

Negative Trials, and What to Do with Them?

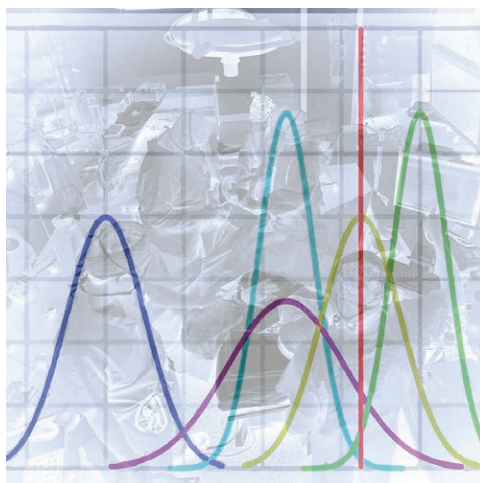
First, Stop Calling Them “Negative”

Daniel I. Sessler, M.D.

Most clinical trials are designed to compare two or more interventions or approaches. Given two or more common approaches to a clinical problem, one is presumably often superior to the other(s). It should then be relatively easy to formally compare various clinical approaches and identify the best. In fact, it has not been easy and many trials report similar outcomes with each tested intervention. Trials that demonstrate that primary results are similar with experimental and reference interventions are often referred to as being “negative”—but that is a suboptimal designation because it encompasses two major causes for results being similar (assuming competent design that limits various sources of bias, missing data, and measurement error).

A common cause of “negative” results is insufficient power, which results because it is statistically challenging to demonstrate substantive changes in relatively uncommon dichotomous events. (It is much easier to demonstrate differences in continuous outcomes such as pain scores, but such outcomes are generally far less important.) Major “hard” outcomes such as death, myocardial infarction, sepsis, reoperation, cardiac arrest, *etc.*, are multifactorial. It is therefore unlikely that any one intervention will reduce the incidence by more than about 20%.

Fortunately, major dichotomous complications are relatively rare with incidences typically ranging from 1 to 10%. But a consequence is that large numbers of patients are required to demonstrate benefit from a given intervention.¹ For example, it takes about 8,600 patients to provide 90% power at a 5% significance level to identify a relative 20% reduction from a baseline incidence of 10% (2% absolute reduction). But as the baseline incidence and treatment effect decrease, the numbers increase exponentially.



“There is nothing ‘negative’ about robust results showing comparable outcomes from various treatments.”

For example, to demonstrate a 10% reduction from a 5% baseline incidence (0.5% absolute reduction) requires 76,000 patients. Such reductions may be clinically important, but are clearly hard to demonstrate.

Trials that claim “no significant difference” among outcomes often lack sufficient power to demonstrate that there actually is no clinically important difference as a function of treatment. Most underpowered trials are small; but trials with many participants and few dichotomous outcomes are also small from a statistical perspective because power is largely derived from the number of events rather than the number of participants. Such results are not truly negative; they are uninformative. Maybe there truly is no important

treatment effect, but maybe there is and it is undetected because the CIs around the effect estimate are large enough to include zero *and clinically important benefit or harm*. In effect, these are failed trials and might better be characterized as “underpowered” rather than “negative” because they do not actually demonstrate lack of treatment effect. Along those lines, even trials reporting statistically significant differences can be underpowered and fragile,² making them functionally similar to underpowered trials with nonsignificant results.

I distinguish here between statistical significance and clinical importance (still assuming that the trials in question are designed properly and well conducted). It is common for trial results to not be statistically significant and have CIs that include differences that are clinically meaningful. For example, in 2005 we published the results of a trial in which about 250 patients were randomly assigned to conservative or aggressive fluid replacement during major surgery.³ The primary result was that the incidence of surgical site infection risk was 11.3% in the low-volume group and

Image: J. P. Rathmell.

Accepted for publication October 14, 2019. Published online first on December 11, 2019. From the Department of Outcomes Research, Anesthesiology Institute, Cleveland Clinic, Cleveland, Ohio.

