Emerging Trends in Clinical Trial Design

Daniel I. Sessler, MD* and P. J. Devereaux, MD, PhD†

Randomized trials are now routine, and large trials are, appropriately, considered the highest level of evidence in research addressing therapy and prevention questions. But it is worth noting that despite thousands of years of medical practice and observation, clinical trials are a relatively new concept. What is credited as the first randomized trial was a trial that evaluated streptomycin in patients with pulmonary tuberculosis in 1948.[1] Therefore, randomized trial methodology was initiated only 6 decades ago. As might be expected from such a young field, the pace of advancement remains high. In this editorial, we discuss 3 emerging trends in clinical trial design.

The techniques we discuss may be helpful to investigators who are designing randomized trials. They may also be helpful to department leaders who need to allocate resources and guide junior investigators. And finally, an understanding of the methods will also enhance clinicians' ability to critically read and understand studies on which practice decisions might be based.

TRIAL SIZE

The first, and perhaps the most important, major trend is that randomized trials are getting larger. It was only a decade ago that anesthesia trials with a few hundred patients were considered large; however, perioperative trials involving thousands of patients are now redefining what represents a large trial. Several factors are driving this change. One major reason is the recognition that in most situations we can only plausibly expect moderately sized treatment effects (i.e., relative risk reductions on the order of 25%). The etiology of most major perioperative complications is multifactorial. Therefore, it is probably unwise to expect an intervention that typically affects one pathway to result in anything larger than a moderate treatment effect.

For example, perioperative myocardial infarctions presumably have many triggers (i.e., inflammation, hypercoagulation, platelet activation, sympathetic activation, and hypoxia). It is thus unlikely that a single intervention such as a β-blocker, which primarily blocks just one of these triggers, would produce more than a moderate treatment effect. Consistent with this theory, the only large trial of perioperative β-blocker administration showed a relative risk reduction of 27% for perioperative myocardial infarction.[2]

Compounding the sample-size requirements for identifying moderate as opposed to large treatment effects is the fact that investigators are now tackling more clinically important, but less common, outcomes. In particular, many of the most important outcomes are both dichotomous and occur in 2% to 10% of patients, such as myocardial infarction, stroke, pneumonia, sepsis, and death. Sample-size estimates in the setting of expected moderately sized treatment effects with event rates ≤10% indicate that large numbers of patients are required under these conditions. For example, 4300 patients are required to have even 80% power for detecting a 25% risk reduction for a dichotomous outcome with a baseline incidence of 8%. The number of patients required increases to 6450 for a 20% risk reduction, and to 10,850 for a 15% risk reduction, either of which might still be clinically important treatment effects.

A commonly unappreciated issue is that small trials, even when statistically significant, often prove to be wrong.[3] This is the concept of *fragility*.[4] Fragility is characterized by substantial changes in *P* values with small changes in the number of patients experiencing an event in the treatment group. Clinicians should be cautious about statistically significant results that demonstrate fragility, because the results might easily no longer be significant if the trial were repeated.[3] Table 1 highlights how a small change of one patient experiencing an event in the treatment group results in a *P* value change from 0.04 to 0.09. Unfortunately, the statistically significant results of many small trials demonstrate substantial fragility and warrant cautious interpretation.

An additional problem is that many trials are statistically significant without providing useful guidance to clinicians.[5] The term "statistically significant" conventionally means there is only a 5% or lower probability that the difference in the outcomes between the treatment and control groups is due to chance (ignoring possible issues of bias). But in marginally powered studies, the 95% confidence limits on a relative risk reduction often extend over large ranges. For example, a clinical trial of a new investigational drug demonstrates a relative risk reduction of 45% with confidence intervals that extend from 4% to 90%. This statistically significant result only suggests that if the trial were repeated 100 times, clinicians could expect that 95 of the trials would demonstrate a relative risk reduction between 4% to 90%.

From the *Department of Outcomes Research, Cleveland Clinic, Cleveland, Ohio; And the †Departments of Clinical Epidemiology and Biostatistics and and Medicine and the Population Health Research Institute, McMaster University, Hamilton, Ontario, Canada.

