

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 $\mu\text{mol/L}$, 6 months ago it was 180 $\mu\text{mol/L}$, and in a blood sample taken a week before this clinic visit, 220 $\mu\text{mol/L}$. 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; Hertzell 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, FACP; 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 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 Medi-

cal 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 cyclophosphamide¹; 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.

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 article⁴ 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;303:1033-1041.
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.