Copyright © 2020, the American Society of Anesthesiologists, Inc. All Rights Reserved. Anesthesiology 2020; 132:221–4. DOI: 10.1097/ALN.0000000000003046

8.5% in the high-volume group ($P = 0.46$). We concluded that infection risk was similar with each fluid administration strategy, and that “supplemental hydration in the tested range does not have a major impact on infection risk.” In fact, a 25% relative risk reduction would be highly clinically meaningful, but the trial was underpowered to detect that difference. The error in our conclusion was nicely illustrated 15 yr later by the Restrictive *versus* Liberal Fluid Therapy for Major Abdominal Surgery (RELIEF) trial (in which we participated). RELIEF randomized 3,000 patients to conservative or aggressive fluid management, using an approach similar to that of the 2005 trial. The incidence of surgical site infection was 16.5% in the high-volume group and 13.6% in the low-volume group (18% risk reduction, $P = 0.02$).⁴ Clearly, volume restriction does worsen infection risk. The initial trial should thus have been considered underpowered and uninformative, rather than “negative.”

The reverse is also possible: statistically significant results can be clinically meaningless. This situation is rare for trials with dichotomous outcomes both because any difference in “hard” outcomes is usually important, and because it is rare for such trials to be so overpowered. But meaningless statistical significance is common with continuous outcomes. For example, there are hundreds of studies reporting statistically significant differences in pain scores that are clinically meaningless. Consider, for example, a trial of postoperative pain in patients randomized to ondansetron and placebo.⁵ The results showed that ondansetron significantly reduced the analgesic effect of acetaminophen, but correctly concluded that “the reduction was of marginal clinical importance” and that “clinicians can use the combination [of ondansetron and acetaminophen] without anticipating much analgesic impairment.”

My point is that statistical significance needs to be interpreted in the context of what changes are clinically meaningful (preferably defined *a priori*). P values are of limited value in this regard, which is a major reason why they should not be interpreted dichotomously (*e.g.*, results “positive” or “negative”). Instead, the 95% CIs around treatment-effect point estimates should be compared to clinically important differences. When CIs extend across clinically meaningful differences, results are underpowered; in contrast, treatment effect estimates provide useful guidance when CIs are small compared with meaningful differences. To put this another way, trial results should mostly be interpreted based on the relationship between CIs and meaningful differences, rather than dichotomously as significant or not.

Figure 1 illustrates the importance of considering CIs. The hypothetical results from trials of various sizes are all statistically significant, but results from the smallest trial ($n = 500$) have CIs that are wide compared to clinically important differences. These results are therefore statistically significant but almost uninterpretable because the true treatment effect might range from biologically implausible benefit to virtually no effect. In this example, trial size

