Clinical Research Methodology 3: Randomized Controlled Trials

Daniel I. Sessler, MD,* and Peter B. Imrey, PhD†

Randomized assignment of treatment excludes reverse causation and selection bias and, in sufficiently large studies, effectively prevents confounding. Well-implemented blinding prevents measurement bias. Studies that include these protections are called randomized, blinded clinical trials and, when conducted with sufficient numbers of patients, provide the most valid results. Although conceptually straightforward, design of clinical trials requires thoughtful tradeoffs among competing approaches—all of which influence the number of patients required, enrollment time, internal and external validity, ability to evaluate interactions among treatments, and cost. (Anesth Analg 2015;121:1052–64)

Observational study designs, as explained in our previous 2 reviews, are inherently vulnerable to systematic errors from selection and measurement biases and confounding; retrospective studies may additionally be subject to reverse causation when the timing of exposure and outcome cannot be precisely determined. Fortunately, 2 study design strategies, randomization and blinding, preclude or mitigate these major sources of error.

Randomized clinical trials (RCTs) are cohort studies—necessarily prospective—in which treatments are allocated randomly to the subjects who agree to participate. In the most rigorous randomized trials, called blinded or masked, knowledge of which treatment each patient receives is concealed from the patients and, when possible, from investigators who evaluate their progress. Blinded RCTs are particularly robust because randomization essentially eliminates the threats of reverse causation and selection bias to study validity, and, considerably mitigates the threat of confounding. Well-executed blinding/masking simultaneously mitigates measurement bias and placebo effects by equalizing their impacts across treatments.[1] We now discuss these benefits in more detail.

RANDOMIZATION

In reviewing patient records, we would not expect 2 treatments for a medical condition to be randomly distributed among patients because care decisions are influenced by numerous factors, including physician and patient preference. Patients given different treatments may therefore differ systematically and substantially in their risks of outcomes.

Randomization eliminates selection bias in treatment comparisons because, by definition, randomized assignments are indifferent to patient characteristics of any sort. For example, investigators reviewing records might find that aggressively treated septic critical care patients do better than those treated conservatively. Improved outcomes might occur because aggressive therapy was more effective. But it might equally well be that patients who appeared stronger were selected for more aggressive therapy because they were thought better able to tolerate it and then did better because they were indeed stronger. Looking back, it is hard to distinguish selection bias from true treatment effect.

The threat of confounding, which can be latent in a population or result from selection or measurement bias, refers to misattribution error because of a third-factor linking treatment and outcome. For example, anesthesiologists may prefer to use neuraxial anesthesia in older patients under the impression that it is safer. Let us say, however, that younger patients in a study cohort actually do better. But did they do better because of they were given general anesthesia (a causal effect of treatment) or simply because they were younger (confounding by age, the third variable)? Looking back in time, as in a retrospective analysis of existing data, in complex medical situations, it is difficult to determine the extent to which even *known* mechanisms contribute causally to outcome differences. And it is essentially impossible to evaluate the potential contributions of unknown mechanisms.

Confounding can only occur when a third factor, the confounder, differs notably between treatment groups in the study sample. Randomization largely prevents confounding because, in a sufficiently large study, patients assigned to each treatment group will most likely be very similar with respect to non–treatment-related factors that might influence the outcome. The tendency of randomization to equalize allocations across treatment groups improves as sample size increases and applies to both known and unknown factors, thus making randomization an exceptionally powerful tool.

Randomized groups in a sufficiently large trial will thus differ substantively only in treatment. The consequence is that, unless there is measurement bias, differences in outcome can be causally and specifically attributed to the treatment itself—which is what we really want to know.

Trials need to be larger than one might expect to provide a reasonable expectation that routine baseline factors are comparably distributed in each treatment group (i.e., $n > 100$ per group). For factors that are uncommon, many more patients are required. For example, consider a feasibility trial testing tight-versus-routine perioperative glucose

From the *Department of Outcomes Research, Anesthesiology Institute, Cleveland Clinic, Cleveland, Ohio; and †Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic, Cleveland, Ohio.

Accepted for publication May 1, 2015.

Funding: Supported by internal sources.

The authors declare no conflicts of interest.

Address correspondence to Daniel I. Sessler, MD, Department of Outcomes Research, Cleveland Clinic, 9500 Euclid Ave./P77, Cleveland, OH 44195. Address e-mail to DS@OR.org; www.OR.org.

DOI: 10.1213/ANE.0000000000000862

control with 25 patients per group. It is unlikely that average age or weight will differ greatly between the 2 groups. But it would not be too surprising if, by bad luck, 1 group had 14 diabetic patients and the other had just 7. Such a large disparity will occur about 1 in 12 trials of this size, just by chance. But diabetes mellitus obviously has substantial potential to influence the investigators' ability to tightly control glucose concentration. Baseline inhomogeneity, that is, the disparate numbers of patients with diabetes mellitus, would then complicate interpretation of the results of an otherwise excellent trial.

One approach to avoiding baseline heterogeneity is increasing the sample size. Larger sample size lowers the risk of any given level of heterogeneity across all potentially important factors, including unknown confounders. But when there are a limited number of obviously important uncommon factors, an alternative is to stratify randomization (i.e., randomize separately in groups distinguished by the values of these factors), using one of several possible devices that restrict the results so that each factor is distributed roughly evenly across treatment groups. This process assures that even uncommon characteristics, if included in the stratification, will be roughly comparably distributed across treatment groups.

With electronic randomization systems, such as those accessed in real time via the Web, it is possible to include various levels of stratification without difficulty. In multicenter studies, permuted-block randomization is virtually always used because site-related variations are assumed to be a substantial potential confounder and a given site may not enroll enough patients to assure baseline homogeneity. In this process, random treatments are assigned in small sequential blocks of each site's patients, so that different sites' patients are comparably distributed across treatment groups at the end of each block. Block sizes are concealed from clinicians enrolling patients, and often changed, so that enrolling clinicians and other investigators are masked to the allocation process and cannot know the next patient's treatment assignment.

Although theoretically straightforward, randomization can be tricky in practice. For example, patients may agree to participate in a randomized trial but then drop out if they are not assigned to the novel treatment they wanted. Similarly, consented patients may remove themselves from a study based on perceived lack of benefit or complications. To the extent that patients drop out of studies nonrandomly, there remains potential for selection bias.

BLINDING

Clinical measurements, no matter how carefully taken, are rarely precisely accurate. Preservation is imperfect, biosamples degrade, batches of reagents and biologics vary, human operators processing samples are inconsistent, and radiologists and pathologists vary in their interpretations of images and biopsy specimens, which are themselves of variable quality. But these sorts of errors occur randomly, with underestimates likely to be balanced by overestimates over large numbers of measurements. Consequently, in clinical trials, such errors are equally likely to favor 1 treatment group over the other and very unlikely to favor either substantially when averaged over large numbers of tests. Random error adds variability to results, which degrades statistical power, but power can always be augmented by increasing the numbers of patients and/or measurements.

Nonrandom errors that affect treatment groups differently are a major threat to study validity. Measurement bias, that is, error resulting from distortions of a measurement process, and thus not expected to average out over many measurements, can compromise even large RCTs. For example, suppose investigators are evaluating the effect of a new drug on postoperative nausea and vomiting. And let us say that both the investigators and patients know whether they have been assigned to the experimental or control group. Does knowing the treatment influence the amount of nausea and vomiting reported? It almost surely does: patients (and some physicians) typically assume that the novel treatment is better, and overestimate its benefits, even when the new treatment is actually a placebo or the same drug as in the control group.[2] They make this assumption even though superiority is actually unknown, and comparing the novel and conventional treatment is the entire point of the study. The effect is so strong that IRBs typically forbid investigators to describe the experimental treatments as novel because patients assume "new" implies "new and improved."

Overestimation of benefits is, of course, most likely with subjective outcomes such as pain or nausea and vomiting. But improvement also occurs with supposedly objective outcomes, possibly because patients expect to do better and that expectation alone improves physical characteristics such as immune function. Improved outcomes are welcome of course, but it would be a mistake under these circumstances to conclude that the novel therapy *caused* the improvement. The same logic applies to complications, which are often underestimated for novel treatments.

