Stopping at Nothing? Some Dilemmas of Data Monitoring in Clinical Trials

Steven N. Goodman, MD, MHS, PhD

This commentary reviews the argument that clinical trials with data monitoring committees that use statistical stopping guidelines should generally not be stopped early for large observed efficacy differences because efficacy estimates may be exaggerated and there is minimal information on treatment harms. Overall, the average of estimates from trials that use these boundaries differs minimally from the true value. Estimates from a given trial that seem implausibly high can be moderated by using Bayesian methods. Data monitoring committees are not ethically required to precisely estimate a large efficacy difference if that difference differs convincingly from zero, and the requirement to detect harms and balance efficacy against harm depends on whether the nature of the harm is known or unknown before the trial.

In this issue, Mueller and colleagues (6) enter this debate squarely in the utilitarian camp, arguing forcefully that the primary purpose of a trial is to get an accurate assessment of the risks and benefits associated with a given treatment. This is a desirable aim, but it is not the goal enshrined in the traditional hypothesis–test framework of study design; rather, that goal is to decide which treatment is more efficacious, with statistical control over how often false-positive and false-negative conclusions are made (7). As we shall see in the ensuing discussion, the goals of error control and accurate estimation can sometimes be in direct conflict.

Are Trials with Stopping Rules Biased?

Much of Mueller and colleagues’ argument rests on claims that trials that stop early for efficacy produce efficacy estimates that are biased, that is, on average higher than the true effect, and they go so far as to declare such results as scientifically invalid. Using bias to judge a stopping rule is akin to moving the goal posts; most stopping rules are not designed to optimize estimation or eliminate bias. However, because accurate assessment of efficacy (and risk) is a worthy scientific goal, it is of interest to see how a trial that exactly follows the dictates of a statistical stopping rule would do on that score.

Mueller and colleagues focus on bias in the effect measure itself, whereas evaluation of estimation bias usually incorporates uncertainty by calculating how often the CI around the estimate includes the true value (8). But because inordinate emphasis is often placed on the observed effect estimate without consideration of the full range of the CI, bias in the effect estimate is of some interest. Claims about bias should be based on all estimates that arise from trials that use stopping guidelines, not just trials that are stopped early. If one considers all such outcomes, bias from trials that use conventional stopping guidelines is small (9–11); such trials do not greatly overstate the effects that they aim to measure. It is therefore reasonable to accept an estimate from such a trial as a valid estimate of effect.
It is also true that the estimates of effect from trials that have stopped early for efficacy tend to be higher than the true value (10–12). How do we reconcile this with the previous claim of minimal bias? The same phenomenon is seen in trials with fixed sample sizes, which cannot be stopped on the basis of an observed efficacy difference and are indubitably unbiased. The key insight is that a trial that has been stopped early for efficacy is by definition statistically significant, usually highly so. If one takes just the significant results in one direction from any set of trial results, their average will necessarily be higher than the average of the whole set. The higher the significance, the larger that difference, and the smaller the sample size of the significant trials, the larger still.

Consider the simple example of a trial with a fixed sample size, designed to detect (with 80% power) a 10% reduction in mortality (from 50% to 40%), and assume sample size, designed to detect (with 80% power) a 10% difference, and the smaller the sample size of the results, their average will necessarily be higher than the average of the whole set. The higher the significance, the larger that difference, and the smaller the sample size of the significant trials, the larger still.

The curves in the Figure assume a known true value for the true effect, showing that the average estimates from trials that use this stopping rule are quite near the truth. The estimate from a trial following a stopping guideline is therefore a pretty good guess if we do not know anything about the true value. But in RCTs, we almost always know something external to the results that may help us judge whether a large observed effect is an overestimate. External evidence includes other research on the treatment in question, findings of related RCTs, evidence supporting the proposed mechanism of effect, studies of other therapies for the same condition, and the design and execution of the trial itself (15). These can be used to construct a distribution of effect sizes that might a priori be considered plausible in a given trial.

Mueller and colleagues implicitly acknowledge the importance of external evidence when they describe the most concerning trials that were stopped early as those with findings that are “implausible” and that require “astute clinicians” to make an appropriate interpretation. For a result to be implausible or surprising, there must be prior evidence that led to a different expectation from what was observed. Conventional approaches to statistical inference do not formally incorporate prior evidence. Bayesian methods do, and they can clarify the issues posed by Mueller and colleagues.

