Published in final edited form as:

Clin Trials. 2016 December ; 13(6): 621–631. doi:10.1177/1740774516649595.

Quantifying over-estimation in early-stopped clinical trials and the “freezing effect” on subsequent research

Hao Wang¹, Gary L Rosner¹, and Steven N Goodman^{2,3}

¹Division of Biostatistics and Bioinformatics, Johns Hopkins University School of Medicine, Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins, Baltimore, MD, USA

²Stanford University School of Medicine, Stanford, CA, USA

³Meta-research Innovation Center at Stanford (METRICS)

Abstract

Background—Despite the wide use of the design with statistical stopping guidelines to stop a randomized clinical trial (RCT) early for efficacy, there are unsettled debates of potential harmful consequences of such designs. These concerns include the possible over-estimation of treatment effects in early-stopped trials, and a newer argument of a “freezing effect” that will halt future RCTs on the same comparison since an early-stopped trial represents an effective declaration that randomization to the un-favored arm is unethical. The purpose of this study is to determine the degree of bias in designs that allow for early stopping, and assess the impact on estimation if indeed future experimentation is “frozen” by an early-stopped trial.

Methods—We perform simulations to study the effect of early stopping. We simulate a collection of trials and contrast the treatment-effect estimates (risk differences and ratios) with the simulation truth. Simulations consider various scenarios of between-study variation, including an empirically derived distribution of effects from the clinical literature.

Results—Across the trials whose true effects are sampled from a uniform distribution, estimates from trials that stop early for efficacy deviate minimally from the simulation truth (median bias of the estimate of risk difference is 0.005). Over-estimation becomes appreciable only when the true effect is close to the null value 0 (median bias of the risk difference estimate is 0.04) or when stopping happens with 40% information or less; however, stopping under these situations is rare. We also find slight reverse bias of the estimated treatment effect (median bias of the risk difference estimate is −0.002) among trials that do not cross the early stopping boundaries but continue to the final analysis. Similar results occur with relative risk estimates. In contrast, Bayesian estimation of the treatment effect shrinks the estimate from trials stopping early and pulls back under-estimation from completed trials, largely rectifying any over-estimation among trials that terminate early. Regarding the so-called “freezing effect,” the pooled effects from meta-analyses that include truncated RCTs show an unimportant deviation from the true value, even when no subsequent trials are conducted after a truncated RCT.

Conclusions—Group sequential designs with stopping rules seek to minimize exposure of patients to a disfavored therapy and speed dissemination of results, and such design do not lead to materially biased estimates. The likelihood and magnitude of a “freezing effect” is minimal. Superiority demonstrated in a RCT stopping early and designed with appropriate statistical stopping rules is likely a valid inference, even if the estimate may be slightly inflated.

Keywords

Clinical trial methodology; early stopping; treatment effect; over-estimate; bias; Bayesian inference; meta-analysis; “freezing effect”

Introduction

Randomized clinical trials often include interim analyses with statistical stopping rules to evaluate emerging evidence of treatment efficacy. Group sequential designs provide a hypothesis-testing framework for making decisions regarding early termination of the study. These methods address pragmatic, ethical, and statistical concerns. These concerns include not expending resources on an experiment beyond the point at which the evidence is compelling that one therapy is superior to another, minimizing the number of patients who will receive a demonstrably inferior therapy, and releasing clinically relevant information from the trial to patients and providers outside the trial as soon as possible, all while controlling the type I error rate inflation caused by multiple looks at the data.