Fortunately, if done effectively, blinding can largely eliminate measurement biases by equalizing errors over treatment groups so they do not skew treatment comparisons. Blinding in research is defined by concealing which treatment a patient actually receives. Full blinding prevents bias in the measurement process because patients, clinicians who treat them, and all who evaluate their health outcomes do not know which treatment—the experimental or control—has been provided. Patient responses, and the processes by which physicians and investigators measure them, are thus not affected by their impressions about the superiority of novel treatments. Any inherent biases in the measurement process, whether random or systematic, are experienced by each study group equally and will tend to cancel out of the intergroup comparisons.

RANDOMIZED AND BLINDED STUDIES

In retrospective analyses of treatment outcomes, which by definition involve looking back in time, it is difficult to determine whether outcome differences result from baseline differences in the group characteristics (selection bias), a third factor (confounding), nonrandom measurement errors (measurement bias), or from the treatment itself (the relationship of interest).

Limited ability to recognize distortions from bias and confounding weakens retrospective studies. But in prospective cohort studies, treatments can be randomly assigned to prevent selection bias and minimize confounding; furthermore, participating patients and clinicians can be blinded to prevent measurement bias. Studies that incorporate the protections of randomization and blinding are called randomized, blinded clinical trials. And because the major sources of error are limited in such trials, they are considered to provide the highest level of clinical evidence.

Before randomized trials, it was difficult to objectively evaluate treatment efficacy, leading Oliver Wendell Holmes to comment: "If all our drugs could be sunk to the bottom of the sea, it would be all the better for mankind—and all the worse for the fishes." Amazingly, the first major randomized trial, which evaluated streptomycin as a treatment for tuberculosis, was not conducted until 1948. Clinical trials have thus been an established part of medical research only for about the past 65 years.

Of course designed experimentation can only be done prospectively. And enrolling, treating, and monitoring a sufficient number of patients may take many years and considerable expense. Thus, trials should be considered a scarce resource and used to address important questions supported by compelling mechanistic understanding and, preferably, reasonable animal and previous human experience.

CONTINUOUS VERSUS DICHOTOMOUS OUTCOMES

Clinical signs of disease risk or progression, and thus prognostic of outcomes, are often measurable as continuous variables, even though the outcomes they predict, and which patients directly perceive, are typically dichotomous or categorical. Examples include QT interval in relation to nonperfusing arrhythmias, such as torsade-de-pointes; diastolic blood pressure in relation to myocardial infarction or ischemic stroke; and creatinine clearance in relation to development of end-stage renal disease. The first example is typical in that QT interval per se cannot be detected by patients and would only interest them to the extent that it might predict something they do care about, such as having a cardiac arrest. That said, many continuous outcomes are clinically and economically important, such as length of critical care, duration of hospitalization, months of severe pain, or quality of life.

Continuous outcomes are easier than dichotomous outcomes to study because continuous markers or mediators can often be usefully evaluated without awaiting ultimate patient outcomes, which can involve extended patient surveillance. Moreover, because comparisons of patients with respect to dichotomous outcomes can only be of 4 sorts: both yes, both no, yes/no, or no/yes, whereas numerical values can be ordered and scaled by distance, there is simply more information in numerical markers than dichotomous outcomes, even when the latter are far more important. A consequence is that more patients must usually be studied to reliably distinguish treatments on the basis of dichotomous as compared with continuous outcomes, a difference that increases for infrequent events.

As thus might be expected, many and perhaps most clinical trials designate continuous variables rather than more important dichotomous variables as their primary outcomes. There is an implicit assumption that what improves the continuous predecessor will similarly benefit the later patient outcome, although this does not always prove to be the case. Despite the costs, ultimately, decisions about comparative benefit to patients are thus best based on unambiguous dichotomous outcomes that damage patients in perceptible ways.

SUBJECT SELECTION

When selecting participants for clinical trials, there is inherent tension between the scientific requirements of demonstrating treatment benefit conclusively and the desire for results that are widely clinically useful. This relates to the distinction between internal and external validity of a study.

Internal validity denotes the strength of design and analysis of a study in protecting from spurious conclusions about similar participants such as might otherwise arise from reverse causation mix-ups, selection and measurement biases, confounding, and chance. External validity refers to applicability of research results to persons different from participants, conditions different from those of the trial, other doses, other routes of administration, or even other agents within a given pharmaceutical class.

Restricting participants to a near-homogeneous group reduces outcome variability, which reduces the sample size, time, and cost needed to obtain statistically significant differences between treatments, assuming one exists. It can also narrow the spread of potential confounders, which limits the distortion confounding can produce. It is also reasonable and ethically appropriate to restrict enrollment to patients who are (1) especially likely to benefit from the experimental intervention and (2) unlikely to suffer complications. A consequence is that most clinical trials evaluate only small subsets of potential populations of interest.

Although there are compelling reasons to restrict trial enrollment, technically results of any trial apply only to patients similar to those who were studied. The problem is that many randomized trials are nonrepresentative, failing to include even substantial minority populations, along with subjects who cannot read, do not speak the dominant language, or have cognitive impairment, all of which complicate obtaining valid consent. Because enrollment is often highly restricted, many trials suffer from poor generalizability; that is, results may extrapolate poorly to the great majority of patients who might benefit from the experimental treatment.

The difficulty, of course, is that clinicians rarely care for patients whose characteristics and background closely match those who participated in relevant trials. In clinical practice, we thus need to make reasonable extrapolations from existing data to our patients. Extrapolation is almost always preferable to the alternative (best clinical judgment or "making it up"), but confidence in various extrapolations, especially among competing study results, is enhanced when studies include broad populations. In practice, real-world applications of new treatments usually prove to be less effective and more toxic than reported in clinical trials. Table 1 summarizes the advantages and disadvantages of loose and tight clinical trial enrollment criteria.

Generally speaking, observational studies such as retrospective cohort analyses tend to include broad populations. Consequently, their results, if correct, tend to generalize well. However, their internal validity is often degraded by unknown amounts of confounding and selection and measurement bias, making it difficult to assert correctness confidently. Tightly controlled clinical trials with restrictive

Table 1. Subject Selection Strategies

- Tight criteria
  - Reduce variability and sample size
  - Exclude subjects at risk of treatment complications
  - Include subjects most likely to benefit
  - May restrict to advanced disease, compliant patients, etc.
  - Slow enrollment
  - Best case results (compliant low-risk patients with ideal disease stage)
- Loose criteria
  - Include more real-world participants
  - Increase variability and sample size
  - Speed enrollment
  - Enhance generalizability

enrollment have the opposite problem. Randomization and blinding limit bias and confounding, resulting in high internal validity; however, trial external validity often suffers as a direct consequence of tight enrollment criteria.

A further factor to consider is that the results of even the best clinical trial apply to the typical patient in each group. But it remains possible—and perhaps probable—that a treatment that on average is beneficial might be harmful for particular members of the group. Similarly, toxicity might be substantial for certain members of a particular study group, even if on average toxicity was less with the designated treatment. In fact, such divergent results are sometimes predictable and are termed practice misalignments.[3] They result when enrollment criteria are broad and the study intervention is one that is typically titrated by clinicians; for example, designating fixed low and high concentrations of vasopressors for treatment of sepsis, rather than allowing clinicians to titrate lower concentrations for less severe disease and higher concentrations in more critical patients.[4]

CROSSOVER TRIALS

The prototypical clinical trial is a randomized and blinded parallel-group prospective cohort study. In such studies, patients are randomly assigned to 1 of ≥2 treatment groups, and patient outcomes are compared among these groups. Statistically, the challenge with this approach is separating background population variability from the treatment effect. If the treatment effect is large (say comparing a highly effective treatment with placebo), it is often easy to isolate the effect of an experimental treatment. But in an era where there are many effective treatments for most conditions, the much more common problem is to identify small incremental benefits. And those can easily be obscured by natural variation in the population.

