

THERAPY

Gordon Guyatt, Deborah Cook, PJ Devereaux,
Maureen Meade, and Sharon Straus

The following EBM Working Group members also made substantive contributions to this section: Peter Wyer, Roman Jaeschke, Daren Heyland, Anne Holbrook, and Luz Maria Letelier

IN THIS SECTION

Finding the Evidence

Are the Results Valid?

Were Patients Randomized?

Was Randomization Concealed?

Were Patients Analyzed in the Groups to Which They Were Randomized?

Were Patients in the Treatment and Control Groups Similar
With Respect to Known Prognostic Variables?

Were Patients Aware of Group Allocation?

Were Clinicians Aware of Group Allocation?

Were Outcome Assessors Aware of Group Allocation?

Was Follow-up Complete?

What Are the Results?

How Large Was the Treatment Effect?

How Precise Was the Estimate of the Treatment Effect?

When Authors Do Not Report the Confidence Interval

How Can I Apply the Results to Patient Care?

Were the Study Patients Similar to the Patient in My Practice?

Were All Clinically Important Outcomes Considered?

Are the Likely Treatment Benefits Worth the Potential Harm and Costs?

Clinical Resolution



CLINICAL SCENARIO

The Internet Tells Me Spironolactone Will Prolong My Life: Doctor, Should I Take It?

You are a general internist reviewing a 66-year-old man with idiopathic dilated cardiomyopathy whom you have been following for 3 years. The patient, who has been very involved in decision making with regard to his care, presents you with an Internet summary of a new study stating, "Spironolactone saves lives in heart failure." He is very encouraged by the summary and believes that spironolactone will prolong his life.

For the preceding 18 months, the patient has been stable with mild symptoms that you classify as New York Heart Association (NYHA) class II. His echocardiogram 3 months ago demonstrated unchanged global left-sided ventricular dysfunction with an ejection fraction of 30%. His current medications include enalapril 10 mg twice a day, metoprolol 50 mg twice a day, and furosemide 20 mg once a day. His blood pressure is 110/70 mm Hg and his heart rate is 60 bpm. His blood work from the previous week reveals a creatinine level of 100 mmol/L and potassium level of 4.1 mmol/L. Since enalapril suppresses aldosterone, you wonder how spironolactone, an aldosterone antagonist, could provide additional benefit. You check the *Physician's Desk Reference* (PDR) and read that simultaneous use of enalapril and spironolactone is relatively contraindicated because of the risk of hyperkalemia.¹ You share with the patient your concerns about spironolactone as well as your determination not to overlook its potential benefits, you inform him that you will review the evidence and offer a recommendation when he returns to see you in 1 week.

FINDING THE EVIDENCE

You begin by formulating your question:

In patients with NYHA class II heart failure and a decreased ejection fraction, what is the impact of spironolactone therapy on mortality and quality of life?

Since the study you are seeking was published during the past couple of months, you know that it will not yet be included in Best Evidence, the database you would normally use to begin such a search. You therefore begin with a MEDLINE search using OVID and the following search strategy: “heart failure, congestive” (MH)—which stands for “MeSH heading”—and “spironolactone” (MH) limited to “clinical trials” and the year “1999.” This search yields only four articles, one of which is evidently your target.²

The article you retrieve reports a trial in which investigators randomized 1663 patients with NYHA class III and class IV heart failure to receive spironolactone 25 mg once daily. In this trial, patients were followed for an average of 2 years. You immediately discern that the patient you are seeing with class II heart failure would not have been eligible for the study. However, you still suspect the trial might be relevant to this patient’s care and you decide to review the report carefully before the patient returns to see you.

Although this book discusses evaluation of articles about therapy, we caution that our definition of therapy is a broad one. The principles apply to therapies designed to ameliorate symptoms or reduce morbidity and mortality in those who are acutely or chronically ill (eg, the therapeutic use of spironolactone for patients with heart failure); to interventions designed to prevent chronologically distant morbid or mortal events in patients with known underlying pathology (eg, beta blockade after myocardial infarction); to interventions designed to prevent morbidity and mortality in those at risk but without current evident illness (eg, treatment of high blood pressure); to interventions designed to improve patient outcome by improving the process of care (see Part 2B1, “Therapy and Validity, Computer Decision Support Systems”); to diagnostic tests designed to reduce morbidity or mortality (eg, gastroscopy in those with acute gastrointestinal bleeding); and to the combination of diagnostic testing and subsequent therapy that make up screening programs (eg, screening for fecal occult blood) (see Part 2F, “Moving From Evidence to Action, Recommendations About Screening”). In each of these situations, you risk doing more harm than good when you intervene. Before acting, therefore, ascertain the benefits and risks of the therapy and seek assurance that the societal resources (usually valued in dollars) consumed in the intervention will not be exorbitant.



ARE THE RESULTS VALID?

As described in “How to Use This Book,” we suggest a three-step approach to using an article from the medical literature to guide patient care. We recommend that you first determine whether the study provides valid results, that you next review the results, and, finally, that you consider how the results can be applied to the patients in your practice (Table 1B-1).

Whether the study will provide valid results depends on whether it was designed and conducted in a way that justifies claims about the benefits or risks of a therapeutic regimen. Tests of study methods break down into two sets of four questions. The first set helps you decide whether persons exposed to the experimental therapy had a similar prognosis to patients exposed to a control intervention at the beginning of the study. The second set helps you confirm that the two groups were still similar with respect to prognostic factors throughout the study.

TABLE 1B-1

Users' Guides for an Article About Therapy

Are the results valid?

Did experimental and control groups begin the study with a similar prognosis?

- Were patients randomized?
- Was randomization concealed (blinded or masked)?
- Were patients analyzed in the groups to which they were randomized?
- Were patients in the treatment and control groups similar with respect to known prognostic factors?

Did experimental and control groups retain a similar prognosis after the study started?

- Were patients aware of group allocation?
- Were clinicians aware of group allocation?
- Were outcome assessors aware of group allocation?
- Was follow-up complete?

What are the results?

- How large was the treatment effect?
- How precise was the estimate of the treatment effect?

How can I apply the results to patient care?

