THERAPY
(RANDOMIZED TRIALS)

Gordon Guyatt, Sharon Straus, Maureen O. Meade, Regina Kunz, Deborah J. Cook, PJ Devereaux, and John Ioannidis

IN THIS CHAPTER:

Clinical Scenario
A Patient with Coronary Disease and a Gastrointestinal Bleed: How Can I Best Help Avoid Vascular Events and Minimize Bleeding Risk?

Finding the Evidence
The Users' Guides

Are the Results Valid?
Did Intervention and Control Groups Start With the Same Prognosis?
Was Prognostic Balance Maintained as the Study Progressed?
Were the Groups Prognostically Balanced at the Study's Completion?

What Are the Results?
How Large Was the Treatment Effect?
How Precise Was the Estimate of the Treatment Effect?

Example 1
Example 2
How Can I Apply the Results to Patient Care?

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

Were All Patient-Important Outcomes Considered?

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

Clinical Resolution

**CLINICAL SCENARIO**

**A Patient With Coronary Disease and a Gastrointestinal Bleed: How Can I Best Help Avoid Vascular Events and Minimize Bleeding Risk?**

You are a general internist following a 62-year-old man with peptic ulcer disease and stable angina for whom you have been prescribing low-dose aspirin, a statin, an angiotensin-converting enzyme inhibitor, and as-needed nitrates. Recently, the patient developed an upper gastrointestinal bleed. Biopsy done at endoscopy was negative for *Helicobacter pylori*. In hospital, the gastroenterologist looking after your patient changed the aspirin to clopidogrel (and supported his action by citing a systematic review of thienopyridine derivatives, including clopidogrel, in high-risk vascular patients that found a decrease in the odds of a gastrointestinal bleed compared with aspirin; odds ratio, 0.71; 95% confidence interval [CI], 0.59-0.86).¹

You use *ACP Journal Club* to browse the medical literature and, reviewing the patient’s story, you recall a recent article that may be relevant. The patient is currently stable and you ask him to return in a week for further review of his medications.

**FINDING THE EVIDENCE**

Evidence from populations with vascular disease suggests that clopidogrel is likely to be similar, if not superior, to aspirin in its ability to prevent vascular events in patients with stable angina,² allowing you to focus on prevention of bleeding. You therefore formulate the relevant question: in a patient with previous aspirin-associated ulcer, is clopidogrel effective in preventing recurrent ulcer bleeding? Searching *ACP Journal Club* in your medical library’s Ovid system with the terms “clopidogrel” and “gastrointestinal bleeding” identifies 3 articles, one of which turns out to be your target: “Aspirin plus esomeprazole reduced recurrent ulcer bleeding more than clopidogrel in high-risk patients.”² You print a copy of this and the original full-text article.³

This article describes a randomized, placebo-controlled trial including 320 patients with endoscopically confirmed ulcer bleeding, either negative test results for *H pylori*
or successful eradication of *H pylori*, and anticipated regular use of antiplatelet therapy. Participants were *randomly allocated* to clopidogrel 75 mg daily and placebo or to aspirin 80 mg and esomeprazole (a potent proton-pump inhibitor) 20 mg twice daily for 12 months. The primary outcome was recurrent ulcer bleeding, and secondary outcomes included lower gastrointestinal bleeding and adverse effects.

**The Users’ Guides**

Table 6-1 presents our usual 3-step approach to using an article from the medical literature to guide your practice. You will find these criteria useful for a variety of therapy-related questions, including treating symptomatic illnesses (eg, asthma or arthritis), preventing distant complications of illness (eg, cardiovascular death after myocardial infarction), and screening for silent but treatable disease (eg, colon cancer screening).

If the answer to one key question (Were patients randomized?) is no, some of the other questions (Was randomization concealed? Were patients analyzed in the groups to which they were randomized?) will lose their relevance. As you will see, nonrandomized observational studies yield far weaker inferences than randomized controlled trials (RCTs). Nevertheless, clinicians must use the best evidence available in managing their patients, even if the quality of that evidence is limited (see Chapter 2, The Philosophy of Evidence-Based Medicine). The criteria in Chapter 12 (Harm [Observational Studies]) will help

<table>
<thead>
<tr>
<th>TABLE 6-1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Users’ Guides for an Article About Therapy</strong></td>
</tr>
</tbody>
</table>

**Are the results valid?**

- Did intervention and control groups start with the same prognosis?
  - Were patients randomized?
  - Was randomization concealed?
  - Were patients in the study groups similar with respect to known prognostic factors?
  - Was prognostic balance maintained as the study progressed?
  - To what extent was the study blinded?
  - Were the groups prognostically balanced at the study’s completion?
  - Was follow-up complete?
  - Were patients analyzed in the groups to which they were randomized?
  - Was the trial stopped early?