Accepted for publication August 16, 2012.

See Disclosures at end of article for author conflicts of interest.

Reprints will not be available from the authors.

Address correspondence to Daniel I. Sessler, MD, Department of Outcomes Research, Cleveland Clinic, 9500 Euclid Ave., P-77, Cleveland, OH 44195. Address e-mail to ds@or.org. On the World Wide Web: www.OR.org.

DOI: 10.1213/ANE.0b013e318270036d

Table 1. Results Demonstrating Fragility of Results in Small Trial

Outcome	Treatment (*n* = 200)	Placebo (*n* = 200)	*P*
Trial 1			
Myocardial infarction	2	10	0.04
Trial 2			
Myocardial infarction	3	10	0.09

Adding just a single infarction to 2 otherwise identical small trials converts the result from statistically significant (*P* = 0.04) to nonsignificant (*P* = 0.09). These results are thus fragile and provide limited guidance as to the reliability of the treatment. Note that fragility remained despite having 400 patients in each trial because the outcome incidence was only approximately 3%.

Table 2. Example of a Factorial Randomized Controlled Trial

	Clonidine active	Clonidine placebo
Aspirin active	Clonidine active Aspirin active	Clonidine placebo Aspirin active
Aspirin placebo	Clonidine active Aspirin placebo	Clonidine placebo Aspirin placebo

Table 3. Effect of Antagonistic Interactions on Sample Size

Primary outcome			Power (2-sided α = 0.05)	
Event rate	Subadditivity	Hazard ratio	*n* = 10,000	*n* = 11,000
5.6%	0%	0.75	81.1%	84.6%
6.1%	0%	0.75	84.3%	87.5%
5.6%	10%	0.78	77.6%	81.4%
	25%	0.80	69.9%	73.9%
6.1%	10%	0.78	81.1%	84.6%
	25%	0.80	73.5%	77.5%

A synergistic interaction increases trial power, whereas an antagonistic interaction decreases power. Twenty-five percent antagonism, for example, means that the relative reduction for 1 treatment is 25% divided by 4 in the factorial cell where both medications are given together with the observed event rate in that cell equal to (1–0.25) × (1–[0.25 × 0.75]) times the double placebo event rate. The net effect is to reduce the relative reduction observed at the margin and to increase the hazard ratio (HR). An HR of 0.75 in the setting of 10% antagonism increases to an HR of 0.76, and to 0.78 with 25% antagonism. Although antagonism should be considered in sample-size estimates if likely, the effect it has on sample-size requirements is substantially less than the sample size required to undertake 2 separate trials. In this table, we model antagonism of 0%, 10%, and 25% on trial power with an outcome event rate typical for major dichotomous outcomes such as myocardial infarction or surgical site infection. The effects are small.

Some clinicians may find this level of precision an unconvincing guide to clinical practice, especially if the treatment is more expensive and has a poorly characterized side effect profile.

Accruing large sample sizes overcomes the problems discussed above; however, large trials usually require multiple centers that increase the complexity and cost of the trials. The inclusion of multiple centers does, however, increase geographic and demographic diversity and improve applicability of the results. The anesthesia community should thus expect and even demand large trials that provide robust and generalizable conclusions on which to base clinical practice.

FACTORIAL RANDOMIZATION

The second major trend in randomized trials is toward factorial randomization.[6] Factorial randomization consists of simultaneously randomizing patients to 2 or more treatments, each with their own control intervention in a single trial. Table 2 reports the interventions we are currently evaluating in a large, international factorial randomized controlled trial. In this trial, the randomization process will allocate 25% of patients to receive active clonidine and active aspirin, 25% of patients to receive active clonidine and aspirin placebo, 25% to receive clonidine placebo and active aspirin, and 25% to receive clonidine placebo and aspirin placebo. Although 2-intervention factorial designs are most common, factorial trials can evaluate 3 or more interventions.[7]

Factorial designs differ from multigroup trials in that factorial designs are substantially more efficient. For example, an alternative to the factorial design in Table 2 would be a trial that randomized patients to 1 of 3 groups (i.e., active clonidine, active aspirin, or placebo). The substantial advantage a factorial design has over this multigroup design is that the factorial trial sample-size requirement is substantially lower. This occurs because all patients in the factorial design are acting as an active or control for each intervention, whereas one third of the patients in the multigroup trial are not acting as an active or control intervention for each of the active interventions.