- Were the study patients similar to my patient?
- Were all clinically important outcomes considered?
- Are the likely treatment benefits worth the potential harm and costs?

Were Patients Randomized?

Consider the question of whether, in very sick people, hospital care prolongs life. A study finds that more sick people die in the hospital than in the community. We would easily reject the naive conclusion that hospital care kills because, intuitively,

we understand that hospitalized patients are generally much sicker than patients in the community. This difference would lead to a biased assessment, a massive underestimation of the beneficial effect of hospital care. An unbiased comparison would require a comparison of those in the hospital with equally sick patients in the community, a study that an institutional review board is unlikely to approve.

During the 1970s and early 1980s, surgeons frequently performed extracranial-intracranial bypass (ie, 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. Comparisons of outcomes among nonrandomized cohorts of patients who, for various reasons, did or did not undergo this operation fueled their conviction. These studies suggested that patients who underwent surgery appeared to fare much better than those who did not undergo surgery. However, to the surgeons' surprise, a large multicenter randomized controlled trial (RCT) in which patients were allocated to surgical or medical treatment using a process analogous to flipping a coin demonstrated that the only effect of surgery was to increase adverse outcomes in the immediate postsurgical period.³

Other surprises generated by randomized trials that contradicted the results of less rigorous trials include the demonstration that steroid injections do not ameliorate facet-joint back pain,⁴ that plasmapheresis does not benefit patients with polymyositis,⁵ and that a variety of initially promising drugs increase mortality in patients with heart failure.⁶⁻¹⁰ Such surprises occur frequently (see Part 2B1, “Therapy, Surprising Results of Randomized Trials”) when treatments are assigned by random allocation, rather than by the conscious decisions of clinicians and patients.

The reason that studies in which patient or physician preference determines whether a patient receives treatment or control (*observational studies*) often yield biased outcomes is that morbidity and mortality result from many causes, of which treatment is only one. Treatment studies attempt to determine the impact of an intervention on such events as stroke, myocardial infarction, or death—occurrences that we call the trial's *target outcomes* or *target events*. A patient's age, the underlying severity of illness, the presence of comorbid conditions, and a host of other factors typically determine the frequency with which a trial's target outcome occurs (*prognostic factors* or determinants of outcome). If prognostic factors—either those we know about or those we don't know about—prove unbalanced between a trial's treatment and control groups, the study's outcome will be biased, either under- or overestimating the treatment's effect. Because known prognostic factors often influence clinicians' recommendations and patients' decisions about taking treatment, observational studies often yield misleading results. Typically, observational studies tend to show larger treatment effects than do randomized trials,¹¹⁻¹⁴ although systematic underestimation of treatment effects also may occur.¹⁵ Observational studies can theoretically match patients, either in selecting patients for study or in the subsequent statistical analysis, for known prognostic factors (see Part 1B2, “Harm”; and see Part 2B, “Therapy and Harm, Why Study



Results Mislead: Bias and Random Error”). The power of randomization is that treatment and control groups are far more likely to be balanced with respect to both the known and the unknown determinants of outcome.

Randomization does not always succeed in its goal of achieving groups with similar prognosis. Investigators may make mistakes that compromise randomization—if those who determine eligibility are aware of the arm of the study to which the patient will be allocated, or if patients are not analyzed in the group to which they were allocated—or they may encounter bad luck.

Was Randomization Concealed?

Some years ago, a group of Australian investigators undertook a randomized trial of open vs laparoscopic appendectomy.¹⁶ The trial ran smoothly during the day. At night, however, the attending surgeon's presence was required for the laparoscopic procedure but not the open one; and the limited operating room availability made the longer laparoscopic procedure an annoyance. Reluctant to call in a consultant, and particularly reluctant with specific senior colleagues, the residents sometimes adopted a practical solution. When an eligible patient appeared, the residents checked the attending staff and the lineup for the operating room and, depending on the personality of the attending surgeon and the length of the lineup, held the translucent envelopes containing orders up to the light. As soon as they found one that dictated an open procedure, they opened that envelope. The first eligible patient in the morning would then be allocated to a laparoscopic appendectomy group according to the passed-over envelope (D. Wall, written communication, June 9, 2000). If patients who presented at night were sicker than those who presented during the day, the residents' behavior would bias the results against the open procedure.

This story demonstrates that if those making the decision about patient eligibility are aware of the arm of the study to which the patient will be allocated—if randomization is unconcealed (unblinded or unmasked)—they may systematically enroll sicker—or less sick—patients to either treatment or control groups. This behavior will defeat the purpose of randomization and the study will yield a biased result.^{17,18} Careful investigators will ensure that randomization is concealed, for example, through (a) preparation of blinded medication in a pharmacy, (b) remote randomization, in which the individual recruiting the patient makes a call to a methods center to discover the arm of the study to which the patient is allocated, or (c) (in our view a much less secure approach) ensuring that the envelope containing the code is sealed.

Were Patients Analyzed in the Groups to Which They Were Randomized?

Investigators can also corrupt randomization by systematically omitting from the results patients who do not take their assigned treatment. Readers might initially agree that such patients who never actually received their assigned treatment should be excluded from the results. Their exclusion, however, will bias the results.

The reasons people do not take their medication are often related to prognosis. In a number of randomized trials, patients who did not adhere to their treatment regimens 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.¹⁹⁻²⁴ 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 because they suffer the outcome of interest (eg, stroke or myocardial infarction) before they get to the operating room. If investigators include such poorly destined patients 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 from the systematic exclusion from the surgical group of those with the poorest prognosis.

This principle of attributing all patients to the group to which they were randomized results in an *intention-to-treat analysis*, which is analysis of outcomes based on the treatment arm to which patients were randomized, rather than which treatment they actually received. This strategy preserves the value of randomization: prognostic factors that we know about—and those we do not know about—will be, on average, equally distributed in the two groups; and the effect we see will result simply from the treatment assigned.

In conclusion, when reviewing a report of a randomized trial, look for evidence that the investigators analyzed all patients in the groups to which they were randomized (see Part 2B1, “Therapy and Validity, The Principle of Intention-to-Treat”).

