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. 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.

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. This is the concept of fragility. 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. 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. 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%.
Emerging Trends in Clinical Trial Design

**Table 1. Results Demonstrating Fragility of Results in Small Trial**

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Treatment (n = 200)</th>
<th>Placebo (n = 200)</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trial 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Myocardial infarction</td>
<td>2</td>
<td>10</td>
<td>0.04</td>
</tr>
<tr>
<td>Trial 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Myocardial infarction</td>
<td>3</td>
<td>10</td>
<td>0.09</td>
</tr>
</tbody>
</table>

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**

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Clonidine active</th>
<th>Clonidine placebo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspirin active</td>
<td>Clonidine active</td>
<td>Clonidine placebo</td>
</tr>
<tr>
<td>Aspirin active</td>
<td>Aspirin active</td>
<td>Aspirin active</td>
</tr>
<tr>
<td>Aspirin placebo</td>
<td>Clonidine active</td>
<td>Clonidine placebo</td>
</tr>
<tr>
<td>Aspirin placebo</td>
<td>Aspirin placebo</td>
<td></td>
</tr>
</tbody>
</table>

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 anesthesiology 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. 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.

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 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) × (1–0.75у) 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.

**Table 3. Effect of Antagonistic Interactions on Sample Size**

<table>
<thead>
<tr>
<th>Primary outcome</th>
<th>Power (2-sided α = 0.05)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event rate</td>
<td>Subadditivity</td>
</tr>
<tr>
<td>5.6%</td>
<td>0.75</td>
</tr>
<tr>
<td>6.1%</td>
<td>0.75</td>
</tr>
<tr>
<td>5.6%</td>
<td>0.78</td>
</tr>
<tr>
<td>6.1%</td>
<td>0.78</td>
</tr>
<tr>
<td>25%</td>
<td>0.80</td>
</tr>
</tbody>
</table>

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) × (1–0.75у) 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.
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 com-
posite 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 com-
ponent 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 out-
comes thus allow investigators to characterize a broad spec-
trum 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 prob-
ably 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 exam-
ple, an infection composite might reasonably include deep
sternal wound infections, organ space infection, ventila-
tor-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 com-
ponent 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 consid-
eration, though, is that atrial fibrillation is far more common
than all the other components combined. Including atrial
fibrillation thus converts a rational composite that accu-
rately characterized serious complications to an essentially
unitary result that hardly differed from atrial fibrillation
alone. A better approach is to retain the original compos-
ite, 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 homogene-
ity of effect across the component outcomes. A remaining
potential problem with composite outcomes is heterogene-
ity of effect when homogeneity is assumed. An example of
heterogeneous outcomes is the POISE trial in which meto-
prolol 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 com-
ponent results for each element of a composite outcome
should always be presented, so readers can judge which

<table>
<thead>
<tr>
<th>Table 4. Major Benefits and Limitations of Large Sample Size, Composite Outcomes, and Factorial Randomization</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Benefits</strong></td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>Ability to evaluate small, but clinically important, treatment effects</td>
</tr>
<tr>
<td>Better estimation of treatment effect</td>
</tr>
<tr>
<td>Less fragile results</td>
</tr>
<tr>
<td>Efficient design (simultaneously testing 2 or more hypotheses)</td>
</tr>
<tr>
<td>Ability to evaluate interactions among treatments</td>
</tr>
<tr>
<td>Better characterization of disease influencing multiple systems</td>
</tr>
<tr>
<td>Lower sample size</td>
</tr>
</tbody>
</table>

---

1. The Open Mind
2. ANESTHESIA & ANALGESIA
3. www.anesthesia-analgesia.org

Emerging Trends in Clinical Trial Design

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. 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. 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

4. Devereaux PJ, Chan MT, Eisenach J, Schriger T, Sessler DI. The need for large clinical studies in perioperative medicine. Anesthesiology 2012;116:1169-75