An alternative to comparing ≥2 groups of different patients is thus to randomly compare ≥2 treatments in the same patients, that is, use a crossover design.

Specifically, a single group of subjects is exposed to ≥2 treatments, with each patient crossed over from one treatment to another in a random order. This approach is statistically efficient because the comparison is between treatments within each patient. Patient variability across the population thus largely drops from the analysis because, instead of comparing patients with other patients to determine treatment effects, the response of each patient to a treatment is first compared with his or her own responses to other treatments, and these within-patient differences in responses are then aggregated across patients.

The more that patients vary from one another, the greater the increase in sensitivity—statistical power—to detect a treatment benefit between paired *t* tests and related methods used in a crossover trials and unpaired *t* tests and related methods used for parallel-group trials. The biggest advantage of crossover trials is thus that, by minimizing the effects of population variability, it is easier to observe specific intervention effects, and with far fewer patients than would be needed than with a parallel-group approach.

Unfortunately, however, substantial limitations of crossover designs often preclude their use. For example, they assume that the underlying disease process is static over the trial's duration and that treatments have no permanent residual effects, so that the condition of a patient when the second treatment is initiated is no longer affected by the first treatment and is similar to when the study started. Surgery, for example, never qualifies and often a wash-out period of no therapy is required to restore the patient to a baseline state. A corollary is that crossover trials can only evaluate transient symptoms and signs, mediators, and soft outcomes such a laboratory tests, patient function, or hemodynamic responses. They are thus perfect for comparing ≥2 analgesics for control of persistent and stable pain or of statins for blood cholesterol reduction. But crossover designs cannot directly address far more important questions, such as which statin most reduces the risk of a heart attack or mortality.

FACTORIAL DESIGNS

Factorial designs have been used in other fields for decades, but only recently have they become well established and increasingly common in medicine.[5] The basic approach is to *simultaneously* randomly assign patients to ≥4 interventions. For example, patients in a single study might be randomly assigned to clonidine or a placebo for clonidine *and* to aspirin or a placebo for aspirin. The randomization might be arranged so that the fraction of patients given 1 treatment, say clonidine, does not depend on whether or not the patient receives the other treatment, say aspirin. This avoids confounding, that is, contaminating, the effect of 1 treatment with that of the other, so that the overall clonidine effect can be evaluated without regard to aspirin, and vice versa.[6,7] These overall results are called marginal effects because, when results are conventionally described in a 2 × 2 table with rows representing 1 treatment axis (e.g., clonidine or clonidine placebo) and columns representing the other (aspirin or aspirin placebo), they appear at the edges, or margins, of the table. Effectively, then, 2 studies are conducted simultaneously in the same group of patients. This approach is obviously efficient, in that 2 hypotheses can be tested with only slightly more patients than would be required for each question alone.

Another important benefit of factorial designs is that they allow investigators to evaluate effect modification, that is, the interaction between treatments, as well as their marginal effects. Consider, for example, a large trial of clonidine versus placebo and consider a separate large trial of aspirin versus placebo. The results of the first would presumably clearly identify the benefit (if any) from clonidine, and the results of the second would identify the effects of aspirin on the designated outcome. But no matter what the results, clinicians would reasonably ask what might be expected when the 2 treatments are combined.

The answer to that important question cannot be determined from 2 independent trials but is readily available from an adequately powered factorial trial. Note, however, that interactions are typically smaller than main effects and thus require larger studies to reliably detect. Moreover, once an interaction is detected, the study, in effect, breaks into pieces because the effect of each treatment now must be evaluated separately with and without the other, with only half the original number of patients available to do this, and thus losing an important benefit of the factorial approach.

A disadvantage of factorial trials is their increased complexity, both in study conduct and potentially in interpretation. Another disadvantage is that each intervention usually has at least slightly different contraindications, but only patients who can be randomly assigned to *each* potential intervention can be included, thus narrowing the group of eligible patients.

A complex anesthesia study using a 6-way randomization of various antiemetic strategies and powered not only for marginal effects but also for second- and third-level interactions among the treatments exemplifies the strengths and limitations of such designs.[8]

BEFORE-AND-AFTER STUDIES AND CLUSTER TRIALS

Many patient characteristics—such as smoking status, race, sex, and obesity—cannot be randomized. Other interventions, such as changes to health systems per se, for example, introduction of electronic records, a new billing system, or a clinical pathway, cannot be randomized on an individual basis. These require extensive system and behavioral changes, months or years to implement, and cannot be undertaken or reversed on a patient-by-patient basis. But such system interventions are arguably among the most important changes to evaluate, with huge potential impact. How then to study them?

The most common approach for evaluating system interventions is a before-and-after analysis. However, this is a weak study design because it is nearly impossible to account for 4 important sources of error. The first is that most aspects of health care improve over time. And although it is tempting to attribute improvement to a specific intervention of interest, many aspects of care inevitably change simultaneously. The intervention of interest is thus confounded by other simultaneous interventions or other changes, which are often unrecognized, such as subtle caregiver behavioral changes.

Consider surgical wound infections. The monthly incidence of surgical wound infections varies considerably, even in large high-quality hospitals. Why the incidence varies remains unknown, which speaks to the many factors contributing to this key safety and quality metric. When wound infections spike, as they do from time to time, high-quality hospitals recognize the problem and evaluate ameliorative approaches. Almost inevitably, this involves discussions among infection-control specialists, nurses, surgeons, and anesthesiologists. But it also involves environmental experts who assess operating room airflow, filter integrity, and bacterial colonization of caregivers and the air-handling system. Simultaneously, there is renewed focus on hand washing, surgical-site preparation, and proper draping.

The attentions of all these experts and efforts among caregivers are invariably successful, and the surgical wound infection incidence soon returns to the expected level. But why? A likely reason is that the high infection incidence that triggered concern was simply random variation and never reflected a systematic increase that would be expected to persist. If so, the infection incidence's subsequent return to baseline level was simply regression-to-the-mean, the third major problem with before-and-after studies.

Let us say, however, that the original increase in the wound infection incidence was real and that the subsequent return to baseline risk reflected enhanced care. Even assuming real improvement, a major problem with before-and-after studies is that it remains unrealistic to attribute improvement to a single intervention because several were introduced simultaneously, and it is essentially impossible to determine the extent to which the benefit of one was confounded with the benefits of others.

The fourth weakness of before-and-after studies is the Hawthorne effect. The term was introduced by Henry Landsberger based on studies conducted at the Hawthorne Works, a Western Electric factory near Chicago.[9] What Landsberger noticed was that tiny environmental changes (such as more illumination) improved productivity and satisfaction, but only briefly. In fact, most any change briefly improved productivity—including subsequently reducing illumination levels! But because the improvements were not sustained, he correctly concluded that they resulted from employee engagement in the process rather than the environmental change per se.

The Hawthorne effect is undoubtedly useful in that worker engagement improves quality and satisfaction. For example, anesthesiologist-driven changes in departmental call schedules improve satisfaction even when the total workload is unchanged. Skilled chairs thus encourage member-driven process improvement discussions because engagement in this very process improves quality and satisfaction. The difficulty is that before-and-after studies often attribute all improvement to a specific factor, ignoring the very real improvement that comes just from the *process* of discussing changes (to say nothing of other simultaneous improvements).

An alternative to before-and-after studies, with all their weaknesses, is cluster randomized trials. Cluster randomization refers to randomly allocating treatments to entire groups or clusters, in a one-for-all fashion, en masse. Clusters might consist of classes attending continuing medical education events, all patients of individual primary care providers or entire group practices, patients or clinicians at individual outpatient clinics affiliated with a health care system, or groups of employees of several different hospitals.

The difficulty, of course, is that entire care systems within a hospital, or even entire hospitals, need to be randomized and each becomes a unit of analysis. Statistically, each care unit or hospital is treated more-or-less the way a single patient would be in a conventional trial. In other words, statistical analysis is not based on the number of patients but on the number of units that are randomized. Considering the logistical challenges and cost of such trials, it obvious why they are uncommon although well-conducted cluster randomized trials can yield invaluable information.[10,11]

Escalating costs of large RCTs have prompted consideration of other ways to accrue and study large numbers of patients quickly at low cost. One possibility is an alternating intervention trial, a nonrandomized extension of a quality improvement demonstration approach that we mention here because of its utility and importance under special circumstances, especially when combined with electronic data capture.