Because treatment allocation is completely balanced in a factorial design, it is statistically appropriate to assess the marginal effect of each drug on the outcome of interest. Effectively, this consists of testing the effect of each drug on the entire population, which is valid because the alternative factor is intentionally balanced. The alternative of comparing drug A to placebo and drug B to placebo is far less efficient, because only two thirds of the population participates in each analysis. From a practical perspective, investigators can thus answer 2 questions simultaneously in the same population. And because the marginal effects are evaluated independently, the sample-size requirement (assuming no interaction and moderate treatment effects) is not much larger than it would be for a trial of either drug alone.

Synergistic interactions increase trial power, whereas subadditive interactions decrease trial power. However, subadditive interactions are rare.[8] And even when they occur, the total number of patients required in a factorial design is far less than that required for 2 independent trials (Table 3).

An additional advantage of factorial randomization is that it allows investigators to evaluate the interaction among treatments. Even the largest independent trials do not provide information about responses to drug combinations (whether beneficial or harmful). Specifically, it is impossible to determine from independent trials whether drug interactions on a given outcome are antagonistic, additive, or synergistic, which in some circumstances is a critically important clinical question. Adequately powered factorial designs can, however, inform whether there are important interactions.[9] A caveat, though, is that powering trials to assess small- to moderate-sized interactions can require more patients than for marginal effects alone. Investigators designing factorial trials thus need to consider the importance of drug interactions and the extent to which

Table 4. Major Benefits and Limitations of Large Sample Size, Composite Outcomes, and Factorial Randomization

	Benefits	Limitations
Large sample size	Ability to evaluate small, but clinically important, treatment effects	Greater cost
	Better estimation of treatment effect	Longer data-acquisition periods
	Less fragile results	May require multiple centers
Factorial randomization	Efficient design (simultaneously testing 2 or more hypotheses)	Modest increase in sample size for main effects
	Ability to evaluate interactions among treatments	Contraindications to multiple interventions can reduce the number of eligible patients
		Patients may feel less inclined to participate when 2 experimental interventions are undergoing evaluation
		Need to increase the sample size to evaluate negative interactions; however, this sample-size requirement is almost universally less than the sample size required to conduct 2 separate trials.
Composite outcomes	Better characterization of disease influencing multiple systems	Component heterogeneity
	Lower sample size	

sufficiently increasing sample size is justified. Often it is not and, in fact, a recent meta-analysis found that only 18% of factorial design trials were powered to detect interactions.[10]

Factorial trials have substantial advantages in terms of cost, efficiency (answering 2 or more questions at the same time), avoiding competition for patients, and the ability to assess interactions. The trend toward factorial designs is thus an important and positive advance in the conduct of clinical trials.

COMPOSITE OUTCOMES

The third trend in clinical trials is the increasing use of composite outcomes. A composite outcome consists of several component outcomes (e.g., death, nonfatal myocardial infarction, nonfatal stroke). A participant is considered to have experienced the composite outcome if 1 or more component outcomes occur. The major reason researchers use composite outcomes is to reduce sample size and simplify the presentation of data.[11]

The expected benefits and harms of an intervention may best be characterized as a composite. For example, tight glucose control in diabetics may reduce the risk of kidney failure, blindness, amputation, myocardial infarction, and stroke. All are clinically important outcomes and could influence a clinician's decision to use a therapy. There may be little clinical basis for designating any single outcome as the primary outcome, and in all but the largest trials, it would also be prohibitive to take the necessary statistical penalty for defining 5 primary outcomes. Composite outcomes thus allow investigators to characterize a broad spectrum of important treatment-induced benefits or harms.