**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 patient-important outcomes considered?
- Are the likely treatment benefits worth the potential harm and costs?
you assess an observational study addressing a potential treatment that has not yet been evaluated in an RCT.

**ARE THE RESULTS VALID?**

**Did Intervention and Control Groups Start With the Same Prognosis?**

Were Patients Randomized?

Consider the question of whether 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 we understand that hospitalized patients are sicker than patients in the community.

Although the logic of prognostic balance is clear in comparing hospitalized patients with those in the community, it may be less obvious in other contexts. Until recently, clinicians and epidemiologists (and almost everyone else) believed that hormone replacement therapy (HRT) could decrease the risk of coronary events (death and myocardial infarction) in postmenopausal women. The belief arose from the results of many studies that found women taking HRT to have a decreased risk of coronary events. Results of the first large randomized trial of women with established coronary artery disease (CAD) provided a surprise: HRT failed to reduce the risk of coronary events. Even more recently, the Women’s Health Initiative demonstrated that HRT also failed in the primary prevention of CAD.

Other surprises generated by randomized trials include the demonstration that antioxidant vitamins fail to reduce gastrointestinal cancer—and one such agent, vitamin E, may actually increase all-cause mortality—and that a variety of initially promising drugs increase mortality in patients with heart failure. Such surprises occur periodically when investigators conduct randomized trials to test the observations from studies in which patients and physicians determine which treatment a patient receives (see Chapter 9.2, Surprising Results of Randomized Trials).

The reason that studies in which patient or physician preference determines whether a patient receives treatment or control (observational studies) often yield misleading results 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, and death—occurrences that we call the trial’s target outcomes. A patient’s age, the underlying severity of illness, the presence of comorbidity, 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 do not know about—prove unbalanced between a trial’s treatment and control groups, the study’s outcome will be biased, either underestimating or overestimating the treatment’s effect. Because known prognostic factors often influence clinicians’ recommendations and patients’ decisions about taking treatment, observational studies often yield biased results.
Observational studies can theoretically match patients, either in the selection of patients for study or in the subsequent statistical analysis, for known prognostic factors (see Chapter 12, Harm [Observational Studies], and Chapter 5, Why Study Results Mislead: Bias and Random Error). The power of randomization is that treatment and control groups are more likely to be balanced with respect to both known and unknown determinants of outcome.

What was the cause of bias in the HRT observational studies? Evidence suggests that women who took HRT enjoyed a higher socioeconomic status. Their apparent benefit from HRT was probably due to factors such as a healthier lifestyle and a greater sense of control over life. Whatever the explanation, we are now confident that it was their previous prognosis, rather than the HRT, that led to lower rates of CAD.

Although randomization is a powerful technique, it does not always succeed in creating groups with similar prognosis. Investigators may make mistakes that compromise randomization, or randomization may fail because of simple bad luck. The next 2 sections address these issues.

**Was Randomization Concealed?**

Some years ago, a group of Australian investigators undertook a randomized trial of open vs laparoscopic appendectomy. 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 limited operating room availability made the longer laparoscopic procedure an annoyance. Reluctant to call in a consultant, the residents sometimes adopted what they saw as a practical solution. When an eligible patient appeared, the residents held the semiopaque envelopes containing the study assignment up to the light. They opened the first envelope that dictated an open procedure. The first eligible patient in the morning would then be allocated to the laparoscopic appendectomy group according to the passed-over envelope (D. Wall, written communication, June 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.

When those enrolling patients are unaware and cannot control the arm to which the patient is allocated, we refer to randomization as concealed. In unconcealed trials, those responsible for recruitment 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. Careful investigators will ensure that randomization is concealed through strategies such as 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 assigned.