Randomized controlled trials (RCTs) that stop early for benefit may yield estimates of treatment effects that are high. It has been claimed that this is a substantial bias that must be taken into account in meta-analyses and interpretation of trials that stop early. Through a systematic review and meta-analysis, Bassler et al.¹ reported that the pooled ratio of relative risk (RR) of the experimental therapy compared to control therapy in the truncated trials that stopped early versus matching non-truncated trials that either did not stop early or did not have planned interim looks was 0.71 (95% confidence interval: 0.65–0.77), a 29% exaggeration of effect. The authors concluded that “their results represent a conservative estimate of the exaggeration in treatment benefit associated with stopping early.” The same authors also suggested that early-stopped trials (dubbed tRCTs for “truncated” clinical trials) are ethically problematic because they produce seriously inflated estimates of the treatment effect, that early stopping should be included in “risk of bias” instruments, and that early stopping biases meta-analyses as well.^{2,3}

Subsequent commentaries and simulations^{4–6} and empirical research⁷ showed that the investigators had erred by comparing point estimates of statistically significant findings to those of non-significant findings and declaring the difference to be a “bias.” Therefore, most of what was labeled “bias” was, in fact, random variation, and the true “bias” of estimates from trials stopped early was shown to be quite small, with minimal impact on meta-analyses.^{6,8} This was acknowledged in a 2013 paper⁹ by the original authors of the Bassler paper,¹ in which they withdrew the bias claim.

Unfortunately, belief in the original claim of substantively biased estimates from early-stopped trials persists, partly because it appeared in a high impact general medical journal

(JAMA), was preceded by another JAMA article by the same group suggesting the same,¹⁰ and was reiterated by the authors in other articles through 2013.^{11,12} Further, it had both face validity and some element of truth, in that such estimates are subject to over-shoot. At the time of this writing, the papers suggesting a problem have been cited over 990 times, versus 12 citations of the 2013 corrective article.

In several publications, these authors have also raised concern about a phenomenon they dubbed the “freezing effect”.^{1,9} This is a concern that future RCTs making the same comparison will be halted (“frozen”) after an early-stopped trial, because stopping early represents an effective declaration that randomization to the unfavored arm is unethical. Such a concern is plausible, as early-stopped trials often attract attention among the media and editorialists, in part because of the inherent drama implied by a data monitoring committee declaring that the estimate of benefit (or harm) from a particular therapy was so compelling that a trial could not continue. Indeed, there are anecdotal reports of sponsors taking advantage of this phenomenon by intentionally planning trials with sample sizes larger than needed in the hope that an “early stop” will garner headlines.

In this paper, we conduct simulations under a wide variety of scenarios to explore further the degree of over-estimation of treatment effect estimates from trial designs with interimmonitoring boundaries that allow for early stopping for efficacy. We provide quantitative insight into conditions under which such over-estimation is likely to be of greatest concern, as well as how frequently these circumstances are likely to arise. We use the empirical distribution of treatment effects reported by Bassler et al.¹ to re-examine the impact of early stopping for efficacy. Finally, we conduct the first simulation study of the quantitative impact of the postulated “freezing effect” of early-stopped trials on subsequent research. We end with a discussion of how to interpret interim analysis results from a particular trial that stopped early for benefit and present an approach for correcting the treatment effect estimate.

Methods

We study the effect of early stopping via simulations. We consider two-arm randomized clinical trials with a binary primary outcome. The simulated trials incorporate interim analyses with early stopping rules for efficacy based on O’Brien-Fleming boundaries¹³ having a one-sided type I error probability of 0.025. We examine a design with four equally spaced interim analyses and a final analysis. The stopping boundaries, based on observed Z scores, are 4.88 (look 1), 3.36 (look 2), 2.68 (look 3), 2.29 (look 4) and 2.03 (final analysis.) (R package GroupSeq is used to compute the boundaries.)

The simulations follow a two-stage random-effects model. We generate each trial-specific treatment effect at random from a specified distribution representing a hypothetical collection of clinical trials.¹⁴ The observed outcomes of that trial, designed with five looks, are generated randomly based on this effect. We then contrast the trial’s estimate with its true effect and average over the distribution of possible trial effects.

We repeat these same simulations assuming three hypothetical distributions for the true treatment effect: a uniform distribution; several normal distributions centered on the null value of no treatment difference, with several between-study treatment-effect variances from large to small; and a non-null distribution based on a range of effects estimated from the literature.

