliance, for instance), the availability of monitoring, or patient characteristics such as age, comorbidity, or concomitant treatment may affect estimates of benefit and risk that come from the published literature. The clinician must further consider if the available studies have measured all important outcomes, if patients were followed up for a sufficient length of time, and if experimental treatment was compared with the most compelling alternatives. While our Users' Guide on treatment applicability will help clinicians define the general issues that they need to consider when advising the individual patient,[34] nothing can substitute for clinical expertise in determining the specific considerations relevant to that person.

Thus, knowing the tools of evidence-based practice is necessary but not sufficient for delivering the highest-quality patient care. In addition to clinical expertise, the clinician requires compassion, sensitive listening skills, and broad perspectives from the humanities and social sciences. These attributes allow understanding of patients' illnesses in the context of their experience, personalities, and cultures.

The sensitive understanding of the patient links to evidence-based practice in a number of ways. For some patients, incorporation of patient values for major decisions will mean a full enumeration of the possible benefits, risks, and inconvenience associated with alternative management strategies that are relevant to the particular patient. For some of these patients and problems, this discussion should involve the patients' family. For other problems, such as the discussion of screening with prostate-specific antigen in older male patients, attempts to involve other family members might violate strong cultural norms.

Many patients would be uncomfortable with an explicit discussion of benefits and risks, and object to having what they experience as excessive responsibility for decision making placed on their shoulders.[35] In such patients, who would tell us they want the physician to make the decision on their behalf, the physician's responsibility is to develop insight to ensure that choices will be consistent with patients' values and preferences. Understanding and implementing the sort of decision making process patients desire and effectively communicating the information they need requires skills in understanding the patient's narrative, and the person behind that narrative.[36,37]

Ideally, the technical skills and humane perspective of evidence-based

**Table 2.** A Hierarchy of Preprocessed Evidence

| |
|---|
| Primary studies |
| Preprocessing involves selecting only studies that are both highly relevant and with study designs that minimize bias and thus permit a high strength of inference |
| Summaries |
| Systematic reviews provide clinicians with an overview of all the evidence addressing a focused clinical question |
| Synopses |
| Synopses of individual studies or of systematic reviews encapsulate the key methodologic details and results required to apply the evidence to individual patient care |
| Systems |
| Practice guidelines, clinical pathways, or evidence-based textbook summaries of a clinical area provide the clinician with much of the information needed to guide the care of individual patients |

physicians will lead them to become effective advocates for their patients both in the direct context of the health system in which they work and in broader health policy issues. This advocacy may involve changing the system to facilitate evidence-based practice; for example, improving infrastructure for access to high-quality information to guide clinicians at the bedside. A continuing challenge for EBM, and for medicine in general, will be to better integrate the new science of clinical medicine with the time-honored craft of caring for the sick.

## ADDITIONAL CHALLENGES FOR EBM

In 1992, we identified skills necessary for evidence-based practice. These included the ability to precisely define a patient problem, and what information is required to resolve the problem, conduct an efficient search of the literature, select the best of the relevant studies, apply rules of evidence to determine their validity, and to extract the clinical message and apply it to the patient problem.[1] To these we would now add an understanding of how the patient's values affect the balance between advantages and disadvantages of the available management options, and the ability to appropriately involve the patient in the decision. Studying the process of eliciting and understanding patient values, and the best ways of incorporating them in the clinical decision making process, constitutes 1 important challenge for EBM.

Time limitation remains the biggest obstacle to evidence-based practice. Fortunately, new resources to assist clinicians are available, and the pace of innovation is rapid. One can consider a classification of information sources that comes with the mnemonic 4S: (1) the individual study, (2) the systematic review of all the available studies on a given problem, (3) a synopsis of that summary, and (4) systems of information. By systems we mean summaries that link a number of synopses related to the care of a particular patient problem (acute upper gastrointestinal tract bleeding) or type of patient (the diabetic outpatient) (TABLE 2).

Evidence-based selection and summarization is becoming increasingly available at each level. Secondary journals such as *ACP Journal Club* and *Evidence-based Medicine* review a large number of primary journals and include only articles that are both relevant and have passed a methodological filter. Clinicians can therefore be confident that any data they gather from these sources is already high on the hierarchy of evidence in Table 1. These secondary journals not only restrict themselves to studies of superior design, but present the information as structured abstracts that provide a synopsis of the individual studies and systematic reviews from the primary journals. The structure of the abstract is crucial: evidence-based synopses provide critical information about a study that are necessary for determining validity and for applying results to individual patients. While not always the case, these synopses often provide most of the information clinicians need to incorporate the results of a new study into their clinical practice.