**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 outcomes, is similar. Sometimes, through bad luck, randomization will
fail to achieve this goal. The smaller the sample size, the more likely the trial will have 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 8 patients. One would not be surprised if all 4 class III patients were allocated to the treatment group and all 4 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.

You can check how effectively randomization has balanced prognostic factors by looking for a display of patient characteristics of the treatment and control groups at the study’s commencement—the baseline or entry prognostic features. Although we will never know whether similarity exists for the unknown prognostic factors, we are reassured when the known prognostic factors are well balanced.

All is not lost if the treatment groups are not similar at baseline. Statistical techniques permit adjustment of the study result for baseline differences. Adjusted analyses may not be preferable to unadjusted analyses, but when both analyses generate the same conclusion, readers gain confidence in the validity of the study result.

**Was Prognostic Balance Maintained as the Study Progressed?**

**To What Extent Was the Study Blinded?**

If randomization succeeds, treatment and control groups in a study begin with a similar prognosis. Randomization, however, provides no guarantees that the 2 groups will remain prognostically balanced. **Blinding** is, if possible, the optimal strategy for maintaining prognostic balance.

Table 6-2 describes 5 groups involved in clinical trials that, ideally, will remain unaware of whether patients are receiving the experimental therapy or control therapy. You are probably aware that patients who take a treatment that they believe is effective may feel and perform better than those who do not, even if the treatment has no biologic activity. Although the magnitude and consistency of this

<table>
<thead>
<tr>
<th>Patients</th>
<th>To avoid placebo effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinicians</td>
<td>To prevent differential administration of therapies that affect the outcome of interest (cointervention)</td>
</tr>
<tr>
<td>Data collectors</td>
<td>To prevent bias in data collection</td>
</tr>
<tr>
<td>Adjudicators of outcome</td>
<td>To prevent bias in decisions about whether or not a patient has had an outcome of interest</td>
</tr>
<tr>
<td>Data analysts</td>
<td>To avoid bias in decisions regarding data analysis</td>
</tr>
</tbody>
</table>
placebo effect remain uncertain, investigators interested in determining the biologic impact of a pharmacologic or nonpharmacologic treatment will ensure patients are blind to treatment allocation. Similarly, rigorous research designs will ensure blinding of those collecting, evaluating, and analyzing data (Table 6-2). Demonstrations of bias introduced by unblinding—such as the results of a trial in multiple sclerosis in which a treatment benefit judged by unblinded outcome assessors disappeared when adjudicators of outcome were blinded—highlight the importance of blinding. The more that judgment is involved in determining whether a patient has had a target outcome (blinding is less crucial in studies in which the outcome is all-cause mortality, for instance), the more important blinding becomes.

Finally, differences in patient care other than the intervention under study—cointervention—can, if they affect study outcomes, bias the results. Effective blinding eliminates the possibility of either conscious or unconscious differential administration of effective interventions to treatment and control groups. When effective blinding is not possible, documentation of potential cointervention becomes important.

Were the Groups Prognostically Balanced at the Study’s Completion?
Unfortunately, investigators can ensure concealed random allocation and effective blinding and still fail to achieve an unbiased result.

Was Follow-up Complete?
Ideally, at the conclusion of a trial, you will know the status of each patient with respect to the target outcome. The greater the number of patients whose outcome is unknown—patients 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—they may disappear because they have adverse outcomes or because they are doing well and so did not return for assessment.

When does loss to follow-up seriously threaten validity? Rules of thumb (you may run across thresholds such as 20%) are misleading. Consider 2 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 6-3). In trial A, treated patients die at half the rate of the control group (200 vs 400), a relative risk reduction (RRR) 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)/400 or 170/400, or 43%. Thus, even assuming the worst makes little difference to the best estimate of the magnitude of the treatment effect. Our inference is therefore secure.
Contrast this with trial B. Here, the reduction in the relative risk (RR) 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. Because 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% RRR if we ignore those lost to follow-up; 0% RRR 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 RRR.

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 lost to follow-up.

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, the extent to which validity is compromised depends on how likely it is that treatment patients lost to follow-up did badly while control patients lost to follow-up did well. That decision is a matter of judgment.