1. *“Anything is possible”*: The simulation assumes a 40% event probability in the control group. The event probability in the treatment group corresponds to a random risk difference (RD) drawn from a uniform $(-0.4, 0.6)$ distribution. This scenario represents a population of treatment contrasts that are equally likely to be very large, modest, or small. (Figure 1a)
2. *Varying effects, biased toward skepticism*: The risk difference is sampled from a truncated mean zero normal distribution with differing standard deviations, viz., 0.05, 0.1, 0.25, 0.5, and 1. (Figure 1b) Large standard deviations represent large between-study variation, and small standard deviations represent small variation between studies.
3. *Between-study variation based on an empirically observed distribution of effects from the clinical literature*: The Bassler study¹ showed a mean relative risk (RR) = 0.82 in non-truncated trials and approximately 95% of estimates falling between 0.56 and 1.2. We model this as a lognormal distribution for RRs with mean $\ln(0.82)$ and variance 0.19². (Figure 1c)

For scenarios (1) and (2), we consider a maximum sample size of 518 patients per group based on traditional sample size formulae, to achieve 90% power to detect an increase in event probability from 40% to 50% with a one-sided Z test at significance level 0.025. For scenario (3), we simulate up to 589 patients per group in each trial to allow 80% power to detect a relative risk of 0.8, assuming that the risk in the control group is 0.4. Positive differences or relative risks smaller than 1 indicate superiority of the experimental treatment.

We obtain the distribution of the estimates at each of the 5 analysis times. We estimate RDs or RRs from simulated trials for two groups of the trials: those in which the stopping boundaries are crossed at any of the four interim analyses, and those that do not stop early. We compare the simulated trial's estimated RD or RR to the trial's simulated true effect, and compute the difference or ratio. Bias of the design is computed by averaging over all simulated trials including early stopped trials and those that do not cross the early stopping boundaries and continue to the fifth analysis. We compute “over-” or “under-” estimation for trials that stop early or stop at a particular interim analysis as the average of the corresponding subset of results. We also carry out Bayesian analyses, using various prior distributions centered at the simulation's overall effect distribution used in the simulation and centered at a different value. We calculate a Bayesian posterior mean estimate of $\ln(\text{RR})$, computed as the weighted mean of the observed $\ln(\text{RR})$ values and the prior mean, with the weight being their relative precisions (inverse of the variance).¹⁵

Assessing the “freezing effect”

Finally, we use similar procedures to estimate the impact of a “freezing effect.” We focus on the meta-analyses involving early-stopped trials about which the 2013 corrective article of Bassler et al. raised a concern.⁹ We assume there could be up to 10 randomized trials for a particular experimental therapy. If any trial stops early for efficacy, no additional trials in the series would occur. The meta-analytic summary estimates are based on the Mantel-Haenszel fixed-effect model. The event risk is fixed at 0.6 in the control group. To assess the degree of relative bias in the pooled estimate, we take the ratio of the RR from the meta-analysis to the true RR. The simulation is repeated 10,000 times. The first set of simulations are with a fixed RR value, and the second simulation set represents a more empirically-supported scenario in which the RR of each trial could either be null (RR=1) or the effect quantified by Bassler et al.¹ We model this by sampling each RR from a mixture distribution that gives 50% probability to the null hypothesis and 50% probability to the previously described empirical distribution. Mathematically, $\ln(\text{RR}) \sim 0.5 \cdot I_0 + 0.5 \cdot \mathcal{N}(\ln(0.82), 0.19^2)$, where I_0 indicates the null value 0 (corresponding to RR=1), and $\mathcal{N}(\ln(0.82), 0.19^2)$ is the empirical distribution of RRs from the Bassler et al. study.¹

Defining “bias”