Presume in the previous example that investigators found a 30% improvement in the mortality rate at the first interim look. This exceeds the 23% boundary set for the first look and might be regarded as surprising and perhaps implausible. A Bayesian approach encodes the a priori plausible range of results in the form of a prior probability distribution. Suppose that prior evidence indicated that the

A BAYESIAN APPROACH TO INTERPRETING INTERIM RESULTS

The curves in the Figure assume a known true value for the true effect, showing that the average estimates from trials that use this stopping rule are quite near the truth. The estimate from a trial following a stopping guideline is therefore a pretty good guess if we do not know anything about the true value. But in RCTs, we almost always know something external to the results that may help us judge whether a large observed effect is an overestimate. External evidence includes other research on the treatment in question, findings of related RCTs, evidence supporting the proposed mechanism of effect, studies of other therapies for the same condition, and the design and execution of the trial itself (15). These can be used to construct a distribution of effect sizes that might a priori be considered plausible in a given trial.

Mueller and colleagues implicitly acknowledge the importance of external evidence when they describe the most concerning trials that were stopped early as those with findings that are “implausible” and that require “astute clinicians” to make an appropriate interpretation. For a result to be implausible or surprising, there must be prior evidence that led to a different expectation from what was observed. Conventional approaches to statistical inference do not formally incorporate prior evidence. Bayesian methods do, and they can clarify the issues posed by Mueller and colleagues.

Presume in the previous example that investigators found a 30% improvement in the mortality rate at the first interim look. This exceeds the 23% boundary set for the first look and might be regarded as surprising and perhaps implausible. A Bayesian approach encodes the a priori plausible range of results in the form of a prior probability distribution. Suppose that prior evidence indicated that the
The trials were designed to have 90% power to detect a 10% mortality benefit (for example, 50% vs. 40%). Each panel corresponds to a different underlying true difference: no difference (top), 10% difference (middle), and 20% difference (bottom). The distribution of results is shown for trials of 2 designs: 1 using a 4-look O’Brien–Fleming stopping rule (“stopping”) and 1 using a fixed sample size (“no stopping”). Median effect size and 2.5% and 97.5% percentiles of each estimate are reported in parentheses. The mean sample size is reported for the “stopping” trial only: $n = 1040$ for the fixed sample size design.
Stopping at Nothing? Some Dilemmas of Data Monitoring in Clinical Trials

PERSPECTIVE

10% difference used in the sample size calculation was the expected effect, with the 95% plausible range extending from 7% harm to 25% benefit. A Bayesian estimate that combined this prior distribution with the observed data would yield an estimate of a 23% difference (95% credible interval, 14% to 31%)—substantially less than and almost excluding the observed effect of 30%. This moderated estimate could be reported as the investigators’ best guess of the true difference, much lower than the effect observed but still different from zero. There are also non-Bayesian approaches to adjusting the point estimate, but they are complex, are dependent on the stopping rule, and are not clearly related to subject-specific knowledge (11, 16).

The Bayesian approach formalizes the notion that surprising effects are probably overestimations and provides a more tempered estimate of the true difference (17–20). If the main criticism against estimates from trials stopped very early is that the point estimate is exaggerated, that problem can be ameliorated without extending the trial by using Bayesian methods, which formally incorporate evidence-based skepticism that Mueller and colleagues suggest be used informally. In smaller trials, interim boundary points will necessarily represent large effects, but because they are based on relatively few participants, the Bayesian adjustment will be considerable.

The best correction of an implausible observed effect is achieved by combining a trial result with those of other similar trials. Bayesian adjustments are necessary only if other trials don’t exist or further trials cannot be justified without the adjustment. A Bayesian correction with a prior distribution based on previous RCTs is mathematically equivalent to a standard meta-analysis (21).

STOPPING TRIALS TO SAVE LIVES

The primary statistical goal of standard methods of clinical trial design is to decide which treatment is better with regard to the trial’s primary end point. One goal of data monitoring is to expedite that decision if the data indicate a large difference and of the stopping guideline to do so in a way that will not increase the number of erroneous decisions. If the observed benefit is large, we are willing to estimate it imprecisely (that is, based on fewer patients) if we are confident that the true effect differs from zero. This is where decision and estimation goals can conflict. As one statistician has declared (22):

If reliable estimates are required for each treatment then it seems inevitable that a substantial number of patients must receive the inferior treatment . . . Then it must be recognized that the risks undertaken by volunteers in the experiment are mainly associated with estimation, rather than the need to discover which of the treatments is superior.