Were Patients in the Treatment and Control Groups Similar With Respect to Known Prognostic Factors?

The purpose of randomization is to create groups whose prognosis, with respect to the target outcome, is similar. Sometimes, through bad luck, randomization will fail to achieve this goal. The smaller the sample size, the more likely the trial will suffer from prognostic imbalance.

Picture a trial testing a new treatment for heart failure enrolling patients in New York Heart Association functional class III and class IV. Patients in class IV have a much worse prognosis than those in class III. The trial is small, with only eight patients. One would not be terribly surprised if all four class III patients were allocated to the treatment group and all four class IV patients were allocated to the control group. Such a result of the allocation process would seriously bias the study in favor of the treatment. Were the trial to enroll 800 patients, one would be startled if randomization placed all 400 class III patients in the treatment arm. The larger the sample size, the more likely randomization will achieve its goal of prognostic balance.

Investigators can check how well randomization has done its job by examining the distribution of all prognostic factors in treatment and control groups.



Clinicians should look for a display of prognostic features of the treatment and control patients at the study's commencement—the baseline or entry prognostic features. Although we never will know whether similarity exists for the unknown prognostic factors, we are reassured when the known prognostic factors are well balanced.

The issue here is not whether there are statistically significant differences in known prognostic factors between treatment groups (eg, in a randomized trial, one knows in advance that any differences that did occur happened by chance, making the frequently cited *P* values unhelpful), but, rather, the magnitude of these differences. If the differences are large, the validity of the study may be compromised. The stronger the relationship between the prognostic factors and outcome, and the greater the differences in distribution between groups, the more the differences will weaken the strength of any inference about treatment impact (ie, you will have less confidence in the study results).

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, clinicians should look for documentation of similarity for relevant baseline characteristics; if substantial differences exist, they should note whether the investigators conducted an analysis that adjusted for those differences. When both unadjusted and adjusted analyses generate the same conclusion, readers justifiably gain confidence in the validity of the study result.

Were Patients Aware of Group Allocation?

Patients who take a treatment that they believe is efficacious may feel and perform better than those who do not, even if the treatment has no biologic action. Although we know relatively little about the magnitude and consistency of this *placebo effect*,²⁵ its possible presence can mislead clinicians interested in determining the biologic impact of a pharmacologic treatment. Even in the absence of placebo effects, patients might answer questions or perform functional tests differently, depending on whether they believe they are taking active medication.

The best way to avoid these problems is to ensure that patients are unaware of whether they are receiving the experimental treatment. For instance, in a trial of a new drug, control group patients can receive an inert tablet or capsule that is identical in color, taste, and consistency to the active medication administered to the treatment group patients. These placebos can ensure that control group patients benefit from placebo effects to the same extent as actively treated patients.

Were Clinicians Aware of Group Allocation?

If randomization succeeds, treatment and control groups in a study begin with a very similar prognosis. However, randomization provides no guarantees that the two groups will remain prognostically balanced. Differences in patient care other than the intervention under study can bias the results. For example, returning to the spironolactone trial described earlier in this section, if treatment group

patients received more intensive treatment with angiotensin-converting enzyme inhibitors or beta blockers than control group patients did, the results would yield an overestimate of the treatment effect. The reason is that both of these classes of cointervention drugs prolong life in heart failure patients.

Clinicians gain greatest confidence in study results when investigators document that all cointerventions that may plausibly impact on the outcome are administered more or less equally in treatment and control groups. The absence of such documentation is a much less serious problem if clinicians are blind to whether patients are receiving active treatment or are part of a control group. Effective blinding eliminates the possibility of either conscious or unconscious differential administration of effective interventions to treatment and control groups.

Were Outcome Assessors Aware of Group Allocation?

If either the treatment or the control group receives closer follow-up, target outcome events may be reported more frequently. In addition, unblinded study personnel who are measuring or recording outcomes such as physiologic tests, clinical status, or quality of life may provide different interpretations of marginal findings or may offer differential encouragement during performance tests, either one of which can distort results.²⁶ The study personnel assessing outcome can almost always be kept blind to group allocation, even if (as is the case for surgical therapies or health services interventions) patients and treating clinicians cannot. Investigators can take additional precautions by constructing a blinded adjudication committee to review clinical data and decide issues such as whether a patient has had a stroke or myocardial infarction, or whether a death can be attributed to cancer or cardiovascular disease. The more judgment is involved in determining whether a patient has suffered a target outcome (blinding is less crucial in studies in which the outcome is all-cause mortality, for instance) the more important blinding becomes.

Was Follow-up Complete?

Ideally, at the conclusion of a trial investigators will know the status of each patient with respect to the target outcome. We often refer to patients whose status is unknown as lost to follow-up. The greater the number of patients who are *lost to follow-up*, the more a study's validity is potentially compromised. The reason is that patients who are lost often have different prognoses from those who are retained; these patients may disappear because they suffer adverse outcomes (even death) or because they are doing well (and so did not return to be assessed). The situation is completely analogous to the reason for the necessity for an intention-to-treat analysis: patients who discontinue their medication may be less or (usually) more likely to suffer the target adverse event of interest.

When does loss to follow-up seriously threaten validity? Rules of thumb (you may run across thresholds such as 20%) are misleading. Consider two hypothetical randomized trials, each of which enters 1000 patients into both treatment and



control groups, of whom 30 (3%) are lost to follow-up (Table 1B-2). In trial A, treated patients die at half the rate of the control group (200 vs 400), a reduction in relative risk of 50%. To what extent does the loss to follow-up potentially threaten our inference that treatment reduces the death rate by half? If we assume the worst, ie, that all treated patients lost to follow-up died, the number of deaths in the experimental group would be 230 (23%). If there were no deaths among the control patients who were lost to follow-up, our best estimate of the effect of treatment in reducing the risk of death drops from 200/400, or 50%, to (400 – 230) or 170/400, or 43%. Thus, even assuming the worst makes little difference in the best estimate of the magnitude of the treatment effect. Our inference is therefore secure.

TABLE 1B-2

When Does Loss to Follow-up Seriously Threaten Validity?