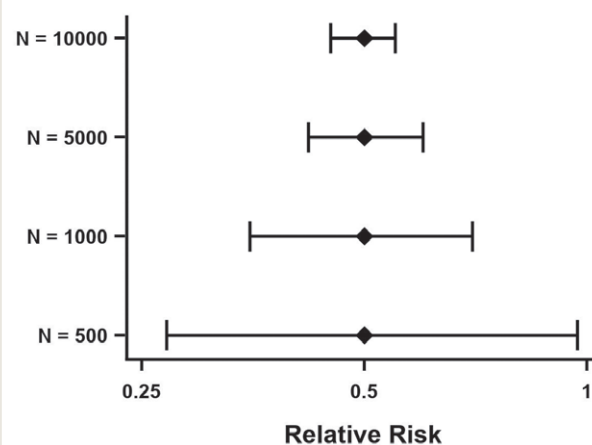


Fig. 1. The point estimate and 95% CIs for a trial with a dichotomous outcome that is reduced from 10 to 5% by an intervention. For perspective, the incidence of major adverse cardiac events is roughly 10% in moderate-to-high-risk patients who have inpatient surgery. All results shown are statistically significant, but their value differs considerably. Consider the results with 500 patients—which is already a substantial trial. The CIs extend from nearly a factor of four reduction in the relative risk, which is biologically implausible for nearly all interventions, to nearly 1, which indicates no benefit. It is necessary to increase trial size to between 1,000 and 5,000 patients to shrink the CIs to a range that provides useful guidance to clinicians considering whether to implement the experimental treatment. This figure illustrates the danger of treating “statistically significant” dichotomously rather than considering the CIs and how useful the results might actually be. The reverse can be true as well: Results that are not statistically significant may nonetheless suggest potentially important differences that should not be equated with “no difference.”

needs to be between 1,000 and 5,000 patients to shrink the CIs to a range that provides useful guidance to clinicians considering whether to implement the experimental treatment. Of course the CIs in any given trial depend on whether the outcome is continuous or dichotomous, baseline variability or incidence, and treatment effect. But whatever the CIs prove to be should be compared to clinically meaningful differences.

I do not mean that underpowered trials have no value. Many pilot trials, for example, are designed for feasibility and safety with no expectation of concluding anything from limited data. Other pilot trials are underpowered by design, but nonetheless provide information that allows investigators to refine designs and more accurately estimate sample size for a future full trial. It is also common for even well-designed trials to end up being underpowered because the baseline incidence or treatment effect turn out to be smaller than anticipated. Such results, while technically “negative” (no significant difference), can still be incorporated into meta-analyses and inform clinicians and future

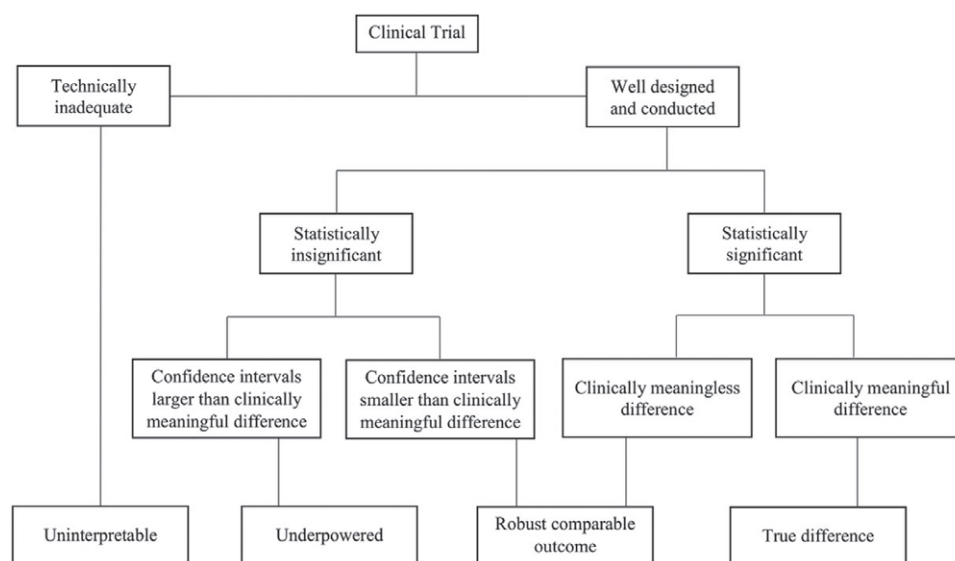


Fig. 2. Schematic dividing clinical trial results into four broad categories based on statistical significance and the relationship between 95% CIs and clinically meaningful differences. These rough categories much oversimplify the complexities of interpreting trial results, but provide a framework for analysis.

investigators, even though they provide little guidance on their own.