### TABLE 6-3
When Does Loss to Follow-up Seriously Threaten Validity?

<table>
<thead>
<tr>
<th></th>
<th>Trial A</th>
<th></th>
<th>Trial B</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Treatment</td>
<td>Control</td>
<td>Treatment</td>
<td>Control</td>
</tr>
<tr>
<td>Number of patients randomized</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
</tr>
<tr>
<td>Number (%) lost to follow-up</td>
<td>30 (3)</td>
<td>30 (3)</td>
<td>30 (3)</td>
<td>30 (3)</td>
</tr>
<tr>
<td>Number (%) of deaths</td>
<td>200 (20)</td>
<td>400 (40)</td>
<td>30 (3)</td>
<td>60 (6)</td>
</tr>
<tr>
<td>RRR not counting patients lost to follow-up</td>
<td>0.2/0.4 = 0.50</td>
<td></td>
<td>0.03/0.06 = 0.50</td>
<td></td>
</tr>
<tr>
<td>RRR—worst-case scenarioa</td>
<td>0.17/0.4 = 0.43</td>
<td></td>
<td>0.00/0.06 = 0</td>
<td></td>
</tr>
</tbody>
</table>

Abbreviation: RRR, relative risk reduction.

aThe 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.
Was the Trial Stopped Early?
Although it is becoming increasingly popular, stopping trials early when one sees an apparent large benefit is risky. Trials terminated early will compromise randomization if they stop at a “random high” when prognostic factors temporarily favor the intervention group. Particularly when sample size and the number of events are small, trials stopped early run the risk of greatly overestimating the treatment effect (see Chapter 9.3, Randomized Trials Stopped Early for Benefit).

Were Patients Analyzed in the Groups to Which They Were Randomized?
Investigators can also undermine randomization if they omit from the analysis patients who do not receive their assigned treatment or, worse yet, count events that occur in nonadherent patients who were assigned to treatment against the control group. Such analyses will bias the results if the reasons for nonadherence are related to prognosis. In a number of randomized trials, patients who did not adhere to their assigned drug 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. When adherent patients are destined to have a better outcome, omitting those who do not receive assigned treatment undermines the unbiased comparison provided by randomization. Investigators prevent this bias when they follow the intention-to-treat principle and analyze all patients in the group to which they were randomized (see Chapter 9.4, The Principle of Intention to Treat).

USING THE GUIDE
Returning to our opening clinical scenario, did the experimental and control groups begin the study with a similar prognosis? The study was randomized and allocation was concealed; 320 patients participated and 99% were followed up. The investigators followed the intention-to-treat principle, including all patients in the arm to which they were randomized, and stopped when they reached the planned sample size. There were more patients who smoked (13% vs 8.2%) and regularly consumed alcohol (8.1% vs 5%) in the clopidogrel group compared with the aspirin-esomeprazole group. This could bias the results in favor of the aspirin-esomeprazole, and the investigators do not provide an adjusted analysis for the baseline differences. Clinicians, patients, data collectors, outcomes assessors, and data analysts were all blind to allocation.

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 baseline differences between the groups, we conclude that the methods were strong.
What Are the Results?

How Large Was the Treatment Effect?
Most frequently, RCTs 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 have an event or they do not, 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 (Table 6-4). How might one express these results?

One possibility would be the absolute difference (known as the absolute risk reduction [ARR], or risk difference), between the proportion who died in the control group (baseline risk or control event rate [CER]) and the proportion who died in the treatment group (experimental event rate [EER]), or CER – EER = 0.20 – 0.15 = 0.05. Another way to express the impact of treatment is as an RR: the risk of events among patients receiving the new treatment relative to that risk among patients in the control group, or EER/CER = 0.15/0.20 = 0.75.

The most commonly reported measure of dichotomous treatment effects is the complement of the RR, the RRR. It is expressed as a percentage: 1 – (EER/CER) × 100% = (1 – 0.75) × 100% = 25%. An RRR of 25% means that the new treatment reduced the risk of death by 25% relative to that occurring among control patients; the greater the RRR, the more effective the therapy. Investigators may compute the RR over a period of time, as in a survival analysis, and call it a hazard ratio (see Chapter 7, Does Treatment Lower Risk? Understanding the Results). When people do not specify whether they are talking about RRR or ARR—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 RRR (see Chapter 7, Does Treatment Lower Risk? Understanding the Results, for more detail about how the RRR results in a subjective impression of a larger treatment effect than do other ways of expressing treatment effects).