We reserve the term “bias” for the unconditional difference between the average of all trial estimates and the true parameter being estimated over all possible trial outcomes.¹⁶ In this paper, we define $\text{Bias} = E \left(E \left(\hat{\theta}^* - \theta | \theta \right) \right)$, where θ represents the risk difference and $\hat{\theta}^*$ is its estimate at each trial’s last analysis, which is either at the interim look when the stopping boundary is crossed or at the planned final analysis if the trial continues to completion, θ follows the distributions described in scenarios 1–3. The ratio of the relative risk estimate to the truth is $E \left(E \left(\hat{\theta}^* / \theta | \theta \right) \right)$ when θ is the true relative risk. We use the terms “over-estimate” or “underestimate” for pre-specified subsets of those results (e.g., at a particular interim look) if the average deviates from the true parameter value. Some statisticians have used the term “conditional bias” to describe such errors, but we avoid that phrase because it is not widely accepted. Also, the term invites confusion of terminology and concept, in that many outcome subsets of unconditionally unbiased procedures can be shown to have “conditional bias.” Furthermore, as we demonstrate, the unconditional bias of a procedure is fairly robust to the distribution of true effects, but over- and under-estimation within subsets of results (e.g. by interim analysis) can vary widely depending on variation in such distributions.

Results

Over-estimation in early-stopped trials

Across the trials with effects sampled from the uniform distribution, estimates from trials that stop early for efficacy have minimal deviation from the underlying effect. Median over-estimation (estimated risk difference minus truth) is 0.005 (10th and 90th percentiles: −0.059, 0.074) (Table 1) in this case. For example, if the true value is 30%, the average of the estimates is 30.5%. We divide each trial-specific bias by the true value and find the median relative bias to be 1.5% (10th and 90th percentiles: −17.8%, 28.4%) (Table 1). Estimated and

true treatment effects fall randomly around the line of equality, indicating no tendency to over- or underestimate the true effect as that effect varies. Figure 2 shows the magnitude of over-estimation in simulated trials for each interim analysis where the efficacy stopping boundary is crossed and for the trials continuing to the final analysis. Over-estimation is largest among trials that stop at the first interim analysis (i.e., after 20% information), which occurs around 1 time in 4 (25.4%). Median bias at each of the four interim analyses that lead to early termination is 0.014, -0.001, -0.002, and -0.006 (corresponding to relative over-estimation of 2.7%, -0.6%, -1.4%, and -6.5%), respectively, when the trials terminate early for benefit. The chance of stopping early is 25.4%, 16.4%, 5.1%, and 3.1%, respectively. Median over-estimation among all simulated trials (truncated or not) is a negligible 0.0004 (10th and 90th percentiles: -0.048, 0.056).

When we assume the true RD follows a Normal (0, 1) distribution (Table 1), we find slight over-estimation in trials that stop early. Over-estimation appears appreciable when the underlying treatment effects are small and when the between-study treatment effects' standard deviations (SDs) are little. This phenomenon is to be expected, because if the true difference lies in a range below the stopping boundary, then by definition any estimate based on hitting the boundary will be an over-estimate. If the RDs follow a Normal(0, 0.1²) sampling distribution, with 95% of true differences falling between -0.2 and 0.2, the median degree of over-estimation is as high as 0.02 for trials stopping early. Only 24% of trials under this distribution hit the stopping boundary at an interim look, however.

For the scenario in which the distribution of trial-specific treatment effects is centered away from zero, we consider a log-normal sampling distribution for $\ln(RR)$, with mean = $\ln(0.82)$ and SD = 0.19. This empirically-based distribution of treatment effects comes from Bassler et al.,¹ based on estimates from the literature. This right-skewed distribution (Figure 1c) implies that trials have a median risk ratio of 0.82, with 14% of trials having RR estimates above 1, and a 41% chance that a trial's RR lies between 0.8 and 1. With this distribution, the median ratio of RR estimates to the true RRs is 0.95 for early-stopped trials and 1.02 for trials not stopping early. Although the estimate of RR at the first look substantially over-estimates the underlying effect, stopping at the first interim analysis is rare (<0.1%, Table 2). Inflation of effect estimates is moderate in the trials that stop at the second interim look (40% information), and gets even smaller if trials stop at later interim analyses (i.e., after 60% or more of planned information accrued) (Table 2). In other words, we expect the greatest degree of over-estimation in situations with the lowest probability of occurring, which is why the bias over all trial estimates is small. Trials that continue to the final analysis result in a slight under-estimate of the treatment effect (Table 2). The magnitude of inflation depends on the control group's risk (Table 2). When the risk of an event among controls is 0.2, estimates from early stopped trials exhibit greater deviation.