	Trial A		Trial B	
	Treatment	Control	Treatment	Control
Number of patients randomized	1000	1000	1000	1000
Number (%) lost to follow-up	30 (3%)	30 (3%)	30 (3%)	30 (3%)
Number (%) of deaths	200 (20%)	400 (40%)	30 (3%)	60 (6%)
RRR not counting patients lost to follow-up	0.2/0.4 = 0.50		0.03/0.06 = 0.50	
RRR for worst-case scenario*	0.17/0.4 = 0.43		0.00/0.06 = 0	

* The worst-case scenario assumes that all patients allocated to the treatment group and lost to follow-up died and all patients allocated to the control group and lost to follow-up survived.

RRR indicates relative risk reduction.

Contrast this with trial B. Here, the reduction in the relative risk of death is also 50%. In this case, however, the total number of deaths is much lower; of the treated patients, 30 die—and the number of deaths in control patients is 60. In trial B, if we make the same worst-case assumption about the fate of the patients lost to follow-up, the results would change markedly. If we assume that all patients initially allocated to treatment—but subsequently lost to follow-up—die, the number of deaths among treated patients rises from 30 to 60, which is exactly equal to the number of control group deaths. Let us assume that this assumption is accurate. Since we would have 60 deaths in both treatment and control groups, the effect of treatment drops to 0. Because of this dramatic change in the treatment effect (50% relative risk reduction if we ignore those lost to follow-up; 0% relative risk reduction if we assume all patients in the treatment group who were lost to follow-up died), the 3% loss to follow-up in trial B threatens our inference about the magnitude of the relative risk reduction.

Of course, this worst-case scenario is unlikely. When a worst-case scenario, were it true, substantially alters the results, you must judge the plausibility of a markedly different outcome event rate in the treatment and control group patients

USING THE GUIDE

Returning to our opening clinical scenario, how well did the study of spironolactone achieve the goal of creating groups with similar prognostic factors? The investigators tell us the study was randomized, but they do not explicitly address the issue of concealment. Of the 822 treated patients, 214 discontinued treatment because of a lack of response, adverse events, or administrative reasons, as did 200 of 841 patients in the control group. The investigators appear to have included all these patients in the analysis, which they state followed intention-to-treat principles. They document the two groups' similarity with respect to age, sex, race, blood pressure, heart rate, ejection fraction, cause of heart failure, and medication use. The one variable for which there was some imbalance is the severity of underlying heart failure: 31% of control patients, vs 27% of treated patients, had NYHA class IV symptoms. This could potentially bias the results in favor of the treatment group. However, the effect is likely to be small, and we are reassured by the investigators' report of an analysis that adjusted for baseline differences in known prognostic factors.

As in many reports of randomized trials, the authors describe their study as "double-blind." Unfortunately, neither clinical epidemiologists nor readers are certain what this term signifies in terms of who is blind to allocation.²⁷ We will therefore avoid its use and instead, we will specify which groups were unaware of treatment allocation. The spironolactone report implies that patients, caregivers, and those adjudicating outcome were all blinded to allocation, and the editors of Best Evidence have conferred with the authors and reassure us that this is the case.²⁸

The authors make no explicit statement about loss to follow-up and their presentation of the data suggests they did not lose any patients. While this is possible for other outcomes, for the outcome of mortality, it seems unlikely.

The final assessment of validity is never a "yes" or "no" decision. Rather, think of validity as a continuum ranging from strong studies that are very likely to yield an accurate estimate of the treatment effect to weak studies that are very likely to yield a biased estimate of effect. Inevitably, the judgment as to where a study lies in this continuum involves some subjectivity. In this case, despite uncertainty about loss to follow-up, we judge that overall, the methods were strong. The study is thus high on the continuum between very low and very high validity, likely provides a minimally biased assessment of spironolactone's impact on heart failure patients, and can help us decide whether to recommend spironolactone to the patient under consideration.



who have not been followed. Investigators' demonstration that patients lost to follow-up are similar with respect to important prognostic variables such as age and disease severity decreases—but does not eliminate—the possibility of a different rate of target events.

In conclusion, loss to follow-up potentially threatens a study's validity. If assuming a worst-case scenario does not change the inferences arising from study results, then loss to follow-up is not a problem. If such an assumption would significantly alter the results, validity is compromised. The extent of that compromise remains a matter of judgment and will depend on how likely it is that treatment patients lost to follow-up did poorly, while control patients lost to follow-up did well.

WHAT ARE 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—ones that either happen or do not 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% of a control group died, but only 15% of those receiving a new treatment died. How might these results be expressed?

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*: the risk of events among patients on the new treatment, relative to that risk among patients in the control group, or $y/x = 0.15 / 0.20 = 0.75$.

The most commonly reported measure of dichotomous treatment effects is the complement of this relative risk, 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; and the greater the relative risk reduction, the more effective the therapy. Investigators may compute the relative risk over a period of time, as in a *survival analysis*, and call it a *hazard ratio* (see Part 2B2, “Therapy and Understanding the Results, Measures of Association”). When people do not specify whether they are talking about relative or absolute risk reduction—for instance, “Drug X was 30% effective in reducing the risk of death,” or “The efficacy of the vaccine was 92%,” they are almost invariably talking about relative risk reduction. Pharmaceutical advertisements, whether they make it explicit or not, almost invariably cite relative risk. See Part 2B2, “Therapy and Understanding the Results, Measures of Association,” for more detail about how the relative risk reduction results

in a subjective impression of a larger treatment effect than do other ways of expressing treatment effects.

How Precise Was the Estimate of the Treatment Effect?

Realistically, the true risk reduction can never be known. The best 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, a single value calculated from observations of the sample that is used to estimate a population value or parameter. The *point estimate* reminds us that, although the true value lies somewhere in its neighborhood, it is unlikely to be precisely correct. Investigators often tell us the neighborhood within which the true effect likely lies by the statistical strategy of calculating *confidence intervals*, a range of values within which one can be confident that that a population parameter is estimated to lie.²⁹

We usually (though arbitrarily) use the 95% confidence interval (see Part 2B2, “Therapy, Confidence Intervals”). You can consider the 95% confidence interval as defining the range that includes the true relative risk reduction 95% of the time. You will seldom find the true RRR toward the extremes of this interval, and you will find the true RRR beyond these extremes only 5% of the time, a property of the confidence interval that relates closely to the conventional level of “statistical significance” of $P < .05$ (see Part 2B2, “Therapy and Understanding the Results, Hypothesis Testing”). We illustrate the use of confidence intervals in the following examples.