When treatments differ substantially in efficacy, monitored trials can dramatically reduce the number of deaths incurred during the study and speed dissemination of the result (23). The Figure shows us the relevant numbers for the example already cited. The potential reduction in sample size is 25% with a 10% true difference and about 55% when the true treatment difference is 20%. It is very difficult to predict how treatment of patients outside the trial will be affected by a DMC decision. Stopping early (with a larger effect) or late (with a larger sample) might have more impact or no impact at all. A utilitarian perspective can justify either decision.

Mueller and colleagues’ position can be recast in a decision-making framework. These authors would like trials to ascertain not merely when relative efficacy is established, but also when relative efficacy is shown to exceed harm. This could be accomplished by defining an end point that combines harm and benefit, testing efficacy and safety separately, or defining a nonzero efficacy threshold that would offset a given degree of harm (24–26). This requires that trial designs be able to measure harm, and it could increase sample size requirements by changing the null hypothesis from a zero effect to a nonzero degree of benefit needed to exceed the harm. If we observe a 20% benefit and want to be sure that this is statistically distinguishable not from zero but from a 10% threshold benefit, the required sample size increases substantially with or without stopping rules.

The implications of Mueller and colleagues’ proposal to explicitly include treatment harms in data monitoring considerations depends critically on whether, before the study, the nature of the harm is known or unknown and whether this harm is likely to appear during the trial. If the harm is unknown and has not appeared once efficacy is established, the DMC should not be expected to support administration of a less effective therapy because its benefit might be offset by a harm that might be discovered if the trial were continued longer. Society would probably not tolerate that. Unsuspected late or rare adverse effects are often better ascertained through observational studies, continued follow-up of trial participants, postmarketing or outcomes research, or meta-analyses (27, 28). The capture and reporting of adverse event data in RCTs in general need to be improved so that risk signals can reliably emerge from meta-analyses (29–31).

When the nature of the harm is known before the study but its frequency is uncertain, as in anticoagulation and stroke or estrogen therapy and breast cancer, the approach to data monitoring can include the harm (24–26, 32, 33) and the plan can be prospectively discussed with institutional review boards, investigators, DMC members, and the patients themselves. In the Women’s Health Initiative, the possibility of asynchronous harm and benefit was built into the stopping guidelines, and extensive preliminary work was done with the DMC to elicit their reactions to different possible observed patterns (34, 35).
In conclusion, Mueller and colleagues cast their argument in stark terms, such as “scientifically invalid,” “biased,” and “unethical.” A more nuanced view is that RCTs and DMCs must balance many competing and worthwhile medical, statistical, ethical, and social goals, which is why the literature in this area is so rich and why DMC deliberations that have been described are so difficult (35-39). The DMCs typically weigh all of the concerns articulated by Mueller and colleagues, and more (36, 40, 41). As with juries, review panels, and other groups empowered with making difficult decisions (42), the outcomes of such deliberations can always be second-guessed, but algorithmic solutions to improve the process rarely do so. However, as implied by Mueller and colleagues’ comments, not all DMCs have a sophisticated understanding of methodological issues and not all function optimally. The number of potential DMC participants with training and experience in the DMC process is relatively small, and efforts to expand that pool are badly needed (43).

This discussion should be viewed as part of a broader debate about the acceptable speed of medical progress. This pace is a conscious social choice that implicitly balances the interests of individuals against that of the broader population (4). If society perceives individual interests to have been excessively compromised for the collective good, investigators risk a social response that can seriously harm the interests of individuals against that of the broader population (43). If society perceives individual interests to have been excessively compromised for the collective good, investigators risk a social response that can seriously harm both the scientific enterprise and, paradoxically, that collective good (44). How heavy the hand should be on each side of the individual versus collective ethical scale feels personal to DMC members, but it is ultimately a societal choice to be determined through public discussion. Mueller and colleagues’ perspective contributes to that discussion and will stimulate yet more conversation among scientists and the public on this critical issue.

From Johns Hopkins University Schools of Medicine and Public Health, Baltimore, Maryland.

Disclaimer: The views and content herein are solely the responsibility of the author.

Acknowledgments: The author thanks Drs. Donald Berry, Thomas Louis, and Joel Greenhouse for their helpful comments on earlier versions of the manuscript.

Potential Financial Conflicts of Interest: None disclosed.

Requests for Single Reprints: Steven N. Goodman, MD, MHS, PhD, Johns Hopkins University Schools of Medicine and Public Health, 550 North Broadway, Suite 1103, Baltimore, MD 21205; e-mail, sg Goodman @jhmi.edu.

References


READERS’ COMMENTS

Readers wishing to comment on published articles should use the “Send comment/rapid response letter” option at www.annals.org. While this service is free to Annals subscribers, readers without subscriptions who wish to comment on articles may purchase temporary access.