JAMA Guide to Statistics and Methods

Interpretation of Clinical Trials That Stopped Early

Kert Viele, PhD; Anna McGlothlin, PhD; Kristine Broglio, MS

Clinical trials require significantresources to complete in terms of patients, investigators, and time and should be carefully designed and conducted so that they use the minimum amount of resources necessary to answer the motivating clinical question. The size of a clinical trial is typically based on the minimum number of patients required to have high probability of detecting the anticipated treatment effect. However, it is possible that strong evidence could emerge earlier in the trial either in favor of or against the benefit of the novel treatment. If early trial results are compelling, stopping the trial before the maximum planned sample size is reached presents ethical advantages for patients inside and outside the trial and can save resources that can be redirected to other clinical questions. This advantage must be balanced against the potential for overestimation of the treatment effect and other limitations of smaller trials (eg, limited safety data, less information about treatment effects in subgroups).

Many methods have been proposed to allow formal incorporation of early stopping into a clinical trial.[1,2] All of these methods allow a trial to stop at a prespecified interim analysis while maintaining good statistical properties. Data monitoring committees or other similar governing bodies may also monitor the progress of a trial and recommend stopping the trial early in the absence of a prespecified formal rule. An overwhelmingly positive treatment effect might lead to a recommendation for unplanned early stopping but, more commonly, unplanned early stopping results from concerns for participant safety, lack of observed benefit, or concerns about the feasibility of continuing the trial due to slow patient accrual or new external information. Trials stopped for success in an ad hoc manner are challenging to interpret rigorously. In this article, we focus on early stopping for success or futility based on formal, prespecified stopping rules.

In the December 15, 2015, issue of *JAMA*, Stupp et al[3] reported the results of a trial assessing electric tumor-treating fields plus temozolomide vs temozolomide alone in patients with glioblastoma. The trial design included a preplanned interim analysis defined according to an early stopping procedure. The trial was stopped for success at the interim analysis, reporting a hazard ratio of 0.62 for the primary end point of progression-free survival.

Use of the Method

Why Is Early Stopping Used?

When 2 treatments are compared in a randomized clinical trial, the treatment effects observed both during the trial and when the trial ends are subject to random highs and lows that depart from the true treatment effect. Sample sizes for trials are selected to reliably detect an anticipated treatment effect even if a modest, random low observed treatment effect occurs at the final analysis. If such a random low value does not occur or the true treatment effect is larger than anticipated, the extra study participants required to provide this protection against a false-negative result may not be necessary. During the course of a trial, strong evidence may accumulate that the experimental treatment offers a benefit. This may be from a large observed treatment effect emerging early in a trial or from the anticipated treatment effect being observed as early as two-thirds of the way through a trial.

Conversely, evidence could accumulate early in a trial that the experimental treatment performs no better than the control. In a trial with no provision for early stopping, patients would continue to be exposed to the potential harms of the experimental therapy with no hope of benefit. Interim analyses to stop trials early for futility may avoid this risk. Trials may also stop early for futility if there is a limited likelihood of eventual success.[4]

What Are the Limitations of Early Stopping?

One key statistical issue with early stopping, particularly early stopping for success, is accounting for multiple "looks" at the data. Accumulating data, particularly early in the trial with a smaller number of observations, is likely to exhibit larger random highs and lows of values for the treatment effects. The more frequently the data are analyzed as they accumulate, the greater the chance of observing one of these fluctuations. Rules allowing early stopping therefore require a higher level of evidence, such as a lower *P* value, at each interim analysis than would be required at the end of a trial with no potential for early stopping. Taken together, the multiple looks at the data, each requiring a higher bar for success, lead to the same overall chance of falsely declaring success (type I error) as a trial with the usual criterion for success (eg, a $P<.05$) and no potential for early stopping.

Early stopping for futility requires no such adjustment. There are no added opportunities to declare a success; thus, no statistical adjustment to the success threshold is required. However, futility stopping may reduce the power of the trial by stopping trials based on a random low value for the treatment effect that could have gone on to be successful. This reduction in power is usually quite small.

Success thresholds are typically chosen to be more conservative for interim analyses than for the final analysis should the trial continue to completion. The O'Brien-Fleming method, for example, requires very small *P* values to declare success early in the trial and then maintains a final *P* value very close to the traditional .05 level at the final analysis.[1] Using this method, very few trials could be successful at the interim analyses that would not have been successful at the final analysis. Thus, there is a minimal "penalty" for the interim analyses. The more conservative the early stopping criteria, the more assurance there is that an early stop for success is not a false-positive result.