Example 1. 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 $1 - y/x = [1 - 0.75] \times 100 = 25\%$). You might guess, however, that the true RRR might be much smaller or much greater than this 25%, based on a difference of only five deaths. In fact, you might surmise that the treatment might provide no benefit (an RRR of 0%) or might even do 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 those who are not treated). In other words, the 95% confidence interval on this RRR is -38% to 59%, and the trial really has not helped us decide whether or not to offer the new treatment.

Example 2. What if the trial enrolled 1000 patients per group rather than 100 patients per group, and the same event rates were observed 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% ($1 - y/x = 1 - [0.15/0.20] \times 100 = 25\%$). In this larger trial, you might think



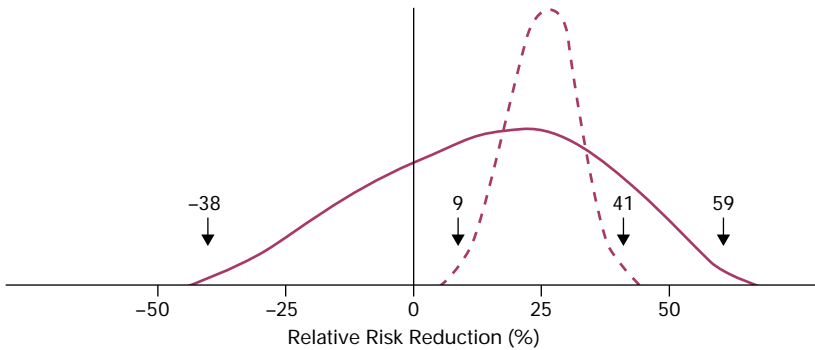
that the true reduction in risk is much closer to 25% and, again, you would be right. The 95% confidence interval on the RRR for this set of results is all on the positive side of zero and runs from 9% to 41%.

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 relative risk reduction (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 was 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% confidence interval, the values are extremely unlikely to represent the true RRR, given the point estimate (that is, the observed RRR).

Figure 1B-1 represents the confidence intervals around the point estimate of a 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 confidence interval is far narrower in the second scenario.

FIGURE 1B-1

Confidence Intervals Around Relative Risk Reduction



Two studies with the same point estimate, a 25% relative risk reduction, but different sample sizes and correspondingly different confidence intervals. The solid line represents the confidence interval around the first example, in which there were 100 patients per group and the numbers of events in active and control were 15 and 20, respectively. The broken line represents the confidence interval around the first example in which there were 1000 patients per group, and the numbers of events in active and control were 150 and 200, respectively.

It is evident that the larger the sample size, the narrower the confidence interval. When is the sample size big enough³⁰ (see Part 2B2, “Therapy and Understanding the Results, Confidence Intervals”)? 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 confidence interval. In the second example, this lower boundary was +9%. If this RRR (the lowest RRR that is consistent with the study results) is still important (that is, it is large enough for you to recommend the treat-

ment to the patient), then the investigators have enrolled sufficient patients. If, on the other hand, you do not consider an RRR of 9% important, 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 confidence interval is only 2.5%, and that a different criterion for the confidence interval (a 90% confidence interval, for instance) might be as or more appropriate.

The confidence interval 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 confidence interval. If the RRR at this upper boundary would, if true, be clinically important, the study has failed to exclude an important treatment effect. For example, consider the first example we presented in this section—the study with 100 patients in each group. This study does not exclude the possibility of harm (indeed, it is consistent with a 38% increase in relative risk), the associated *P* value would be greater than .05, and the study would be considered negative in that it failed to show a convincing treatment effect (see Figure 1B-1). Recall, however, that the upper boundary of the confidence interval was an RRR of 59%. Clearly, if this large relative risk reduction represented the truth, the benefit of the treatment would be substantial. We can conclude that, although the investigators have failed to prove that experimental treatment was better than placebo, they have also failed to prove that it is not; they have not excluded a large, positive treatment effect. Once again, you must bear in mind the proviso that the choice of a 95% confidence interval is arbitrary. A reasonable alternative, a 90% confidence interval, would be somewhat narrower.

When Authors Do Not Report the Confidence Interval

What can you do if the confidence interval around the RRR is not reported in the article? The easiest approach is to examine the *P* value. If it is exactly .05, then the lower bound of the 95% confidence limit for the RRR has to lie exactly at zero (a relative risk 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 zero.

A second approach involves calculating the confidence intervals yourself³¹ or asking the help of someone else (a statistician, for instance) to do so. Once you obtain the confidence intervals, 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, the authors of the spironolactone study might have reported changes in exercise capacity or quality of life with the drug. 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.³² This 6-minute walk improved from an average of 406 to 416 m (up 10 m) in the



experimental group receiving respiratory muscle training, and from 409 to 429 m (up 20 m) in the control group. The point estimate for improvement in the 6-minute walk due to respiratory muscle training therefore was negative, at -10 m (or a 10-m difference in favor of the control group).

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

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

USING THE GUIDE

Of the 822 treated patients, 284 died during the mean 2-year follow-up (35%), as did 386 of 841 control patients (46%). The investigators conducted a survival analysis that takes into account not only the number of people who died by the end of the trial, but the timing of their deaths along the way. Using only the raw death rate, we would compute a relative risk of 76% and a relative risk reduction of 24%. Because patients in the control group not only died more often but at an earlier point than those in the treatment group, the survival analysis yields a relative risk reduction (or hazard ratio) of 35% (95% confidence interval, 18%-48%). The survival curves started to diverge after 6 months of follow-up and showed increasing separation thereafter.