Bias with Haybittle-Peto boundaries is of a similar magnitude as the popular O'Brien-Fleming boundaries. Under scenario 1, when true effects are sampled from the uniform distribution, mean and median over-estimation among the early stopped trials are 0.008 and 0.007 (corresponding to relative over-estimate of 6.5% and 2.2%), respectively, with 47.8% trials stopping early at interim analysis. At each of the four interim analyses that lead to early termination for benefit, median biases in risk difference estimates are 0.001, -0.002,

0.002, 0.0004 (corresponding to relative over-estimate of 2.7%, -0.7%, 1.1%, 0.3%), respectively. Fractions of trials stopping early at each of the four interim analyses is 38%, 5.6%, 2.5%, 1.7%, respectively. The median over-estimate among all the simulated trials (truncated and completed trials) is 0.001 (10th and 90th percentiles: -0.050, 0.059). For scenario 3 (effect distribution based on Bassler et al.¹) at each of the four interim analyses, ratios of relative risk estimates to the true relative risks are 0.76, 0.91, 0.95, 0.98 (chance of early stopping is 6.9%, 12.9%, 10.6%, 7.7%), respectively, when 40% event risk is assumed for the control group.

Adjusting the estimate of individual trial

Among published methods for interpreting interim results,^{4,14,17-22} the Bayesian approach takes advantage of known external evidence of the treatment effect and summarizes it in the prior distribution.^{4,21,22} The degree of correction depends on the information characterized via the prior distribution. With a prior distribution for the RR that matches the sampling distribution in the simulations presented in the previous section (i.e., $\ln(\text{RR}) \sim N[\ln(0.82), 0.19^2]$), the correction is perfect. The ratio of the Bayesian posterior RR estimate to the underlying true RR among trials stopping early has median 1.00 (10th and 90th percentiles: 0.89 and 1.14). Among trials that continue to the final analysis, the median ratio of the Bayesian estimate to the underlying truth is 1.00 (10th and 90th percentiles: 0.92 and 1.10, respectively). Bayesian posterior estimation shrinks the over-estimate of early stopped trials and pulls back the under-estimate in trials not stopping early (Figure 3).

Table 3 presents Bayesian estimates of treatment effects when the prior distributions differ from the sampling distribution used in the simulations. A weaker prior with a mean of $\ln(0.82)$ and standard error of 0.25 reflects the opinion of a higher chance (46%) of large treatment effects (i.e., the RR is smaller than 0.8), and a smaller percentage (33%) having RR between 0.8 and 1. With this prior, posterior estimation places relatively larger weight on the data, leading to less shrinkage. With this second prior, the median RR ratios are 0.98 and 1.01 for early stopped trials and not-truncated trials, respectively, values not of substantive importance.

We also consider analyses based on a pessimistic prior, biased towards the null, with a mean of the log normal distribution set to zero (i.e., $\ln(\text{RR})=1$). One may consider such a prior to be appropriate when there is no evidence to indicate that the treatment is beneficial. For example, a $N(\ln(1), 0.35^2)$ prior considers that the RR in 95% of trials would be between 0.5 and 2.0. The prior with mean $\ln(1)$ typically leads to substantial shrinkage to the null value. This shrinkage largely rectifies over-estimation for trials stopping early, even though this prior differs from the true distribution used to generate treatment differences. For those trials that do not stop early and in which the true treatment effect is moderate, this prior does not correct the existing underestimate, because the shrinkage of the treatment effect is in the same direction as the null.

Impact of the “freezing effect”

Our simulation studies suggest tRCTs do not lead to a materially inflated effect estimate as long as all trials up to the point when experimentation stops are pooled for a meta-analysis.