If there is any chance it may be available, clinicians whose priority is efficient evidence-based practice should seek a high-quality systematic review rather than the primary studies addressing their clinical question. For issues of therapy, published systematic reviews, including the Cochrane Collaboration database, provide a rapidly growing repository of clinically useful summaries.

Clinicians often seek answers to questions about a whole process of care rather than a focused clinical question. Rather than "What is the impact of digoxin on my CHF patient's longevity?" the clinician may ask "Can I prolong my CHF patient's life?" or even "How can I optimize the management of my CHF patient?" Increasingly, clinicians asking these sort of questions can look to high-quality evidence-based practice guidelines or clinical pathways to provide, in effect, a series of synopses that summarize available evidence. The best systems use computer technology to match the patient or problem characteristics with an evidence-based knowledge repository and provide patient-specific recommendations. Evidence suggests that these computerized decision support systems may change clinician behavior and improve patient outcome.[38] At the same time, we must remember that recommendations can be made only for average patients, and the circumstances and values of the patient before us may differ. One way of dealing with this might be to bring the tools of decision analysis to the bedside. Whatever the ultimate solution, this exploration remains a frontier for EBM.

These developments emphasize that evidence-based practice involves not only being able to distinguish high from low quality in primary studies, but also in systematic reviews, practice guidelines, and other integrative research focused on management recommendations. That is the reason the Users' Guides have included articles that show clinicians how to use systematic reviews,[26] decision analyses,[4,39] practice guidelines,[5,40] economic analyses,[6,10] and any articles that make treatment recommendations.[8] The summary tables from each Users' Guide provide a checklist that clinicians can use to ensure that synopses of each type of study include the key information required

to assess both validity and applicability to their practice.

The last decade has seen publication of a plethora of high-quality systematic reviews and there is no slowing in sight. Most practice guidelines, however, remain methodologically weak.[41] Evidence-based systems have great potential, and are beginning to appear. Efficient production of evidence-based systems of information, increasingly user-friendly synopses, and further advances in easy electronic access to all levels of evidence-based resources should dramatically increase the feasibility of evidence-based practice in the next decade.

This article, and indeed the Users' Guides as a whole, have dealt primarily with decision making at the level of the individual patient. Evidence-based approaches can also inform health policy making,[42] day-to-day decisions in public health, and systems level decisions such as those facing managers at the hospital level. In each of these arenas, EBM can support the appropriate goal of gaining the greatest health benefit from limited resources. On the other hand, evidence as an ideology, rather than a focus for reasoned debate, has been used as a justification for many agendas in health care, ranging from crude cost-cutting to the promotion of extremely expensive technologies with minimal marginal returns. In the policy arena, dealing with differing values poses even more challenges than in the arena of individual patient care. Should we restrict ourselves to alternative resource allocation within a fixed pool of health care resources, or be trading off health care services against, for instance, lower tax rates for individuals or lower health care costs for corporations? How should we deal with the large body of observational studies suggesting that social and economic factors may have a larger impact on the health of populations than health care delivery? How should we deal with the tension between what may be best for an individual, or for the society to which that individual belongs? The debate about such issues is at the heart of evidence-based health policy making, but inevitably has implications for decision making at the individual patient level.

## CONCLUSION

The Users' Guides to the Medical Literature provide clinicians with the tools to distinguish stronger from weaker evidence, stronger from weaker syntheses, and stronger from weaker recommendations for moving from evidence to action. Much of the Users' Guides are devoted to helping clinicians understand study results and enumerate the benefits, adverse effects, toxic effects, inconvenience, and costs of treatment options, both for patients in general and for individual patients under their care. A clear understanding of the principles underlying evidence-based practice will aid clinicians in applying the Users' Guides to facilitate their patient care. Foremost among these principles are that value judgments underlie every clinical decision, that clinicians should seek evidence from as high in the appropriate hierarchy as possible, and that every clinical decision demands attention to the particular circumstances of the patient. Clinicians facile in using the Users' Guides will complete a review of the evidence regarding a clinical problem with the best estimate of benefits and risks of management options and a good sense of the strength of inference concerning those benefits and risks. This leaves clinicians in an excellent position for the final—and still inadequately explored—steps in providing evidence-based care, which is consideration of the individual patient's circumstances and values.