The numbers of patients hospitalized for cardiac causes in the treatment and control groups were, respectively, 336 and 260 (relative risk [RR] 0.70, 95% confidence interval [CI] 0.59-0.82). In the placebo group, about 33% of the patients improved by one or more NYHA functional classes, about 18% were unchanged, and about 48% deteriorated. In the spironolactone group, the comparable percentages were 41%, 21%, and 38%. Chance is a very unlikely explanation for the difference in changes in NYHA functional class ($P < .001$).

In terms of adverse events, nine (1%) of the patients in the control group and 61 (10%) of the treated men developed gynecomastia or breast pain ($P < .001$), and serious hyperkalemia occurred in 10 (1%) and 14 (2%) of the control and treated patients, respectively. The difference in the frequency of serious hyperkalemia was not significant.

HOW CAN I APPLY THE RESULTS TO PATIENT CARE?

Were the Study Patients Similar to the Patient in My Practice?

Often, the patient before you has different attributes or characteristics from those enrolled in the trial. She may be older, sicker or less sick, or may suffer from comorbid disease that would have excluded her from participation in the research study. If the patient had qualified for enrollment in the study—that is, if she had met all inclusion criteria and had violated none of the exclusion criteria—you can apply the results with considerable confidence.

Even here, however, there is a limitation. Treatments are not uniformly effective in every individual. Typically, some patients respond extremely well, whereas others achieve no benefit whatsoever. Conventional randomized trials estimate average treatment effects. Applying these average effects means that the clinician will likely be exposing some patients to the cost and toxicity of the treatment without benefit.

Randomized trials in individual patients offer a solution to this dilemma. In these trials, clinicians use the same strategies that minimize bias in conventional trials of therapy involving multiple patients to guard against misleading results in studies involving single patients.³³ In the N of 1 randomized controlled trial, a single patient undertakes a pair of treatment periods in which the patient receives a target treatment in one period of each pair and a placebo or alternative in the other (see Part 2B1, “Therapy and Validity, N of 1 Randomized Controlled Trials”). The patient and clinician are blinded to allocation, the order of the target treatment and control are randomized, and the patient makes quantitative ratings of his or her 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. When the conditions are right, N of 1 RCTs (a) are feasible,^{34,35} (b) can provide definitive evidence of treatment effectiveness in individual patients, and (c) may lead to long-term differences in treatment administration.³⁶

On the other hand, N of 1 RCTs are unsuitable for short-term problems, for therapies that cure, or for ascertaining effects on long-term outcomes or those that occur infrequently. Furthermore, they are possible only when patients and clinicians have the interest and time required. In most instances, clinicians must content themselves with applying results of conventional trials of other patients to the individual before them.

What if that individual does not meet a study’s eligibility criteria? 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 will not be found, and most often you can generalize the results to the patient with confidence (see Part 2B3, “Therapy and Applying the Results, Applying Results to Individual Patients”).



A related issue has to do with the extent to which we can generalize findings from a study using a particular drug to another closely (or not so closely) related agent. This is the issue of drug class effects. The issue of how conservative one should be in assuming class effects is controversial (see Part 2B3, “Therapy and Applying the Results, Drug Class Effects”).

A final issue arises when a 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 does not 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. For example, those who are older or sicker may be held up as benefiting substantially more or less than other subgroups of patients in the trial.

We encourage you to be skeptical of subgroup analyses³⁷ (see Part 2E, “Summarizing the Evidence, When to Believe a Subgroup Analysis”). The treatment is really likely to benefit the subgroup more or less than the other patients only if the difference in the effects of treatment in the subgroups is large and very unlikely to occur by chance. Even when these conditions apply, the results may be misleading if investigators did not specify their hypotheses before the study began, if they had a very large number of hypotheses, or if other studies fail to replicate the finding.

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 as substitute or *surrogate endpoints* or outcomes (see Part 2B3, “Therapy and Applying the Results, Surrogate Outcomes”). That is, investigators have chosen to substitute these variables for those that patients would consider important, usually because to confirm benefit on the latter, they would have had to enroll many more patients and follow them for far longer periods of time.

Trials of the impact of antiarrhythmic drugs following myocardial infarction provide a dramatic example of the danger of using substitute endpoints. Because such drugs have been shown to reduce abnormal ventricular depolarizations (the substitute endpoints) in the short run, it made 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 endpoint of abnormal ventricular depolarizations 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 placebo.^{38,39} Clinicians relying on the substitute endpoint 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, you must consider whether there may be deleterious effects on other outcomes. For instance, it is likely that a class of lipid-lowering agents, while reducing cardiovascular mortality, increases mortality from other causes.⁴⁰ Cancer chemotherapy may lengthen life but decreases its quality (see Part 2B2, “Therapy and Understanding the Results, Quality of Life”). Surgical trials often document prolonged life for those who survive the operation (yielding a higher 3-year survival rate 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. The most common limitation of randomized trials with regard to reporting important outcomes is the omission of documentation of drug toxicity or adverse effects.

Another long-neglected outcome is the resource implications of alternative management strategies. Few randomized trials measure either direct costs, such as drug or program expenses and health care worker salaries, or indirect costs, such as patients’ loss of income due to illness. Nevertheless, the increasing resource constraints that health care systems face mandate careful attention to economic analysis, particularly of resource-intense interventions (see Part 2F, “Moving From Evidence to Action, Economic Analysis”).

Are the Likely Treatment Benefits Worth the Potential Harm and Costs?

If you can apply the study’s results to a patient, and its outcomes are important, the next question concerns whether the probable treatment benefits are worth the effort that you and the patient must put into the enterprise. A 25% reduction in the relative risk of death may sound quite impressive, but its impact on the patient and your practice may nevertheless be minimal. This notion is illustrated using a concept called *number needed to treat* (NNT), the number of patients who must receive an intervention of therapy during a specific period of time to prevent one adverse outcome or produce one positive outcome.⁴¹

The impact of a treatment is related not only to its relative risk reduction, but also to the risk of the adverse outcome it is designed to prevent. One large trial suggests that tissue plasminogen activator (tPA) administration reduces the relative risk of death following myocardial infarction by approximately 12% in comparison to streptokinase in the setting of acute myocardial infarction.⁴² Table 1B-3 considers two patients presenting with acute myocardial infarction associated with elevation of ST segments on their electrocardiograms.