<table>
<thead>
<tr>
<th>TABLE 6-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Results From a Hypothetical Randomized Trial</td>
</tr>
<tr>
<td>Exposure</td>
</tr>
<tr>
<td>Treatment</td>
</tr>
<tr>
<td>Control</td>
</tr>
</tbody>
</table>

Control event rate (CER): 20/100 = 20%.
Experimental event rate (EER): 15/100 = 15%.
Absolute risk reduction or risk difference: CER – EER, 20% – 15% = 5%.
Relative risk: EER/CER = (15/100)/(20/100) × 100% = 75%.
Relative risk reduction: 1 – (EER/CER) × 100% = 1 – 75% = 25%.
How Precise Was the Estimate of the Treatment Effect?

We can never be sure of the true risk reduction; the best estimate of the true treatment effect is what we observe in a well-designed randomized trial. This estimate is called a point estimate to remind us that, although the true value lies somewhere in its neighborhood, it is unlikely to be precisely correct. Investigators often tell us the neighborhood within which the true effect likely lies by calculating CIs, a range of values within which one can be confident the true effect lies. We usually use the 95% CI (see Chapter 8, Confidence Intervals). You can consider the 95% CI as defining the range that—assuming the study was well conducted and has minimal bias—includes the true RRR 95% of the time. The true RRR will generally lie beyond these extremes only 5% of the time, a property of the CI that relates closely to the conventional level of statistical significance of $P < .05$ (see Chapter 10.1, Hypothesis Testing). We illustrate the use of CIs 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% (CER = 20/100 or 0.20, EER = 15/100 or 0.15, and $1 – EER/CER = (1 – 0.75) \times 100 = 25\%$). You might guess, however, that the true RRR might be much smaller or much greater than 25%, based on a difference of only 5 deaths. In fact, you might surmise that the treatment might provide no benefit (an RRR of 0%) or might even do harm.

![Figure 6-1: Confidence Intervals in Trials of Various Sample Size](image)

**Figure 6-1**

Confidence Intervals in Trials of Various Sample Size

- **Study A:** 100 patients/group
- **Study B:** 1000 patients/group

Abbreviations: CI, confidence interval; RRR, relative risk reduction.

Two studies with the same point estimate, a 25% RRR, but different sample sizes and correspondingly different CIs. The x-axis represents the different possible RRR, and the y-axis represents the likelihood of the true RRR having that particular value. The solid line represents the CI around the first example, in which there were 100 patients per group, and the number of events in active and control was 15 and 20, respectively. The broken line represents the CI around the second example in which there were 1000 patients per group, and the number of events in active and control was 150 and 200, respectively.
(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 those who are not treated). In other words, the 95% CI 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 (CER = 200/1000 = 0.20) and 150 deaths in the treatment group (EER = 150/1000 = 0.15)? Again, the point estimate of the RRR is 25% (1 – EER/CER = 1 – (0.15/0.20) × 100 = 25%).

In this larger trial, you might think that our confidence that the true reduction in risk is close to 25% is much greater, and, again, you would be right. The 95% CI 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 RRR (or any other measure of effect) is close to what we have observed. In the second example, 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% CI, the values are very unlikely to represent the true RRR, given the point estimate (that is, the observed RRR). All this, of course, assumes the study has satisfied the validity criteria we discussed earlier.

Figure 6-1 represents the CIs around the point estimate of an RRR of 25% in these 2 examples, with a risk reduction of 0 representing no treatment effect. In both scenarios, the point estimate of the RRR is 25%, but the CI is far narrower in the second scenario.

Not all randomized trials have dichotomous outcomes, nor should they. 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.35 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% CIs around this difference in changes in exercise capacity and consider their implications. The investigators tell us that
the lower boundary of the 95% CI was –26 (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, patients are unlikely to perceive adding 5 m to the 400 recorded at the start of the trial as important, and this result effectively excludes an important benefit of respiratory muscle training as applied in this study.

It will not surprise you that the larger the sample size, the narrower the CI. If you want to learn more about CIs, including finding out when the sample size is sufficiently large, see Chapter 8, Confidence Intervals.

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 their patients.