The statistical effect is 2-fold: (1) composite outcomes eliminate the statistical penalty that would otherwise be required for considering multiple outcomes simultaneously; and (2) combining various outcomes increases the event rate. Both effects reduce the number of patients required for a given degree of statistical power. Improved power is probably the major factor accounting for increased popularity of composite outcomes; however, as noted above, small trials have substantial disadvantages.

Ideally, component outcomes in a composite should be of comparable severity. It makes no sense to effectively average together serious and minor outcomes. For example, an infection composite might reasonably include deep sternal wound infections, organ space infection, ventilator-associated pneumonia, and sepsis, all of which are life threatening. But it would be unwise to add urinary tract infections, which are usually minor and are easily treated.

Components of a composite outcome should also occur at approximately similar rates. If the incidence of 1 component is especially large, it will overwhelm the others, effectively becoming the primary outcome. For example, a composite of adverse events after cardiac surgery might reasonably include postoperative mechanical circulatory support, serious infection, new-onset dialysis, and stroke. Each is not only of comparable severity but may also occur at similar rates. Adding atrial fibrillation, as is often done,[12] weakens the composite in 2 ways. First, it is far less serious than the other complications. An equally important consideration, though, is that atrial fibrillation is far more common than all the other components combined. Including atrial fibrillation thus converts a rational composite that accurately characterized serious complications to an essentially unitary result that hardly differed from atrial fibrillation alone. A better approach is to retain the original composite, presumably as the primary outcome, and consider atrial fibrillation and other outcomes of interest to be secondary outcomes. Although components of varying severity can be included in a composite outcome through weighting for severity using statistical techniques,[13] the weighting may not be accepted by clinicians or their patients.

The majority of composite outcomes assume homogeneity of effect across the component outcomes. A remaining potential problem with composite outcomes is heterogeneity of effect when homogeneity is assumed. An example of heterogeneous outcomes is the POISE trial in which metoprolol reduced the risk of myocardial infarction (relative risk reduction 27%), but simultaneously doubled the risk of devastating strokes.[2] Simply arithmetically combining these divergent results into a single composite ignores the underlying heterogeneity of the treatment effect and could easily lead to incorrect clinical conclusions. Individual component results for each element of a composite outcome should always be presented, so readers can judge which components contributed most and identify serious heterogeneity. The difficulty is that few trials with composite outcomes are sufficiently powered to allow valid quantification of divergent responses, which once again speaks to the importance and benefit of the first trend we discussed, large trials.

CONCLUSIONS

In summary, 3 major trends in clinical trials are toward large size, factorial randomization, and composite outcomes. Aside from cost and coordination issues, large sample size always improves the reliability of trial results. There is now compelling evidence that small trials, even when statistically significant, are often fragile, thus potentially misleading clinicians.

Factorial designs are statistically efficient in that 2 or more interventions can be simultaneously evaluated with only slightly more patients than would otherwise be required to test either intervention alone. When adequately powered, factorial trials can characterize interactions among drugs, information that cannot otherwise be determined from even the largest separate trials. The use of composite outcomes in clinical trials is increasing in frequency. Several factors need to be considered in selecting a composite to ensure a rational composite to inform clinical practice. The major benefits and limitations of each approach are shown in Table 4.

We note that factorial designs and composite outcomes are powerful but complicated techniques. This brief review is designed to alert investigators and department leaders to the methods; it is not a comprehensive tutorial and does not identify all potential pitfalls. Investigators planning to use either approach should seek additional information and, preferably, collaborate with experts.