TABLE 1B-3

Considerations in the Decision to Treat Two Patients With Myocardial Infarction With Tissue Plasminogen Activator or Streptokinase

	Risk of Death Year After Myocardial Infarction (MI) With Streptokinase	Risk With tPA (Absolute Risk Reduction)	Number Needed to Treat
40-year-old man with small MI	2%	1.76% (0.24% or 0.0024)	417
70-year-old man, large MI and heart failure	40%	35.2% (4.8% or 0.048)	21

MI indicates myocardial infarction; tPA, tissue plasminogen activator

In the first case, a 40-year-old man presents with electrocardiographic findings suggesting an inferior myocardial infarction. You find no signs of heart failure, and the patient is in normal sinus rhythm with a rate of 90 beats per minute. This individual's risk of death in the first year after infarction may be as low as 2%. In comparison to streptokinase, tPA would reduce this risk by 12% to 1.76%, an absolute risk reduction of 0.24% (0.0024). The inverse of this absolute risk reduction (ARR) (that is, 1 divided by the ARR) is equal to the number of such patients we would have to treat to prevent one event (in this case, to prevent one death following a mild heart attack in a low-risk patient)—the number needed to treat (NNT). In this case, we would have to treat approximately 417 such patients to save a single life ($1 / 0.0024 = 417$). Given the small increased risk of intracerebral hemorrhage associated with tPA, and its additional cost, many clinicians might prefer streptokinase in this patient.

In the second case, a 70-year-old man presents with electrocardiographic signs of anterior myocardial infarction with pulmonary edema. His risk of dying in the next year is approximately 40%. A 12% RRR of death in such a high-risk patient generates an ARR of 4.8% (0.048), and we would have to treat only 21 such individuals to avert a premature death ($1 / 0.048 = 20.8$). Many clinicians would consider tPA the preferable agent in this man.

A key element of the decision to start therapy, therefore, is to consider the 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 do not treat, the more likely the patient will benefit from treatment and the fewer such patients we need to treat to prevent one adverse outcome (see Part 2B2, "Therapy and Understanding the Results, Measures of Association"). Knowing the NNT helps clinicians in the process of weighing the benefits and downsides associated with the management options (see Part 2B3, "Therapy and Applying the Results, Applying Results to Individual Patients"). Part 2B3, "Therapy and Applying the Results, Example Numbers Needed to Treat," presents NNTs associated with clearly defined risk groups in a number of common therapeutic situations.

Trading off benefit and risk requires an accurate assessment of medication adverse effects. Randomized trials, with relatively small sample sizes, are unsuitable for detecting rare but catastrophic adverse effects of therapy. Although RCTs are the correct vehicle for reporting commonly occurring side effects, reports regularly neglect to include these outcomes. Clinicians must often look to other sources of information—often characterized by weaker methodology—to obtain an estimate of the adverse effects of therapy.

The preferences or values that determine the correct choice when weighing benefit and risk are those of the individual patient. Clinicians should attend to the growing literature concerning patients' response to illness (see Part 2B3, "Therapy and Applying the Results, Qualitative Research"). Great uncertainty about how best to communicate information to patients, and how to incorporate their values into clinical decision making, remains. Vigorous investigation of this frontier of evidence-based decision making is, however, under way (see Part 2F, "Moving From Evidence to Action, Incorporating Patient Values").

CLINICAL RESOLUTION

The spironolactone study addressed a wide variety of relevant endpoints, including mortality, hospitalization rate, and day-to-day function. In addition, the study documents substantial increases in gynecomastia and breast pain in treated men, and a small and nonsignificant increase in episodes of serious hyperkalemia is reported.

For the group as a whole, the ARR of dying from 46% to 35% corresponds to an NNT of $1 / 0.11$, or approximately 9. However, not all patients with heart failure have the same prognosis. Class IV patients may have a mortality rate over 2 years of as high as 60%, whereas approximately 40% of class III patients may die during this time period. We would anticipate the mortality rate in class II patients to be approximately 20% during the same period.

Table 1B-4 presents some of the benefits and risks that patients with heart failure might anticipate with spironolactone. Using the point estimate of the RRR, the NNT for those with class IV failure and a higher mortality would be 6; class III and class II, with a lower baseline risk, have higher NNTs of 9 and 17, respectively. This table also highlights the smallest RRR consistent with the data (RRR, 18%), the extreme boundary of the 95% confidence interval and, hence, the largest plausible NNT. For the NYHA class IV, III, and II patients, these NNTs prove to be 9, 14, and 27, respectively. Since breast pain and gynecomastia are likely independent of NYHA functional class, given an incidence of 9%, the number needed to harm (NNH) would be 11 in all three groups (we calculate the NNH in the same way as the NNT: 1 divided by the risk difference; in this case, 1 divided by 0.09). Finally, the drug is inexpensive; the cost of a year's treatment is approximately \$25.



TABLE 1B-4

Trading Off Benefits and Risks of Spironolactone Treatment in Three Different Patients With Heart Failure

NYHA Classification for Heart Failure	Risk of Dying During 2 Years if Untreated	Likely Absolute Risk Reduction and NNT During 2 Years if Treated (30% RRR)	Smallest Plausible Absolute Risk Reduction (18% RRR)*	Risk of Breast Pain or Gynecomastia in Men and NNH
Class IV	60%	18% NNT 6	11% NNT 9	9% NNH 11
Class III	40%	12% NNT 9	7% NNT 14	9% NNH 11
Class II	20%	6% NNT 17	3.6% NNT 27	9% NNH 11

NNT indicates number needed to treat; RRR, relative risk reduction; NNH, number needed to harm.

* Calculated using lower boundary of 95% confidence interval around the RRR of 18%.

We anticipate that, given these risks and benefits, most patients would choose spironolactone treatment. This is particularly so since, if breast pain or gynecomastia develops, men can always stop the medication. However, there were virtually no class II patients who participated in this trial. Can we assume that we would see the same reduction in relative risk in these class II patients as in those with worse heart failure?