When the investigational treatment has a large effect, $RR = 0.6\text{--}0.75$, even with a small sample size, an average of two trials (10th and 90th percentile ranges 1–6) would be conducted before future investigations are halted (“frozen”), and the inflation of the estimate of treatment effect is only 4.3% (Table 4). When the effect is modest ($RR = 0.8\text{--}0.9$), the meta-analysis produces a 3.7% over-estimate. As the true effect gets close to the null, meta-analyses over-estimate the RR by less than 5%. In the simulations with a more realistic scenario where the effect is either sampled from the empirical results of Bassler et al.¹ ($\ln(RR) \sim N(\ln(0.82), 0.19^2)$) or has no effect ($RR = 1$), the upwards bias of the meta-analysis is as small as 1.9% when the number of patients in the trial is large and 4.7% when sample size is small (Table 5).

Discussion

We have presented a comprehensive simulation study evaluating the degree of over-estimation of treatment effects in tRCTs and the associated “freezing effect.” The deviation of efficacy estimates from randomized-controlled trials with designs that allow for stop early for benefit is minimal, compared to the true effect, when we consider all possible effect sizes that can occur in a trial. This suggests that estimates from group sequential designs with stopping rules are not meaningfully “biased.” If we have no clue about the likely treatment effect, it is perfectly reasonable to accept the results from early stopped trials.

Researchers are concerned with different kinds of bias. There are the biases that arise because of a study’s design or conduct. Some examples of design bias include selection bias, ascertainment bias, and outcome assessment bias resulting from, say, a lack of masking the outcome assessor to the patient’s treatment. It is often very difficult to remove or even reduce design bias after the study is over. If one wants to assess the bias of a procedure or design, one needs to consider all possible outcomes of the procedure before implementing the procedure. If there is no bias from a clinical trial design, then there is no reason to label any outcomes from that design as low, high, or “biased” without information about the true value. Since the bias following a group sequential design is not design bias, there is no reason to call such a design “biased.”

On the other hand, from a purely statistical standpoint, conventional estimates of treatment effects from trials stopping early can be slightly high under certain conditions. One may be able to apply mathematical corrections to mitigate estimation bias once the data are in hand.^{4,14,17–22} Although it is well recognized that the usual maximum likelihood estimate following a sequential trial that stopped at an interim analysis for beneficial effect may be inflated, a recent study suggested that the bias on estimation assessed relative to a comparable set of positive fixed sample-size trials is negligible if stopping occurs after 50% information has been observed.¹⁶ Our simulations indicated that over-estimation of the effect becomes appreciable only when the true difference is close to the null hypothesis, or when early stopping occurs at an interim analysis with 40% of planned information or less. Trials that stopped with insufficient information accounted for less than 10% of all the simulated trials. We also found the observed treatment effect to be slightly *underestimated* among trials that do not cross early stopping boundaries and continue to the final analysis.

Whether a trial is stopped early or not, a Bayesian analysis may be useful to adjust the treatment effect. If a large effect observed from an early stopped trial seems implausible, there must be previous evidence suggesting why it is implausible. A prior distribution of the effect size can be constructed using this external information. The Bayesian approach combines the prior evidence and the current trial data, providing a moderated estimate of the treatment difference.

The primary goal of stopping rules is to ensure early release of trial results when warranted and reduce the number of patients receiving inferior therapy, not to estimate the treatment effect accurately. We must consider whether such over-estimation would lead to different treatment policies or change clinical practice. First, concerns about a “freezing” effect impeding the conduct of future trials as a consequence of tRCTs and thereby leading to over-estimation are not supported by our simulation results. We do not observe serious over-estimation of treatment effects in meta-analyses using pooled trials up to the first tRCT. Unless the true treatment difference is large, it is very unlikely that an RCT will stop at very early interim analyses, when the estimate may be appreciably optimistic. This is borne out by empirical observation. Among the 27 National Cancer Institute Cooperative Group trials from 1990 to 2005 that stopped early or whose results were reported early with positive results, only one trial stopped before at least 25% of information was observed; half of these trials crossed the early stopping boundary after more than 60% of the planned information accrued (information range: 15% to 90%).⁷ The Guidance for Industry issued by US Food and Drug Administration stresses the need for independent substantiation, recommending more than one clinical trial to establish a conclusion of effectiveness.²³ Correction of implausibly large effects would be achieved by additional studies, as shown by our simulations.