Quality improvement projects typically at most use a before-and-after approach to evaluating treatment effect. For example, surgeons might switch to a new skin preparation routine and then evaluate the incidence of wound infection before and after the change. Very often, the switch will appear salutary—but quite possibly because of unrecognized concomitant changes, regression to the mean, or the Hawthorne effect. The true benefit (if any) of the new skin preparation technique is essentially impossible to determine with this approach.

Now consider an alternative and stronger study design. Say an anesthesia department is considering switching from sevoflurane (more expensive but shorter acting) to isoflurane (less expensive but longer acting) but is concerned the switch might prolong hospital stays. Instead of simply changing anesthetics and a before-and-after comparison, 1 anesthetic can be used exclusively for a couple of weeks and then the other for a couple of weeks, with continued alternation for several months. Note that this is not an individual patient randomization; in fact, the trial is not randomized at all or necessarily blinded. But in the course of say a dozen switches, other process improvements and any regression to the mean may reasonably be expected to average out, letting the investigators isolate the nearly pure effect of anesthetic choice on duration of hospitalization. This expectation is most warranted when the intervention involves a relatively inconspicuous and noncontroversial aspect of an overall treatment process, in contrast to a major change in treatment approach.

Alternating intervention approaches work especially well when the intervention treatment is easy to switch, and the studies can be inexpensive when the outcomes are electronically recorded or otherwise easy to obtain. However, because of the lack of randomization and blinding, such studies can be completely invalidated if some clinicians preferentially schedule patients or surgeries or if discharge decisions are influenced by which treatment is in effect that week or by secular factors correlated, for whatever reason, with both the alternation cycle and the outcome—here, duration of stay (for an example of an alternating intervention trial, see the study by Kopyeva et al.[12]). Note that a stepped-wedge approach to introducing an intervention at multiple sites can gain some, but not all, benefits of alternating intervention approaches.

Although alternating intervention trials share aspects of quality improvement studies, they are very much real research and absolutely require IRB approval. And of course, only selected quality-related interventions are appropriate. But when the novel intervention is reasonably expected to be at least as safe and effective, and possibly better or less expensive, many IRBs will approve this sort of trial with waived consent.

COMPOSITE OUTCOMES

Large trial sizes, factorial designs, and composite outcomes are among the major trends in clinical trials.[5] The use of composite outcomes is increasing because they offer distinct advantages for studies with dichotomous outcomes—which include most hard outcomes such as major complications and deaths. Composites combine ≥2 components into a single summary, typically dichotomous, which then becomes the basis for analysis.

There are 2 major advantages to composite outcomes. The first is that a single measure may poorly characterize the anticipated effect of an experimental intervention. Consider guided fluid management, which may help prevent respiratory failure, protect the kidneys and reduce the risk of wound complications. There is little clinical basis for choosing just 1 of these complications as the primary end point of a trial because each is important. Furthermore, an outcome such as wound complications includes infection, dehiscence, anastomotic leak, abscess, etc. A composite outcome that includes all these potential clinical events thus has better construct validity, meaning that it more fully characterizes the potential overall benefit of guided fluid management, than its single components.

The second major advantage of composite outcomes relates to sample size. Sample size for studies with dichotomous outcomes is determined by baseline incidence of the outcome and the expected treatment effect. Treatment effect can be enhanced by optimal patient selection but is otherwise a function of the intervention. Baseline outcome incidence, however, is increased in a composite because each component contributes. Thus rare outcomes can be evaluated if a sufficient number are combined into a composite. The alternative approach, considering each component as an independent primary outcome and taking a statistical penalty for using multiple parallel significant tests, each incurring separate risk of false-positive error, vastly increases the number of patients required. Composites thus allow investigators to simultaneously combine clinical relevant outcomes and reduce sample size.

The most common form of composite is a simple collapsed composite in which the first occurrence of any component makes the composite positive. The general rule for a collapsed composite is that its components should be of comparable severity and have roughly similar incidences. For example, it would be a poor idea to use an infection composite that combines organ space infection, deep sternal infection, abscess, sepsis, and urinary tract infection. Urinary tract infections are far more common than all the other components combined, to say nothing of being far less serious. The proposed composite is thus essentially just a measure of urinary tract infection, which is not what the investigators really want to assess.

A further assumption of composite outcomes is that treatments at least change component outcomes in the same direction (i.e., between no change and improved). When outcomes respond quite heterogeneously to a treatment, composites become difficult to interpret. Consider, for example, a composite of vascular complications that includes myocardial infarctions and stroke. β-Blockers affect each differently, significantly reducing myocardial infarctions while increasing stroke.[13] Simply adding the incidence of each (1 positive and 1 negative) might wrongly imply that β-blockers have little overall effect, whereas they in fact have clinically important, but divergent, actions.

Fortunately, there are alternatives to simple collapsed composites. In fact, there are many ways to construct and analyze composites, some of which avoid the general requirement for components to have comparable incidences and severities (for more detail about composite outcomes and associated statistical complexities, see Mascha and Imrey[14] and Mascha and Sessler[15]).

SAMPLE SIZE, INTERIM ANALYSES, AND STOPPING RULES

Before the mid-1970s, clinical trialists were faced with an uncomfortable choice between adequate trial monitoring and controlling chance error. Well-designed trials were planned to control the potential for false-positive and false-negative results within what were considered acceptable limits, often 5% for the chance of false-positive results and 10% or 20% for the chance of false-negative results. But such performance characteristics were based on the assumption that trials would be analyzed only once when results from all patients had been obtained.

Investigators, however, would often monitor accumulating results that, depending on how trends developed, might raise concerns about the ethical basis or financial advisability of continuing to enroll patients. Investigators might also informally evaluate results as they accrued and stop trials when their efficacy results reached statistical significance. The difficulty with this approach is that such multiple unprotected looks at the data for efficacy signals substantially increase a study's chance of false-positive error. Similarly, multiple looks for possible safety concerns increase the chance of stopping too early because of transient random imbalances in adverse events and thus the chance of a false-negative error.

Multiple looks are problematic because trial results are influenced not only by true population differences but also by the vagaries of chance in how these play out over time, as results from a particular clinical trial sample accumulate. Observed values at various times during the trial are thus sometimes greater than the true population differences, if any, and at other times less. The degree to which observed values differ from true population values is a function of sample size, with more variability being observed in small studies, which is why power improves with larger size.

The trouble is that investigators who evaluate data as it accrues, even informally, and stop a study when the results look good, are effectively choosing a random high point on the difference curve as it evolves over time. Because the chance that a random curve will meet any particular fixed criterion at any one of many possible times is, by definition, greater than the chance it will meet that criterion at a single predetermined time, random false-positive and false-negative errors will be far more likely with multiple looks than with a single evaluation at the end, unless the statistical significance criterion is made more stringent to account for multiple interrogations of the data. The risk of overstating treatment effects is not restricted to formal analysis and remains even if data are only informally inspected by investigators. Multiple unprotected evaluations of data may, in part, explain why many studies, especially small ones, are subsequently contradicted.[16]

Group sequential clinical trials, proposed by Pocock et al.[17] and extensively developed since, provide a menu of practical solutions to this problem. Within this framework, multiple looks at trial results are completely acceptable if governed by a formal interim analysis plan incorporated in the clinical trial design. Rather than allowing each additional look to further increase the chances of false-positive and/or false-negative conclusions above the target levels, 5% and 20% for instance, group sequential designs reduce the error probabilities at each look, so that the original, planned overall false-positive and false-negative error proportions for the entire study are maintained.