There are a number of reasons to think we might. The biology of heart failure remains similar throughout its course. The authors of the study postulate that spironolactone prevented progressive heart failure by reducing sodium retention and myocardial fibrosis, and that it prevented sudden death by averting potassium loss and increasing the myocardial uptake of norepinephrine. Spironolactone may prevent myocardial fibrosis by blocking the effects of aldosterone on the formation of collagen. There is little reason to think these mechanisms, if they indeed explain the results, would not be important in patients with NYHA class II heart failure.

Further reassurance comes from the fact that the RRR appeared similar in participating subgroups of patients with ischemic and nonischemic etiology of their heart failure and those receiving and not receiving beta blockers. Finally, other drugs that lower mortality in heart failure—angiotensin-converting enzyme inhibitors and beta blockers—appear to have similar reductions in relative risk across subgroups of patients with NYHA class II, III, and IV heart failure.⁴³

The patient before you is very interested in actively participating in decisions regarding his care. Salient points you must communicate to him include his risk of breast pain or gynecomastia of 9% during a 2-year period and his likely reduction in mortality from 20% to 14%. You must also convey the uncertainty associated with this estimate that arises both from the confidence interval around the estimate of RRR in mortality with spironolactone (which suggests his mortality may drop from 20% to 16%, rather than 14%) and from the exclusion of NYHA class II patients from the trial. When you are satisfied that the patient understands these key concepts, you will be in a position to help him arrive at a final decision about whether he wishes to take the medication.

References

1. Vasotec tablets: enalapril maleate. In: *Physician's Desk Reference*. 52nd ed. Montvale, NJ: Medical Economics; 1998:1771-1774.
2. Pitt B, Zannad F, Remme WJ, et al. The effect of spironolactone on morbidity and mortality in patients with severe heart failure. *N Engl J Med*. 1999;341:709-717.
3. Haynes RB, Mukherjee J, Sackett DL, et al. Functional status changes following medical or surgical treatment for cerebral ischemia: results in the EC/IC Bypass Study. *JAMA*. 1987;257:2043-2046.
4. Crette 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.
5. 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.
6. The Xamoterol in Severe Heart Failure Group. Xamoterol in severe heart failure. *Lancet*. 1990;336:1-6.
7. Packer M, Carver JR, Rodeheffer RJ, et al, for the PROMISE Study Research Group. Effects of oral milrinone on mortality in severe chronic heart failure. *N Engl J Med*. 1991;325:1468-1475.
8. Packer M, Rouleau JL, Svedberg K, Pitt B, Fisher L, and the Profile investigators. Effect of flosequinan on survival in chronic heart failure: preliminary results of the PROFILE study [abstract]. *Circulation*. 1993;88(suppl I):I-301.
9. 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 survival in patients with advanced severe heart failure. *Lancet*. 1997;349:971-977.
10. 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.
11. Sacks HS, Chalmers TC, Smith H Jr. Sensitivity and specificity of clinical trials: randomized v historical controls. *Arch Intern Med*. 1983;143:753-755.
12. 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.
13. Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy, I: medical. *Stat Med*. 1989;8:441-454.
14. 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.



15. Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *BMJ*. 1998;317:1185-1190.
16. Hansen JB, Smithers BM, Schache D, Wall DR, Miller BJ, Menzies BL. Laparoscopic versus open appendectomy: prospective randomized trial. *World J Surg*. 1996;20:17-20.
17. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA*. 1995;273:408-412.
18. Moher D, Jones A, Cook DJ, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet*. 1998;352:609-613.
19. 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.
20. 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.
21. Hogarty GE, Goldberg SC. Drug and sociotherapy in the aftercare of schizophrenic patients. *Arch Gen Psychiatry*. 1973;28:54-64.
22. Fuller R, Roth H, Long S. Compliance with disulfiram treatment of alcoholism. *J Chronic Dis*. 1983;36:161-170.
23. 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.
24. Horwitz RI, Viscoli CM, Berkman L, et al. Treatment adherence and risk of death after myocardial infarction. *Lancet*. 1990;336:542-545.
25. Kaptchuk TJ. Powerful placebo: the dark side of the randomised controlled trial. *Lancet*. 1998;351:1722-1725.
26. Guyatt GH, Pugsley SO, Sullivan MJ, et al. Effect of encouragement on walking test performance. *Thorax*. 1984;39:818-822.
27. Devereaux PJ, Manns BJ, Ghali WA, et al. In the dark: physician interpretation of blinding terminology in randomized controlled trials. *JAMA*. In press.
28. Henderson M, Mulrow CD. Commentary on "The effect of spironolactone on morbidity and mortality in patients with severe heart failure." *ACP J Club*. 2000 Jan-Feb; 132(1):2.
29. 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: British Medical Journal; 1989:83-100.

30. 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.
31. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical Epidemiology: A Basic Science for Clinical Medicine.* 2nd ed. Boston: Little Brown & Co Inc; 1991:218.
32. Guyatt GH, Keller J, Singer J, Halcrow S, Newhouse M. Controlled trial of respiratory muscle training in chronic airflow limitation. *Thorax.* 1992;47:598-602.
33. Guyatt GH, Sackett DL, Taylor DW, et al. Determining optimal therapy: randomized trials in individual patients. *N Engl J Med.* 1986;314:889-892.
34. Guyatt GH, Keller JL, Jaeschke R, et al. The n-of-1 randomized control trial: clinical usefulness. Our three year experience. *Ann Intern Med.* 1990; 112:293-299.
35. Larson EB, Ellsworth AJ, Oas J. Randomized clinical trials in single patients during a 2-year period. *JAMA.* 1993;270:2708-2712.
36. Mahon J, Laupacis A, Donner A, Wood T. Randomised study of n of 1 trials versus standard practice. *BMJ.* 1996;312:1069-1074.
37. Oxman AD, Guyatt GH. A consumer's guide to subgroup analysis. *Ann Intern Med.* 1992;116:78-84.
38. 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.
39. 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.
40. Muldoon MF, Manuck SB, Matthews KA. Lowering cholesterol concentrations and mortality: a quantitative review of primary prevention trials. *BMJ.* 1990;301:309-314.
41. 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.
42. The GUSTO Investigators. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. *N Engl J Med.* 1993;329:673-682.
43. 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.