**USING THE GUIDE**

Using the raw numbers provided in the article, 1 of 159 people (0.6%) in the aspirin-esomeprazole group and 13 of the 161 people (8%) in the clopidogrel group experienced a recurrence of ulcer. The RRR is 92%, and the 95% CI extends from 41% to 99%. The very large effect and the small number of events somewhat reduce your confidence in this result; 4.4% of the aspirin-esomeprazole group and 9.4% of the clopidogrel group had an adverse effect (defined as dyspepsia or an allergy). The investigators also reported that 11 patients in the aspirin-esomeprazole group and 9 patients in the clopidogrel group experienced recurrent ischemic events.

**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. He or she may be older or younger, sicker or less sick, or may have comorbid disease that would have excluded him or her from participation in the research study. If the patient qualified for enrollment in the study, you can apply the results with considerable confidence.

What if that individual does not meet a study’s eligibility criteria? The study result probably applies even if, for example, he or 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 do not apply to the patient. You usually will not find a compelling reason, and most often you can generalize the results to your patient with confidence (see Chapter 11.1, 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. The issue of drug class effects and how conservative one should be in assuming class effects remains controversial (see Chapter 22.5, Drug Class Effects). Generalizing findings of surgical treatment may be even riskier. Randomized trials of carotid endarterectomy, for instance, demonstrate much lower perioperative rates of stroke and death than one might expect in one’s own community.36

A final issue arises when a patient fits the features of a subgroup of patients in the trial report. We encourage you to be skeptical of subgroup analyses37 (see Chapter 20.4, When to Believe a Subgroup Analysis). The treatment is 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 Patient-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 provide a sufficient reason for administering these drugs (see Chapter 11.4, Surrogate Outcomes). Here, investigators have chosen substitute or surrogate outcomes rather than those that patients would consider important. What clinicians and patients require is evidence that the treatments improve outcomes that are important to patients (patient-important outcomes), 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.38

Trials of the impact of antiarrhythmic drugs after myocardial infarction illustrate the danger of using substitute outcomes or endpoints. Because such drugs had demonstrated a reduction in abnormal ventricular depolarizations (the substitute endpoints), it made sense that they should reduce the occurrence of life-threatening arrhythmias. A group of investigators performed randomized trials on 3 agents (encainide, flecainide, and moricizine) that were previously shown to be effective in suppressing the substitute endpoint of abnormal ventricular depolarizations. 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.39,40 Clinicians relying on the substitute endpoint of arrhythmia suppression would have continued to administer the 3 drugs, to the considerable detriment of their patients.

Even when investigators report favorable effects of treatment on one patient-important outcome, you must consider whether there may be deleterious effects on other outcomes. For instance, cancer chemotherapy may lengthen life but decrease its quality (see Chapter 10.5, Measuring Patients’ Experience). Randomized trials often fail to adequately document the toxicity or adverse effects of the experimental intervention.41
Composite endpoints represent a final dangerous trend in presenting outcomes. Like surrogate outcomes, composite endpoints are attractive for reducing sample size and decreasing length of follow-up. Unfortunately, they can mislead. We may find that a trial that reduced a composite outcome of death, renal failure requiring dialysis, and doubling of serum creatinine level actually demonstrated a trend toward increased mortality with the experimental therapy and showed convincing effects only on doubling of serum creatinine level (see Chapter 10.4, Composite Endpoints).

Another long-neglected outcome is the resource implications of alternative management strategies. Health care systems face increasing resource constraints that mandate careful attention to economic analysis (see Chapter 22.1, 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 RR of death may sound quite impressive, but its impact on your patient and practice may nevertheless be minimal. This notion is illustrated by using a concept called number needed to treat (NNT), the number of patients who must receive an intervention of therapy during a specific period to prevent 1 adverse outcome or produce 1 positive outcome.

The impact of a treatment is related not only to its RRR but also to the risk of the adverse outcome it is designed to prevent. One large trial in myocardial infarction suggests that tissue plasminogen activator (tPA) administration reduces the RR of death by approximately 12% in comparison to streptokinase. Table 6-5 considers 2 patients presenting with acute myocardial infarction associated with elevation of ST segments on their electrocardiograms.