In practice, the total false-positive risk (i.e., α = type 1 statistical error probability) is distributed across each of the analyses at different times to keep the total at 5%, for example. The simplest way to do this, although conservative, is to divide the intended α-risk by the number of analyses. So if there are 3 interim looks planned, plus a final evaluation, each might be assessed at $0.05/4 = 0.0125 = 1.25\%$. That is, only P values <0.0125 would be considered statistically significant at any given analysis. A similar process can be used to distribute risk of false-negative error (β = type 2 statistical error probability) over various analysis points (Fig. 1).

Most investigators regard taking equal chances of error at early and late times as placing too much emphasis and risk of error early, when few results are available, at the expense of too little emphasis late, when most or all the results are known. In practice, various formulas are used to distribute risks of false-positive and false-negative errors in ways that allow investigators to better satisfy ethical and financial concerns by stopping trials early when data are already sufficiently conclusive as to relative efficacy and/or safety or when it is clear that continuing the trial cannot achieve the intended objective. Most trials are designed to be increasingly sensitive at later times, with the final test, if the trial does not stop early, only slightly more stringent than its fixed sample size counterpart.

Often, the weighting over time for α- and β-risks differs. For example, investigators may want to stop the trial early if there is no evidence of efficacy but continue to a much larger number of patients if the treatment looks effective. This is equivalent to cutting your losses when the treatment does not appear effective but requiring a high degree of evidence if it does.

Group sequential designs exact a price for their flexibility. The sample size planned for such a design is always larger than that needed for a single-look design with the same error probabilities (α and β) and hence the same statistical power. However, the difference is usually not large and is mitigated because, in practice, the actual sample size accrued is reduced when the trial stops early. Thus, one must plan for a modestly larger sample size than for a 1-look trial; but on average, group sequential trials require fewer patients, often substantially fewer, because many

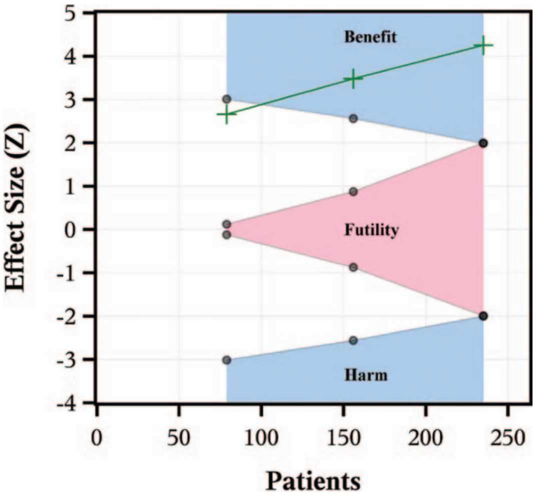

Figure 1. Observed standardized differences in primary postoperative sore throat proportions (green line) at 2 interim analyses (at *N* = 79 and *N* = 179) and the final analysis (at *N* = 235) of a clinical trial comparing an experimental licorice-based gargle to a sugar solution for prevention of sore throat and postextubation coughing. The upper and lower blue areas are stopping regions for benefit and harm of the licorice solution, respectively; the pink region designates early stopping for futility, and non-significant final differences. Although statistically significant efficacy was demonstrated at the second interim analysis based on 150 patients, this minimal risk the study was nevertheless continued to 235 patients to assess treatment benefit more precisely. This trial had a >50% chance of early stopping after 150 patients due to futility, had there been no real licorice benefit. Reprinted from Ruetzler K, et al. A randomized, double-blind comparison of licorice versus sugar-water gargle for prevention of postoperative sore throat and postextubation coughing. Anesth Analg 2013;117:614–21.

such trials can be stopped early. Trials that are stopped early for unplanned reasons overstate statistical significance if knowledge of accumulated results informs the stopping decision. But a study that is stopped per rule at a scheduled interim analysis is not a "stopped early" trial because it was, in fact, stopped as planned by protocol.

Because group sequential designs have so many advantages, many modern trials include 1 or periodic interim analyses. Most use O'Brien-Fleming stopping boundaries, for which statistical penalties are relatively small. However, each interim look at the data requires special data quality control and database activities, a formal statistical analysis, and review by an executive committee and/or data safety monitoring board. There is thus some fixed cost to each interim analysis. The amount of relevant new information that becomes available during any calendar period between looks, and the potential benefits of stopping at the beginning rather than at the end of the period, will vary among trials and during the course of a single trial, depending on the absolute and relative durations of the patient enrollment, treatment, and follow-up periods and the risks and potential benefits of the experimental treatment. In our experience, ≤3 interim looks usually provide sufficient flexibility at the cost of a small increase in the maximum required patient enrollment and a manageable amount of additional statistical and administrative work.

Depending on the size and nature of the trial, interim analyses might be reported to an executive committee or a completely independent data safety monitoring board. Typically, they are presented on a group A versus group B basis because the decision to stop or continue the trial should usually be independent of the group assignments. Interim analysis results should not be shared with investigators involved in data collection to reduce the risk of bias. Although the statistical properties of group sequential trial designs presume strict adherence to their stopping rules, in practice such bodies consider these statistical rules as guidance, taking into account the scientific context of the study and other factors when making stop-versus-continue decisions.

For studies with normally distributed continuous outcomes, the primary determinants of sample size are outcome variability and treatment effect (difference between the study groups). For dichotomous outcomes, the primary determinants are baseline incidence and treatment effect. Furthermore, the number of study subjects needed increases rapidly, in proportion to the inverse square, as treatment effect is reduced. Although less important than whether data are continuous or dichotomous, the statistical analysis approach also influences the sample size.

At a baseline incidence of 20%, for example, it takes only 398 patients—199 per group—to give a balanced 2-arm study 80% power at a 5% α for detecting a 50% reduction in the outcome incidence. But, there are few new interventions that provide anything resembling a 50% treatment effect. Given the same baseline risk, powering the study for a 25% risk reduction would require 1812 patients. And for a 20% risk reduction, 2894 patients are needed. The difficulty is that 20% is often the largest treatment that might realistically be expected, and a trial of that size might miss even smaller effects that would be considered highly clinically important.

Of course, treatment effect is not known when investigators begin a trial. After all, the point of a trial is exactly to determine treatment effect. Nonetheless, comparative clinical trials would rarely meet ethical standards unless based on reasonable mechanistic, animal, and some human, data. Thus investigators usually have at least some basis for expecting effect at a given magnitude. Another consideration is clinical importance; there is no point in powering a study for an effect that would not be considered clinically important. The general approach is thus to power trials for plausibly anticipated treatment effects of clinically important magnitude.

Sample size estimation is straightforward for single-look, 2-group trials. But the calculations quickly get difficult, and numerical simulations may be required when dealing with multivariable analysis, nonparametric distributions, composite outcomes, factorial designs, multiple looks, survival or cumulative incidence analyses, and other design complexities.[15,18] But in all cases, the most problematic part for investigators is developing realistic estimates of baseline incidence, population variability, treatment effect, and potential patient enrollment rate. Too often, these estimates are optimistic, resulting in underpowered trials.

EQUIVALENCE AND NONINFERIORITY TRIALS

A placebo is the natural comparator in a trial conducted simply to show that an experimental treatment provides some degree of benefit. Historically, such placebo-controlled trials have confirmed scientific and clinical breakthroughs against previously untreatable medical problems. When comparing a novel experimental treatment with an inactive placebo, the only relevant question is superiority of the presumably active experimental agent. In practice, such studies often use 2-sided statistical hypothesis testing to recognize harm when it occurs, but the hope is to demonstrate a difference in patient responses that favors the treatment.

Increasingly, however, new treatments emerge for problems for which ≥1 existing therapy is already accepted as effective. In such cases, comparison with placebo may be insufficient to support the use of the new treatment, and the provision of some active therapy to research subjects is usually considered an ethical imperative. Instead of superiority to a placebo, investigators may thus seek to show that a novel intervention is at least as good as an active, accepted alternative. For example, a new patient-warming system that is less expensive and easier to use might be considered an improvement over a conventional system, even if it does not warm patients any better.