Possibly the first illustration of the 3 trends in anesthesia was the IMPACT trial in which a 6-way factorial randomization was used to evaluate the composite of postoperative nausea and vomiting in more than 5000 patients.[7] An ongoing example is the POISE-2 trial in which a 2-way factorial randomization to clonidine and/or aspirin is being evaluated on a composite of death and myocardial infarction outcomes in 10,000 patients.[14] Because of their advantages, investigators are increasingly likely to choose designs that include large sample size and factorial randomization, and where appropriate, composite outcomes. ▪

DISCLOSURES

Name: Daniel I. Sessler, MD.
Contribution: This author helped design and conduct the study, analyze the data, and write the manuscript.
Attestation: Daniel I. Sessler approved the final manuscript.
Conflicts of Interest: The author has no conflicts of interest to declare.
Name: P. J. Devereaux, MD, PhD.
Contribution: This author helped design and conduct the study, analyze the data, and write the manuscript.
Attestation: P. J. Devereaux approved the final manuscript.
Conflicts of Interest: P. J. Devereaux is supported by a Heart and Stroke Foundation of Ontario, Canada, Career Investigator Award but does not have personal financial interest in this work.
This manuscript was handled by: Franklin Dexter, MD, PhD.

REFERENCES

1. MRC Strepromycin in Tuberculosis Trial Committee. Streptomycin treatment for pulmonary tuberculosis. BMJ 1948;ii:769–82
2. 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, Málaga G, Avezum A, Chan M, Montori VM, Jacka M, Choi P; POISE Study Group. Effects of extended-release metoprolol succinate in patients undergoing non-cardiac surgery (POISE trial): a randomised controlled trial. Lancet 2008;371: 1839–47
3. Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. JAMA 2005;294:218–28
4. Devereaux PJ, Chan MT, Eisenach J, Schricker T, Sessler DI. The need for large clinical studies in perioperative medicine. Anesthesiology 2012;116:1169–75
5. Gibbs NM, Weightman WM. Beyond effect size: consideration of the minimum effect size of interest in anesthesia trials. Anesth Analg 2012;114:471–5
6. McAlister FA, Straus SE, Sackett DL, Altman DG. Analysis and reporting of factorial trials: a systematic review. JAMA 2003;289:2545–53
7. 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
8. McAlister FA, Straus SE, Sackett DL, Altman DG. Analysis and reporting of factorial trials: a systematic review. JAMA 2003;289:2545–53
9. Montgomery AA, Peters TJ, Little P. Design, analysis and presentation of factorial randomised controlled trials. BMC Med Res Methodol 2003;3:26
10. Montgomery AA, Astin MP, Peters TJ. Reporting of factorial trials of complex interventions in community settings: a systematic review. Trials 2011;12:179
11. Ferreira-González I, Busse JW, Heels-Ansdell D, Montori VM, Akl EA, Bryant DM, Alonso-Coello P, Alonso J, Worster A, Upadhye S, Jaeschke R, Schünemann HJ, Permanyer-Miralda G, Pacheco-Huergo V, Domingo-Salvany A, Wu P, Mills EJ, Guyatt GH. Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. BMJ 2007;334:786
12. Karthikeyan G, Moncur RA, Levine O, Heels-Ansdell D, Chan MT, Alonso-Coello P, Yusuf S, Sessler D, Villar JC, Berwanger O, McQueen M, Mathew A, Hill S, Gibson S, Berry C, Yeh HM, Devereaux PJ. Is a pre-operative brain natriuretic peptide or N-terminal pro-B-type natriuretic peptide measurement an independent predictor of adverse cardiovascular outcomes within 30 days of noncardiac surgery? A systematic review and meta-analysis of observational studies. J Am Coll Cardiol 2009;54:1599–606
13. Mascha EJ, Sessler DI. Statistical grand rounds: design and analysis of studies with binary- event composite endpoints: guidelines for anesthesia research. Anesth Analg 2011;112: 1461–71
14. Devereaux PJ, Yang H, Guyatt GH, Leslie K, Villar JC, Monteri VM, Choi P, Giles JW, Yusuf S; POISE Trial Investigators. Rationale, design, and organization of the PeriOperative ISchemic Evaluation (POISE) trial: a randomized controlled trial of metoprolol versus placebo in patients undergoing noncardiac surgery. Am Heart J 2006;152:223–30