While methods such as O'Brien-Fleming protect against falsely declaring an ineffective drug successful, the accuracy of estimates of the treatment effect in trials that have stopped early for success remains a concern.[5] When considering the true effect of a treatment, bias is introduced when considering only trials that have observed a large enough treatment effect to meet the critical value for success. By definition, successful trials have larger treatment effects than unsuccessful trials; thus, successful trials include more random highs than random lows. As such, small trials that end in success, either at the end or early, are prone to overestimating the treatment effect. The larger the observed treatment effect, the more likely it is an extreme random high, and the greater the chance for overestimation. If the trial were continued, with the enrollment of additional patients, it is likely that there would be a reduction of the observed treatment effect. In other words, trials with very impressive early results are likely to become less impressive after observing more data, and this should be taken into account when monitoring and interpreting such trials. Extreme attenuation, such as a complete disappearance of the observed treatment benefit, however, is less likely.

Why Did the Authors Use Early Stopping in This Study?

Glioblastoma is an aggressive cancer with few treatment options. In the report by Stupp et al,[3] enrollment was largely complete at the time of the interim analysis. However, the interim analysis allowed the possibility that a beneficial result could be disseminated many months (potentially years) earlier in advance of the fully mature data.

How Should Early Stopping Be Interpreted in This Particular Study?

The primary analysis in this study found a hazard ratio of 0.62 (*P* = .001) based on 18 months of follow-up from the first 315 patients enrolled. This is strong evidence of a treatment benefit for tumor-treating fields plus temozolomide in this population. However, care should be taken when interpreting the estimated benefit corresponding to a hazard ratio of 0.62. Given the potential for an overestimated treatment effect, combined with the general intractability of treating glioblastoma, there is good reason to suspect that the actual benefit of tumor-treating fields, while present, might be smaller than that observed in the study. A robustness analysis (ie, a supplementary or supporting analysis conducted to see how consistent the results are if different approaches were taken in conducting the analysis), based on the then-available data from all participants, illustrates this pattern. That analysis resulted in a hazard ratio of 0.69 (95% CI, 0.55-0.86), also with a *P*<.001. The result remained statistically significant, but the magnitude of the treatment effect was smaller.

Caveats to Consider When Looking at a Trial That Stopped Early

It is important to consider trial design, quality of trial conduct, safety and secondary end points, and other supplementary data when interpreting the results of any clinical trial. For trials that stop early for success, the statistical superiority of an experimental treatment is straightforward when the early stopping was preplanned and it is reasonable to preserve patient resources and time once the primary objective of a trial has been addressed. Early stopping procedures protect against a false conclusion of superiority. However, if the result seems implausibly good, there is a high likelihood that the true effect is smaller than the observed effect. In that light, the benefits of early stopping, to patients both in and out of the trial, must be weighed against how much potential additional knowledge would be gained if the trial were continued.

ARTICLE INFORMATION

Author Affiliations: Berry Consultants LLC, Austin, Texas.

Corresponding Author: Kert Viele, PhD, Berry Consultants LLC, 4301 Westbank Dr, Bldg B, Ste 140, Austin, TX 78746 (kert@berryconsultants.com).

Section Editors: Roger J. Lewis, MD, PhD, Department of Emergency Medicine, Harbor-UCLA Medical Center and David Geffen School of Medicine at UCLA; and Edward H. Livingston, MD, Deputy Editor, *JAMA*.

Conflict of Interest Disclosures: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

REFERENCES

1. Jennison C, Turnbull BW. *Group Sequential Methods With Applications to Clinical Trials*. Boca Raton, FL: Chapman & Hall; 2000.

2. Broglio KR, Connor JT, Berry SM. Not too big, not too small: a Goldilocks approach to sample size selection. *J Biopharm Stat*. 2014;24(3):685-705.

3. Stupp R, Taillibert S, Kanner AA, et al. Maintenance therapy with tumor-treating fields plus temozolomide vs temozolomide alone for glioblastoma: a randomized clinical trial. *JAMA*. 2015;314(23):2535-2543.

4. Saville BR, Connor JT, Ayers GD, Alvarez J. The utility of Bayesian predictive probabilities for interim monitoring of clinical trials. *Clin Trials*. 2014;11(4):485-493.

5. Zhang JJ, Blumenthal GM, He K, Tang S, Cortazar P, Sridhara R. Overestimation of the effect size in group sequential trials. *Clin Cancer Res*. 2012;18(18):4872-4876.