The statistical approach in such cases is to demonstrate that the new treatment, while it may or may not be superior to its competitor, is not enough worse to clinically matter. Not enough worse, euphemistically termed noninferiority, is defined in terms of a clinically important noninferiority margin, a performance decrement defined as the boundary between a reduction in performance that is ignorable and one that is considered unacceptable. Noninferiority is demonstrated when a 1-sided confidence interval for a comparative measure of treatment efficacy excludes underperformance by greater than or equal to the noninferiority margin.

Equivalence trials are symmetric, bidirectional noninferiority trials: "not enough different" must be shown in both directions, such as by a 2-sided confidence interval that excludes sufficiently large differences favoring either of the treatments being compared. Although noninferiority studies are far more common, equivalence studies are needed when the purpose of the treatment is to regulate and maintain a physiologic parameter within an established target range, as for hormonal agents such as thyroxine and insulin or for anticoagulants such as heparin and warfarin. The relationships between superiority, equivalence, and noninferiority approaches are shown in Figure 2 (for additional detail, see a recent review.[19]).

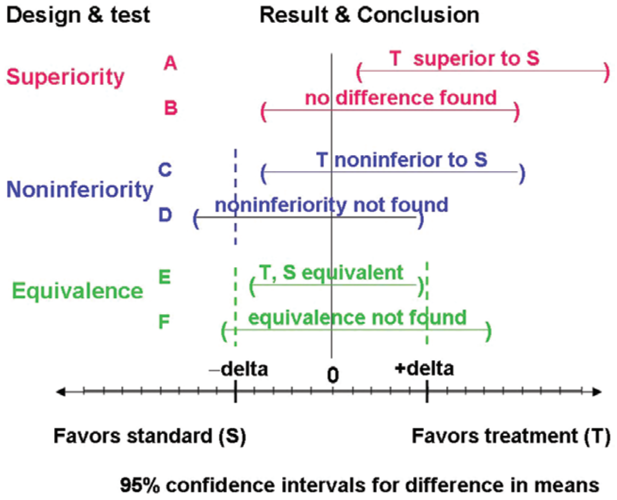

Figure 2. Sample confidence intervals and inference for trials assessing superiority, noninferiority, or equivalence of treatment *T* to standard *S*. NI = noninferiority. Notice, for example, that confidence intervals for results *B* and *C* are identical, but the permitted conclusions quite different because the designs and hypotheses differ. [Reprinted with modifications from Mascha EJ, Sessler DI. Equivalence and noninferiority testing in regression models and repeated-measures designs. Anesth Analg 2011;112:678–87.]

TRIAL SIZE

Trial size matters![20] In 1000 independent flips of a coin, the proportion of heads will very likely fall within ±3% of the expected proportion of 50% and thus give a precise estimate of the truth, meaning the coin's long-run proportion, or chance, of heads. Because flipping the coin 1000 times again would likely produce another result within the same range, the 2 sequences of coin flips will be fairly close, usually within 6% of one another.

But now consider flipping the same coin just 10 times. It would be unsurprising to observe a proportion of heads anywhere from 30% to 70% or for that matter from 20% to 80%. Nor would it be surprising to find that the proportion of heads in a second 10 flips departs considerably from the proportion among the first 10. Such a study (10 coin flips) therefore does not produce either a stable result from one repetition to another or a precise estimate of the truth: its results are fragile. This intuitive comparison illustrates the the probabilistic Central Limit Theorem that, other things being equal, the precision of a study improves in proportion to the square root of the number of subjects. Or, to put this in practical terms, the larger the study, the more likely it is that the results will closely reflect the target characteristics of the underlying biological system and be closely replicated if the study was repeated.[21]

Another reason trial size matters is that randomization only assures that the treatment groups will be comparable in sufficiently large trials. In large trials, treatment groups will almost always be very similar. But in small trials, by pure bad luck, the treatment groups may differ enough for such differences to distort (by confounding) the results. The number of patients required to provide reasonable assurance of baseline homogeneity is considerably larger than generally appreciated. For example, with 100 patients per group in a 2-arm trial in a population with men and women equally represented, the fractions of men and women will differ between groups by >10% in 15% of such trials and by >7% in a third of them.

As noted above, the number of patients needed for various types of studies depends on a host of factors, including the type of outcome, its incidence if dichotomous and variability if quantitative, the duration of follow-up, and the outcome difference between the experimental and control treatments (treatment effect) that the study is intended to detect. But larger trials increase the precision in estimating the treatment effects and always enhance the confidence in the results.[22] Large randomized trials often reverse medical practices that were based on small studies.[23]

It is worth considering that with a *P* value of 0.05, which is generally considered significant, the probability of a repeat trial being significant when the first trial has accurately estimated the true treatment is only 50%. The

P value needs to be 0.005 for this replication probability to reach the conventional power criterion of 80% and 0.0001 to reach 95%.[24] These criteria have recently been suggested as replacements for the 5% and 1% *P* value criteria by which results are currently conventionally labeled statistically significant, or highly statistically significant.[25] Note that significance conventionally means that, were there no real treatment effect, <5% of trials would be expected to produce a difference between outcomes of treatment and control groups as large as that observed (ignoring possible issues of bias). But that is not really what clinicians need to know; instead, they need bounds on the confidence intervals around the observed treatment effect that are narrow enough to be clinically useful. Tight confidence intervals result from sample sizes well exceeding those required to barely confirm statistical significance.

CLINICAL EQUIPOISE

The ethical basis for randomized trials is equipoise. The term means that investigators and oversight bodies collectively judge that the bases for favoring each treatment are roughly balanced, so that the groups receiving each treatment are equally likely to receive greater benefit, with benefit being broadly defined. For example, a novel intervention might be considered beneficial if it provided: (1) superior treatment effect; (2) comparable treatment effect with fewer side effects; (3) comparable treatment effect but is easier for patients or physicians to implement; or (4) comparable treatment effect at lower cost. Of course, many other cost-benefit combinations are possible.

Equipoise is generally achieved by comparing an experimental treatment with the best existing treatment, to which it may or may not prove superior. A major task for IRBs is to judge that there truly is equipoise and that patients in the control group used for comparison are getting the best current treatment and thus are not being disadvantaged by participation in the trial. Note that the control group should receive at least best current local practice and preferably best available treatment. Thus, placebo controls should be restricted to situations in which there is no generally accepted treatment or with suitable rescue with clearly effective drugs (as in many pain studies). For example, it would seem questionable these days to conduct an antiemetic trial in which high-risk control patients were given a placebo now that many safe and effective treatments are available.

An important point is that during a trial, individuals may have strong (and varied) opinions about which treatment is best. Nonetheless, participants and safety monitors may agree that the trial will provide crucial evidence for adjudicating these opinions. In other words, individual opinions and expectations do not preclude equipoise when there is a consensus that additional evidence is required. While it is tempting to assume that a novel treatment will be better (the source of much measurement bias), that turns out to be the case well under half of the time in substantial trials.[26] History documents that it is very hard to predict whether a novel treatment will actually prove superior in terms of efficacy and safety—which is exactly why trials are necessary.

THREATS TO VALIDITY

Randomization and blinding provide excellent protection against selection bias, confounding, and measurement bias. However, there are numerous other threats to trial validity that compromise conclusions. Consequently, the fraction of unrepeatable published trials is probably far higher than generally appreciated.[16,22] Care in design, conduct, and reporting of trials is associated with improved validity.[27]

Inadequate trial size, as discussed above, reduces trial validity. But there are many other subtle threats to individual studies and the medical literature in general. Among the most important is publication bias, which has considerable potential to seriously distort medical literature, and is hardly restricted to clinical trials.[28]

There are 3 main reasons why studies may not be published. The first is that the study might not even be completed. To the extent that investigators see negative results while the study is in progress, they may decide not to complete the project or a sponsor may decline to continue funding the project. (The bias inherent in this decision is why investigators and sponsors should normally be blinded to trial results, and why results should only be evaluated at predefined interim analysis points.) Trials may also be stopped because of inadequate enrollment.[29] The resulting underpowered study may not be publishable even if the investigators wish otherwise, but often they do not even try.