In the first case, a 40-year-old man presents with electrocardiographic findings suggesting an inferior myocardial infarction. You find no signs

| TABLE 6-5 | Considerations in the Decision to Treat 2 Patients With Myocardial Infarction With Tissue Plasminogen Activator or Streptokinase |
|---|---|---|
| Risk of Death 1 Year After MI With Streptokinase (CER) | Risk With tPA (EER) (ARR = CER – EER) | Number Needed to Treat (100/ARR When ARR Is Expressed as a Percentage) |
| 40-Year-old man with small MI | 2% | 1.86% (0.24% or 0.0024) | 417 |
| 70-Year-old man with large MI and heart failure | 40% | 35.2% (4.8% or 0.048) | 21 |

Abbreviations: ARR, absolute risk reduction; CER, control event rate; EER, experimental event rate; tPA, tissue plasminogen activator; MI, myocardial infarction.
of heart failure, and the patient is in normal sinus rhythm, with a rate of 90/min. 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 1.86%, an ARR of 0.24% (0.0024). The inverse of this ARR (that is, 100 divided by the ARR expressed as a percentage) is equal to the number of such patients we would have to treat to prevent 1 event (in this case, to prevent 1 death after a mild heart attack in a low-risk patient), the NNT. In this case, we would have to treat approximately 417 such patients to save a single life (100/0.24 = 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 subsequent 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 (100/4.8 = 20.8). Many clinicians would consider tPA the preferable agent for 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 1 adverse outcome (see Chapter 7, Does Treatment Lower Risk? Understanding the Results). Knowing the NNT helps clinicians in the process of weighing the benefits and downsides associated with the management options (see Chapter 11.1, Applying Results to Individual Patients). Chapter 11.2 (Example Numbers Needed to Treat) presents NNTs associated with clearly defined risk groups in a number of common therapeutic situations.

Tradeoff of benefit and risk also requires an accurate assessment of treatment adverse effects. Randomized trials, with relatively small sample sizes, are unsuitable for detecting rare but catastrophic adverse effects of therapy. Clinicians must often look to other sources of information—often characterized by weaker methodology—to obtain an estimate of the adverse effects of therapy (see Chapter 12, Harm [Observational Studies]).

The preferences or values that determine the correct choice when weighing benefit and risk are those of the individual patient. 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 medicine is, however, under way (see Chapter 22.2, Decision Making and the Patient).

Clinicians may find it tempting to turn to the article’s authors for guidance about tradeoffs between benefits and risks. Because of the possibility of conflict of interest, this can be dangerous. If you are nervous about this danger, check out our strategies to avoid being misled (see Chapter 11.3, Dealing With Misleading Presentations of Clinical Trial Results).
The study that we identified showed a decrease in the recurrence of ulcer bleeding in high-risk patients receiving aspirin-esomeprazole in comparison with those taking clopidogrel. The authors also found that more people in the clopidogrel group experienced an adverse effect from the therapy and that there was no significant difference in the risk of ischemic events, although the small number of outcomes leaves any inferences from this result extremely weak.

Our patient is at a high risk of a recurrent ulcer, given his recent gastrointestinal bleed secondary to an aspirin-induced ulcer. His case is similar to those of patients included in this study. You translate the reduction in risk of bleeding into an NNT of approximately 13 (clopidogrel risk of 8.1% – aspirin/esomeprazole of 6.6% = 7.5%; NNT = 100/7.5). Given the very large effect, the NNT using the more conservative boundary of the CI of an RRR of approximately 40%—and thus an NNT of approximately 30—may be more realistic. In combination with the reduction in less-important adverse effects, this seems to be a clear patient-important benefit.

The patient found his bleeding episode terrifying, and he also believes that lowering his risk of bleeding by even as little as 3% during a year would be worthwhile. He gulps, however, when you tell him that esomeprazole costs $2.20 per pill, and if he takes the drug as administered in the trial, it will cost him more than $1600 in the next year. You then explain that the investigators’ choice of medication leaves some doubt about the best drug to use along with aspirin. Esomeprazole is still under patent, explaining the high cost. The investigators could have chosen omeprazole, a proton-pump inhibitor with marginal differences in effectiveness relative to esomeprazole, which the patient can purchase for approximately half the price. Ultimately, the patient chooses the aspirin/omeprazole combination.

**References**