An alternative cause of “negative” results is that there truly are no clinically meaningful differences as a function of treatment in a trial large enough to be convincing. Such results provide robust guidance to clinicians, namely indicating that either of two or more choices is acceptable. Often such large trials indicate that an experimental intervention is unhelpful. This is a valuable *positive* contribution because experimental interventions are usually more expensive and often have poorly characterized risk profiles. To put this another way, one of several comparably effective treatments is nearly always preferable by virtue of being easier to implement, less expensive, or safer. There is nothing “negative” about robust results showing comparable outcomes from various treatments. Instead, they guide clinicians to select the overall best treatment.

Noninferiority and equivalence trials are special cases since they are designed to show that two or more treatments do *not* differ by more than defined clinically meaningful amounts. A statistically significant result therefore means that the investigators demonstrated that there was *not* a meaningful difference in treatment—which was their goal. Calling this sort of result “negative” is especially confusing because it could refer to the results being significantly noninferior (the intended outcome) or to the results not being statistically significant—which in turn could result from true differences in treatment effect or from insufficient power. “Negative” is equally uninformative when applied to safety and cost studies where comparable outcomes are

often the desired result. As with conventional efficacy or effectiveness trials, “negative” safety and cost studies can be either underpowered or robust.

I thus distinguish between underpowered and unreliable “negative” results, and robust trials that truly show lack of treatment effect. Currently both are often referred to as “negative trials,” a term which is somewhat pejorative and—more seriously—doesn’t distinguish whether nonsignificant outcomes are underpowered or truly similar. Using specific terms to characterize the reliability of nonsignificant outcomes would improve interpretation of trial results. For example, the former might be referred to as “underpowered.” There does not seem to be a single word that captures the concept of a well-powered trial that demonstrates similar outcomes with various treatments (suggestions welcome!). But a term like “robust comparable outcomes” would be reasonable.

Figure 2 divides the results of conventional superiority trials into four broad categories: (1) *uninformative* (trials that are poorly designed or conducted); (2) *underpowered* (wide CIs relative to clinically meaningful difference and not statistically significant); (3) *robust comparable outcomes* (narrow CIs relative to clinically meaningful difference, whether statistically significant or not); and (4) *true differences* (statistically significant and narrow CIs relative to clinically meaningful difference).

In summary, I encourage investigators and clinicians to abandon the uninformative and pejorative term “negative” when describing trials that report comparable outcomes in each treatment group. Instead, use precise terms such

as “underpowered” to indicate when results are essentially uninformative, and something like “robust comparable outcomes” when treatments are convincingly shown to be truly comparable—which is an important and positive trial outcome.

Competing Interests

The author is not supported by, nor maintains any financial interest in, any commercial activity that may be associated with the topic of this article.

Correspondence

Address correspondence to Dr. Sessler: DS@OR.org

References

1. Mascha EJ, Vetter TR: Significance, errors, power, and sample size: The blocking and tackling of statistics. *Anesth Analg* 2018; 126:691–8
2. 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
3. Kabon B, Akça O, Taguchi A, Nagele A, Jebadurai R, Arkilic CF, Sharma N, Ahluwalia A, Galandiuk S, Fleshman J, Sessler DI, Kurz A: Supplemental intravenous crystalloid administration does not reduce the risk of surgical wound infection. *Anesth Analg* 2005; 101:1546–53
4. Myles PS, Bellomo R, Corcoran T, Forbes A, Peyton P, Story D, Christophi C, Leslie K, McGuinness S, Parke R, Serpell J, Chan MTV, Painter T, McCluskey S, Minto G, Wallace S; Australian and New Zealand College of Anaesthetists Clinical Trials Network and the Australian and New Zealand Intensive Care Society Clinical Trials Group: Restrictive *versus* liberal fluid therapy for major abdominal surgery. *N Engl J Med* 2018; 378:2263–74
5. Koyuncu O, Leung S, You J, Oksar M, Turhanoglu S, Akkurt C, Dolapcioglu K, Sahin H, Sessler DI, Turan A: The effect of ondansetron on analgesic efficacy of acetaminophen after hysterectomy: A randomized double blinded placebo controlled trial. *J Clin Anesth* 2017; 40:78–83