The second reason negative studies may not be published is that investigators often do not even submit negative results for peer review, concluding that the results are uninteresting or unlikely to be accepted. And to some extent, they may be correct that their work may not be accepted in their journal(s) of choice, which is the third reason negative results may remain unpublished. A more subtle problem is that corporate sponsors may deliberately restrict submission of negative results.[30]

Competent studies with negative results should be accepted for publication, but they may be given lower priority than positive results by reviewers and editors. The pervasive damage from publication bias is now better understood by editors who seem more willing to publish high-quality negative results. But to the extent that studies are stopped early without statistical justification, results are likely to be underpowered with an ambiguous rather than a truly negative result, reducing a publication's attractiveness to journal editors.

The primary result of publication bias is that, for 1 reason or another, negative results are less likely to appear in print than positive ones. Obviously, this distorts the literature and makes it more likely that a clinician will conclude—based on available evidence—that a treatment is effective when it actually is not. The problem is especially serious when available results are included in systematic reviews and meta-analyses that essentially assume that all results are available for analysis. Nonetheless, evidence suggests that meta-analyses of small trials are generally consistent with subsequent large trials.[31]

Design choices obviously influence the interpretation of study results. For example, using a placebo control rather than the best currently available treatment makes experimental interventions appear superior. Although considered unethical for serious diseases where there are effective alternatives, the practice continues because of pressure from sponsors and because the expected larger treatment effect allows smaller and less expensive trials. A more insidious approach is the use of control and experimental drug doses that are not comparable; to the extent that the experimental dose is functionally higher, its efficacy will presumably be enhanced.

Loss to follow-up (attrition bias) is another subtle way trial results can be compromised, consider, for example, a trial comparing an effective analgesic—say morphine—with an ineffective experimental treatment. Patients randomly assigned to the experimental treatment will find their analgesia inadequate and many will drop out of the study. And once out, they will no longer contribute pain scores or other outcome data. Those who remain will be the ones who have least pain or are otherwise pain tolerant. Their pain scores will typically be considerably lower than those of patients on the experimental therapy who dropped out. Average results for patients who remained in the study will thus far overestimate the analgesic benefit of the experimental treatment, perhaps even making it appear comparable with morphine.

Intention-to-treat analysis is a widely used strategy for countering such postrandomization biases, which, in effect, break the balance between treatment groups fostered by the randomization process. The intent-to-treat concept is neatly captured by the catch-phrase "Where randomized, there analyzed." This analytic philosophy retains patients in the group to which they were randomly assigned, no matter what occurs during the trial. Patients who drop out of the study before the recording of an outcome are analyzed by using an imputed outcome, preferably by a multiple imputation technique grounded in statistical theory but sometimes in practice by an average value in other patients or that patient's last observation carried forward. Patients who remain in a study but are noncompliant with the assigned treatment, even those who switch to another treatment and/or never receive the one randomly assigned, are analyzed as members of their original randomized group.

The intention-to-treat approach is counter intuitive, but well justified when understood. Retaining randomized groups as originally constituted assures that, in the analysis, departures from assigned treatments and differential dropout will not exaggerate true treatment differences. In contrast, if dropouts and treatment crossovers are excluded from analysis, or crossovers analyzed in the groups to which they move, or other ad hoc approaches used, biases of unknown magnitude and direction may occur. Intention-to-treat analyses thus protect against exaggeration of a true treatment effect, or even production of a spurious one, at the expense of potentially underestimating treatment effects and diminishing study power. But because intention-to-treat analyses also attenuate differences in harm, they are generally used only for efficacy comparisons, with adverse events compared between groups reflecting treatments actually received. Comparisons on this basis are termed per-protocol analyses and are often used for secondary analyses.

CONSORT REPORTING GUIDELINES

Various checklists and guidelines have been proposed for reporting of clinical trials, with the Consolidated Standards of Reporting Trials (CONSORT) guidelines being by far the most commonly used. Some journals now require authors to submit a specific checklist documenting that all relevant elements are addressed in their articles. But, with or without a formal checklist, skilled reviewers will expect to see relevant critical components addressed.[32,33]

The CONSORT checklist includes 25 items, many with subelements. Among these, perhaps the most important are (1) type of trial (i.e., parallel group, factorial); (2) important changes after starting enrollment; (3) eligibility criteria; (4) interventions, with sufficient detail to allow replication; (5) completely defined prespecified primary and secondary outcomes; (6) interim analyses, stopping rules, and sample size determination and justification; (7) how randomization was generated and allocation concealed; (8) blinding; (9) statistical methods; (10) subject enrollment, participation, and loss (usually presented as a diagram); (11) why the study ended, if not taken to completion; (12) baseline characteristics for each group (usually a table); (13) outcomes with appropriate measures of variance; (14) harms in each group; and (15) public trial registration (registry and number).

The most obvious way for investigators to be sure of being able to provide each of the CONSORT elements is to incorporate them into protocols. Investigators can also consult the Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT) checklist during the trial design phase.[34]

CONCLUSIONS

Randomized assignment of treatment excludes reverse causation and selection bias and, in sufficiently large studies, effectively prevents confounding. Well-implemented blinding prevents measurement bias. Studies that include these protections are called randomized, blinded clinical trials and, when conducted with sufficient numbers of patients, provide the most valid clinical research results. Although conceptually straightforward, design of clinical trials requires thoughtful trade-offs among competing approaches—all of which influence the number of patients required, enrollment time, internal and external validity, ability to evaluate interactions among treatments, and cost.

Because randomized trials are so expensive and time consuming, the number of patients required is an overriding concern. Many trials evaluate mediators or intermediate outcomes, often characterized by continuous measurements. Hard outcomes (i.e., myocardial infarction, respiratory arrest, or death) that are much more serious and interesting are usually dichotomous and thus require far more patients. Subject selection also influences the sample size by increasing baseline event rates or reducing variability.

Crossover designs reduce required sample size but are limited to short-term interventions with minimal carry-over effects. Factorial designs are efficient, have the capacity to detect interactions, and can be used with any type of outcome. And novel study designs, such as alternating intervention approaches, can speed enrollment and enhance efficiency. The field of clinical trial design is rapidly evolving, and this review has attempted only a basic introduction. Estimating the number of patients required for a trial can range from trivial to extraordinarily complex depending on the trial design and is further complicated by inclusion of interim analyses. Even beyond the basics of selecting a valid design for a particular problem, choices made within the context of a general trial design have considerable potential for enhancing a trial's strength and feasibility. ■

DISCLOSURES

Name: Daniel I. Sessler, MD.
Contribution: This author helped write the manuscript.
Attestation: Daniel I. Sessler approved the final manuscript.
Name: Peter B. Imrey, PhD.
Contribution: This author helped design the study and write the manuscript.
Attestation: Peter B. Imrey approved the final manuscript.
This manuscript was handled by: Steven L. Shafer, MD.