**Acknowledgment:** We thank Deborah Maddock for her extraordinarily skillful and dedicated management of the administrative aspects of preparing this article, and the Users' Guides series as a whole. We thank Sam Auron, MD, FRCPC, for telling us the old rabbinical joke on which we based the Clinical Scenario.

### REFERENCES

1. Guyatt GH, Rennie D. Users' guides to the medical literature. *JAMA*. 1993;270:2096-2097.
2. Evidence-Based Medicine Working Group. Evidence-based medicine: a new approach to teaching the practice of medicine. *JAMA*. 1992;268:2420-2425.
3. Guyatt GH, Meade MO, Jaeschke RJ, Cook DJ, Haynes RB. Practitioners of evidence-based care. *BMJ*. 2000;320:945-955.
4. Hunt DL, Jaeschke R, McKibbon KA, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, XXI: using electronic health information resources in evidence-based practice. *JAMA*. 2000;283:1875-1879.
5. Haynes RB, Sackett DL, Gray JM, Cook DJ, Guyatt GH. Transferring evidence from research into practice, 1: the role of clinical care research evidence in clinical decisions. *ACP J Club*. 1996;125:A14-A16.
6. Guyatt GH, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, II: how to use an article about therapy or prevention, part B: what were the results and will they help me in caring for my patients? *JAMA*. 1994;271:59-63.
7. Jaeschke R, Guyatt GH, Sackett DL, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, III: how to use an article about a diagnostic test, part B: what are the results and will they help me in caring for my patients? *JAMA*. 1994;271:703-707.
8. Richardson WS, Detsky AS, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VII: how to use a clinical decision analysis, part B: what are the results and will they help me in caring for my patients? *JAMA*. 1995;273:1610-1613.
9. Wilson MC, Hayward RS, Tunis SR, Bass EB, Guyatt G, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VIII: how to use clinical practice guidelines, part B: what are the recommendations and will they help you in caring for your patients? *JAMA*. 1995;274:1630-1632.
10. O'Brien BJ, Heyland D, Richardson WS, Levine M, Drummond MF, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, XIII: how to use an article on economic analysis of clinical practice, part B: what are the results and will they help me in caring for my patients? *JAMA*. 1997;277:1802-1806.
11. Guyatt GH, Sackett DL, Sinclair JC, Hayward R, Cook DJ, Cook RJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, IX: a method for grading health care recommendations. *JAMA*. 1995;274:1800-1804.
12. Guyatt G, Sinclair J, Cook D, Glasziou P. Users' guides to the medical literature, XVI: how to use a treatment recommendation. *JAMA*. 1999;281:1836-1843.
13. McAlister FA, Straus SE, Guyatt GH, Haynes RB. Users' guides to the medical literature, XX: integrating research evidence with the care of the individual patient. *JAMA*. 2000;283:2829-2836.
14. Drummond MF, Richardson WS, O'Brien BJ, Levine M, Heyland D, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, XIII: how to use an article on economic analysis of clinical practice, part A: are the results of the study valid? *JAMA*. 1997;277:1552-1557.
15. Feeny D, Furlong W, Boyle M, Torrance GW. Multi-attribute health status classification systems: Health Utilities Index. *Pharmacoeconomics*. 1995;7:490-502.
16. O'Connor AM, Rostom A, Fiset V, et al. Decision aids for patients facing health treatment or screening decisions: systematic review. *BMJ*. 1999;319:731-734.
17. Nisbett R, Ross L. *Human Inference*. Englewood Cliffs, NJ: Prentice-Hall International Inc; 1980.
18. Takala J, Ruokonen E, Webster NR, et al. Increased mortality associated with growth hormone treatment in critically ill adults. *N Engl J Med*. 1999;341:785-792.
19. Hampton JR, van Veldhuisen DJ, Kleber FX, et al, for the Second Prospective Randomized Study of Ibopamine on Mortality and Efficacy (PRIME II) Investigators. Randomised study of effect of ibopamine on sur-