REFERENCES

1. Devereaux PJ, Yusuf S. The evolution of the randomized controlled trial and its role in evidence-based decision making. J Intern Med 2003;254:105–13
2. Kaptchuk TJ, Friedlander E, Kelley JM, Sanchez MN, Kokkotou E, Singer JP, Kowalczykowski M, Miller FG, Kirsch I, Lembo AJ. Placebos without deception: a randomized controlled trial in irritable bowel syndrome. PLoS One 2010;5:e15591
3. Deans KJ, Minneci PC, Danner RL, Eichacker PQ, Natanson C. Practice misalignments in randomized controlled trials: Identification, impact, and potential solutions. Anesth Analg 2010;111:444–50
4. Deans KJ, Minneci PC, Suffredini AF, Danner RL, Hoffman WD, Ciu X, Klein HG, Schechter AN, Banks SM, Eichacker PQ, Natanson C. Randomization in clinical trials of titrated therapies: unintended consequences of using fixed treatment protocols. Crit Care Med 2007;35:1509–16
5. Sessler DI, Devereaux PJ. Emerging trends in clinical trial design. Anesth Analg 2013;116:258–61
6. Devereaux PJ, Mrkobrada M, Sessler DI, Leslie K, Alonso-Coello P, Kurz A, Villar JC, Sigamani A, Biccard BM, Meyhoff CS, Parlow JL, Guyatt G, Robinson A, Garg AX, Rodseth RN, Botto F, Lurati Buse G, Xavier D, Chan MT, Tiboni M, Cook D, Kumar PA, Forget P, Malaga G, Fleischmann E, Amir M, Eikelboom J, Mizera R, Torres D, Wang CY, VanHelder T, Paniagua P, Berwanger O, Srinathan S, Graham M, Pasin L, Le Manach Y, Gao P, Pogue J, Whitlock R, Lamy A, Kearon C, Baigent C, Chow C, Pettit S, Chrolavicius S, Yusuf S; POISE-2 Investigators. Aspirin in patients undergoing noncardiac surgery. N Engl J Med 2014;370:1494–503
7. Devereaux PJ, Sessler DI, Leslie K, Kurz A, Mrkobrada M, Alonso-Coello P, Villar JC, Sigamani A, Biccard BM, Meyhoff CS, Parlow JL, Guyatt G, Robinson A, Garg AX, Rodseth RN, Botto F, Lurati Buse G, Xavier D, Chan MT, Tiboni M, Cook D, Kumar PA, Forget P, Malaga G, Fleischmann E, Amir M, Eikelboom J, Mizera R, Torres D, Wang CY, Vanhelder T, Paniagua P, Berwanger O, Srinathan S, Graham M, Pasin L, Le Manach Y, Gao P, Pogue J, Whitlock R, Lamy A, Kearon C, Chow C, Pettit S, Chrolavicius S, Yusuf S; POISE-2 Investigators. Clonidine in patients undergoing noncardiac surgery. N Engl J Med 2014;370:1504–13
8. Apfel CC, Korttila K, Abdalla M, Kerger H, Turan A, Vedder I, Zernak C, Danner K, Jokela R, Pocock SJ, Trenkler S, Kredel M, Biedler A, Sessler DI, Roewer N; IMPACT Investigators. A factorial trial of six interventions for the prevention of postoperative nausea and vomiting. N Engl J Med 2004;350:2441–51
9. Landsberger H. Hawthorne Revisited: Management and the Worker: Its Critics, and Developments in Human Relations in Industry. Ithaca, NY: Cornell University Press
10. Huang SS, Septimus E, Kleinman K, Moody J, Hickok J, Avery TR, Lankiewicz J, Gombosev A, Terpstra L, Hartford F, Hayden MK, Jernigan JA, Weinstein RA, Fraser VJ, Haffenreffer K, Cui E, Kaganov RE, Lolans K, Perlin JB, Platt R; CDC Prevention Epicenters Program; AHRQ DECIDE Network and Healthcare-Associated Infections Program. Targeted versus universal decolonization to prevent ICU infection. N Engl J Med 2013;368:2255–65
11. Kappen TH, Moons KG, van Wolfswinkel L, Kalkman CJ, Vergouwe Y, van Klei WA. Impact of risk assessments on prophylactic antiemetic prescription and the incidence of postoperative nausea and vomiting: a cluster-randomized trial. Anesthesiology 2014;120:343–54
12. Kopyeva T, Sessler DI, Weiss S, Dalton JE, Mascha EJ, Lee JH, Kiran RP, Udeh B, Kurz A. Effects of volatile anesthetic choice on hospital length-of-stay: a retrospective study and a prospective trial. Anesthesiology 2013;119:61–70
13. Devereaux PJ, Yang H, Yusuf S, Guyatt G, Leslie K, Villar JC, Xavier D, Chrolavicius S, Greenspan L, Pogue J, Pais P, Liu L, Xu S, Malaga G, Avezum A, Chan M, Montori VM, Jacka M, Choi P: Effects of extended-release metoprolol succinate in patients undergoing non-cardiac surgery (POISE trial): a randomised controlled trial. Lancet 2008;371:1839–47
14. Mascha EJ, Imrey PB. Factors affecting power of tests for multiple binary outcomes. Stat Med 2010;29:2890–904
15. Mascha EJ, Sessler DI: Design and analysis of studies with binary-event composite endpoints. Anesth Analg 2011;112:1461–71
16. Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. JAMA 2005;294:218–28
17. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Stat Med 2002;21:2917–30
18. Mascha EJ, Turan A. Joint hypothesis testing and gatekeeping procedures for studies with multiple endpoints. Anesth Analg 2012;114:1304–17
19. Mascha EJ, Sessler DI. Equivalence and noninferiority testing in regression models and repeated-measures designs. Anesth Analg 2011;112:678–87
20. Devereaux PJ, Chan MT, Eisenach J, Schricker T, Sessler DI. The need for large clinical studies in perioperative medicine. Anesthesiology 2012;116:1169–75
21. Walsh M, Srinathan SK, McAuley DF, Mrkobrada M, Levine O, Ribic C, Molnar AO, Dattani ND, Burke A, Guyatt G, Thabane L, Walter SD, Pogue J, Devereaux PJ. The statistical significance of randomized controlled trial results is frequently fragile: a case for a Fragility Index. J Clin Epidemiol 2014;67:622–8
22. Ioannidis JP. Why most published research findings are false. PLoS Med 2005;2:e124
23. Prasad V, Vandross A, Toomey C, Cheung M, Rho J, Quinn S, Chacko SJ, Borkar D, Gall V, Selvaraj S, Ho N, Cifu A. A decade of reversal: an analysis of 146 contradicted medical practices. Mayo Clin Proc 2013;88:790–8
24. Goodman SN. A comment on replication, p-values and evidence. Stat Med 1992;11:875–9
25. Johnson VE. Revised standards for statistical evidence. Proc Natl Acad Sci U S A 2013;110:19313–7
26. Gordon D, Taddei-Peters W, Mascette A, Antman M, Kaufmann PG, Lauer MS. Publication of trials funded by the National Heart, Lung, and Blood Institute. N Engl J Med 2013;369:1926–34
27. Nowbar AN, Mielewczik M, Karavassilis M, Dehbi HM, Shun-Shin MJ, Jones S, Howard JP, Cole GD, Francis DP; DAMASCENE Writing Group. Discrepancies in autologous bone marrow stem cell trials and enhancement of ejection fraction (DAMASCENE): weighted regression and meta-analysis. BMJ 2014;348:g2688
28. Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan AW, Cronin E, Decullier E, Easterbrook PJ, Von Elm E, Gamble C, Ghersi D, Ioannidis JP, Simes J, Williamson PR. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. PLoS One 2008;3:e3081
29. Kasenda B, von Elm E, You J, Blümle A, Tomonaga Y, Saccilotto R, Amstutz A, Bengough T, Meerpohl JJ, Stegert M, Tikkinen KA, Neumann I, Carrasco-Labra A, Faulhaber M, Mulla SM, Mertz D, Akl EA, Bassler D, Busse JW, Ferreira-González I, Lamontagne F, Nordmann A, Gloy V, Raatz H, Moja L, Rosenthal R, Ebrahim S, Schandelmaier S, Xin S, Vandvik PO, Johnston BC, Walter MA, Burnand B, Schwenkglenks M, Hemkens LG, Bucher HC, Guyatt GH, Briel M. Prevalence, characteristics, and publication of discontinued randomized trials. JAMA 2014;311:1045–51

30. Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R. Selective publication of antidepressant trials and its influence on apparent efficacy. N Engl J Med 2008;358:252–60
31. Cappelleri JC, Ioannidis JP, Schmid CH, de Ferranti SD, Aubert M, Chalmers TC, Lau J. Large trials vs meta-analysis of smaller trials: how do their results compare? JAMA 1996;276:1332–8
32. Schulz KF, Altman DG, Moher D; CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. BMC Med 2010;8:18
33. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, Elbourne D, Egger M, Altman DG. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c869
34. Chan AW, Tetzlaff JM, Gøtzsche PC, Altman DG, Mann H, Berlin JA, Dickersin K, Hróbjartsson A, Schulz KF, Parulekar WR, Krleza-Jeric K, Laupacis A, Moher D. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. BMJ 2013;346:e7586