vival in patients with advanced severe heart failure. *Lancet.* 1997;349:971-977.
**20.** Califf RM, Adams KF, McKenna WJ, et al. A randomized controlled trial of epoprostenol therapy for severe congestive heart failure: the Flolan International Randomized Survival Trial (FIRST). *Am Heart J.* 1997;134:44-54.
**21.** Rapola JM, Virtamo J, Ripatti S, et al. Randomised trial of α-tocopherol and β-carotene supplements on incidence of major coronary events in men with previous myocardial infarction. *Lancet.* 1997; 349:1715-1720.
**22.** CIBIS-II Investigators and Committees. The Cardiac Insufficiency Bisoprolol Study II (CIBIS- II): a randomised trial. *Lancet.* 1999;353:9-13.
**23.** Guyatt GH, Sackett DL, Taylor DW, et al. Determining optimal therapy: randomized trials in individual patients. *N Engl J Med.* 1986;314:889-892.
**24.** Guyatt GH, Keller JL, Jaeschke R, et al. Clinical usefulness of N of 1 randomized control trials: three year experience. *Ann Intern Med.* 1990;112:293-299.
**25.** Larson EB, Ellsworth AJ, Oas J. Randomized clinical trials in single patients during a 2-year period. *JAMA.* 1993;270:2708-2712.
**26.** Mahon J, Laupacis A, Donner A, Wood T. Randomised study of N of 1 trials versus standard practice. *BMJ.* 1996;312:1069-1074.
**27.** Guyatt GH, DiCenso A, Farewell V, Willan A, Griffith L. Randomized trials versus observational studies in adolescent pregnancy prevention. *J Clin Epidemiol.* 2000;53:167-174.
**28.** Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *BMJ.* 1998;317: 1185-1190.
**29.** Guyatt GH, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, II: how to use an article about therapy or prevention, part A: are the results of the study valid? *JAMA.* 1993;270:2598-2601.
**30.** Levine M, Walter S, Lee H, Haines T, Holbrook A, Moyer V, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, IV: how to use an article about harm. *JAMA.* 1994;271: 1615-1619.
**31.** Oxman AD, Cook DJ, Guyatt GH, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VI: how to use an overview. *JAMA.* 1994;272:1367-1371.
**32.** Hulley S, Grady D, Bush T, et al, for the Heart and Estrogen/progestin Replacement Study (HERS) Research Group. Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. *JAMA.* 1998;280: 605-613.
**33.** Sackett DL. A primer on the precision and accuracy of the clinical examination. *JAMA.* 1992;267: 2638-2644.
**34.** Dans AL, Dans LF, Guyatt GH, Richardson S, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, XIV: how to decide on the applicability of clinical trial results to your patient. *JAMA.* 1998;279:545-549.
**35.** Sutherland HJ, Llewellyn-Thomas HA, Lockwood GA, Tritchler DL, Till JE. Cancer patients: their desire for information and participation in treatment decisions. *J R Soc Med.* 1989;82:260-263.
**36.** Greenhalgh T. Narrative based medicine: narrative based medicine in an evidence based world. *BMJ.* 1999;318:323-325.
**37.** Greenhalgh T, Hurwitz B. Narrative based medicine: why study narrative? *BMJ.* 1999;318:48-50.
**38.** Hunt DL, Haynes RB, Hanna SE, Smith K. Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. *JAMA.* 1998;280:1339-1346.
**39.** Richardson WS, Detsky AS, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VII: how to use a clinical decision analysis, part A: are the results of the study valid? *JAMA.* 1995;273:1292-1295.
**40.** Hayward R, Wilson MC, Tunis SR, Bass EB, Guyatt GH, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, VIII: how to use clinical practice guidelines, part A: are the recommendations valid? *JAMA.* 1995;274:570-574.
**41.** Shaneyfelt TM, Mayo-Smith MF, Rothwangl J. Are guidelines following guidelines? the methodological quality of clinical practice guidelines in the peer-reviewed medical literature. *JAMA.* 1999;281:1900-1905.
**42.** Haynes RB, Sackett DL, Guyatt GH, Cook DJ, Gray JA. Transferring evidence from research into practice, part 4: overcoming barriers to application. *ACP J Club.* 1997;126:A14-A15.

> There are trivial truths and the great truths. The opposite of a trivial truth is plainly false. The opposite of a great truth is also true.
> —Niels Bohr (1885-1962)

# Users' Guides to the Medical Literature: XXV. Evidence-Based Medicine: Principles for Applying the Users' Guides to Patient Care

Gordon H. Guyatt; R. Brian Haynes; Roman Z. Jaeschke; et al.

*JAMA*. 2000;284(10):1290-1296 (doi:10.1001/jama.284.10.1290)

http://jama.ama-assn.org/cgi/content/full/284/10/1290