Second, even though early stopped trials may sometimes produce slightly inflated estimates of treatment effects, the chance for the effect to be null in those situations is very small. Korn et al. found that 24 out of 27 trials (89%) that crossed the early stopping boundaries for efficacy retained statistical significance based on the final analysis.⁷ Moreover, among 18 trials that had more than 80% follow-up information at the last analysis, only one trial (<6%) reached a different conclusion at the final analysis than the inference made at the time the trial crossed the early stopping boundary. The empirical evidence shows that concerns about excess false positives resulting from designs allowing early stopping are unwarranted if the trials are properly designed to control type I errors. Since the superiority of the new treatment demonstrated by the trials that stop early for efficacy is likely a true positive finding, it is unethical to conduct or continue trials just to slightly change point estimates of efficacy.

Finally, the likelihood and magnitude of a “freezing effect” is minimal. If there are external reasons to believe that a true effect is small or that a given trial was conducted in a way that might have amplified that effect, then one should apply corrections or conduct other trials. If the first trial was very large or seemingly definitive, running another trial may be difficult, but the earliest trials in a series usually are not the largest. In general, these results justify the rule requiring a minimum of two trials should be done before concluding efficacy.

In summary, group sequential designs with stopping rules are not materially biased. The pooled effects from meta-analyses that include truncated RCTs do not show a problem of upward bias even if the truncated RCT “freezes” future experiments; therefore, truncated RCTs should not be omitted from meta-analyses evaluating treatment effects. The superiority of one treatment over the other that is demonstrated at an interim analysis of a randomized clinical trial designed with appropriate stopping rules is likely a valid inference, even if the effect is slightly inflated. When making decisions to stop a trial before observing 50% of planned information, we recommend weighing the potential benefit of reducing the number of patients who receive inferior treatment against the potential cost of slight overestimation of the treatment effect. It is wise to apply statistical procedures to moderate over-estimation of treatment differences in trials that stop early, and if the effect is surprisingly large, to conduct additional studies if the estimated effect would change treatment policy.

Acknowledgments

Grant support:

Laura and John Arnold Foundation (S. Goodman)

NIH/NCI P30CA006973 (H. Wang, G. Rosner)

References

1. Bassler D, Briel M, Montori VM, et al. Stopping randomized trials early for benefit and estimation of treatment effects. *JAMA*. 2010; 303:1180–1187. [PubMed: 20332404]
2. Mueller PS, Montori VM, Bassler D, et al. Ethical issues in stopping randomized trials early because of apparent benefit. *Ann Intern Med*. 2007; 146:878–881. [PubMed: 17577007]
3. Bassler D, Ferreira-Gonzalez I, Briel M, et al. Systematic reviewers neglect bias that results from trials stopped early for benefit. *J Clin Epidemiol*. 2007; 60:869–873. [PubMed: 17689802]
4. Goodman SN. Stopping at nothing? Some dilemmas of data monitoring in clinical trials. *Ann Intern Med*. 2007; 46:882–887.
5. Goodman S, Berry D, Wittes J. Bias and trial stopped early for benefit. *JAMA*. 2010; 304:157.
6. Schou IM, Marschner IC. Meta-analysis of clinical trials with early stopping: an investigation of potential bias. *Stat Med*. 2013; 32:4859–4874. [PubMed: 23824994]
7. Korn EL, Freidlin B, Mooney M. Stopping or reporting early for positive results in randomized clinical trials: The National Cancer Institute cooperative group experience from 1990 to 2005. *J Clin Oncol*. 2009; 27:1712–1721. [PubMed: 19237631]
8. Goodman SN. Systematic reviews are not biased by results from trials stopped early for benefit. *J Clin Epidemiol*. 2008; 61:95–96. author reply 96. [PubMed: 18083466]
9. Bassler D, Montori VM, Briel M, et al. Reflections on meta-analyses involving trials stopped early for benefit: Is there a problem and if so, what is it? *Stat Methods Med Res*. 2013; 22:159–168. [PubMed: 22170891]
10. Montori VM, Devereaux PJ, Adhikari NK, et al. Randomized trials stopped early for benefit: a systematic review. *JAMA*. 2005; 294:2203–2209. [PubMed: 16264162]
11. Briel M, Bassler D, Wang AT, et al. The dangers of stopping a trial too early. *J Bone Joint Surg Am*. 2012; 94(Suppl 1):56–60. [PubMed: 22810449]
12. Prutsky GJ, Domecq JP, Erwin PJ, et al. Initiation and continuation of randomized trials after the publication of a trial stopped early for benefit asking the same study question: STOPIT-3 study design. *Trials*. 2013; 14:335. [PubMed: 24131702]

13. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics*. 1979; 35:549–556. [PubMed: 497341]
14. Pocock SJ, Hughes MD. Practical problems in interim analyses, with particular regard to estimation. *Control Clin Trials*. 1989; 10:209S–221S. [PubMed: 2605969]
15. Carlin, BP.; Louis, TA. Bayesian methods for data analysis. 3tf. Boca Raton, Florida: Chapman & Hall/CRC; 2008.
16. Freidlin B, Korn EL. Stopping clinical trials early for benefit: impact on estimation. *Clin Trials*. 2009; 6:119–125. [PubMed: 19342463]
17. Tsiatis AA, Rosner GL, Mehta CR. Exact confidence intervals following a group sequential test. *Biometrics*. 1984; 40:797–803. [PubMed: 6518248]
18. Whitehead J. On the bias of maximum likelihood estimation following a sequential test. *Biometrika*. 1986; 73:573–581.
19. Emerson SS, Fleming TR. Parameter estimation following group sequential hypothesis testing. *Biometrika*. 1990; 77:875–892.
20. Todd S, Whitehead J, Facey KM. Point and interval estimation following s sequential clinical trial. *Biometrika*. 1996; 83:453–461.
21. Liu A, Troendle JF, Yu KF, et al. Conditional maximum likelihood estimation following a group sequential test. *Biom J*. 2004; 46:760–768.
22. Pocock SJ, Hughes MD. Estimation issues in clinical trials and overviews. *Stat Med*. 1990; 9:657–671. [PubMed: 2145623]
23. USHHS, FDA. CDER, CBER. Guidance for industry. Providing clinical evidence of effectiveness for human drugs and biological products. 1998. <http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm078749.pdf> accessed 18 April 2016

Figure 1. Sampling distributions of treatment effect. a) Uniform distribution for the risk difference between the two arms, with a known risk of 40% in the control arm; b) Truncated normal distribution of the risk difference, with a known risk of 40% in the control arm; c) Risk ratio based on empirically observed distribution of effects from the clinical literature: $\ln(RR) \sim N(\ln(0.82), 0.192)$

Figure 2.

Box plots of over-estimate in the treatment effect estimate at analyses 1, 2, ..., 5. The trial outcomes of risk difference are sampled from uniform $(-0.4, 0.6)$ with the risk in control group = 0.4. The proportion of trials that stop at each interim analysis is included in the parenthesis. a) Absolute over-estimate. b) Relative over-estimate.

Figure 3.

Ratio of observed risk ratio (RR) versus true RR (in white color), and ratio of Bayesian estimate of RR versus true RR (in grey color). RR is sampled from $\ln(RR) \sim N(\ln(0.82), 0.192)$, with a known rate in control group=0.4. Bayesian analysis uses prior distribution $\ln(RR) \sim N(\ln(0.82), 0.192)$. a) Early stopped vs. completed trials. b) By interim analysis.

Simulation studies of the bias in the estimate of risk difference from truncated trials. Event risk is fixed at 40% in the control group.

Table 1

Distribution of treatment effects	Over-estimate ^a		Relative Over-estimate ^b		% trial that hit the boundary at any interim analysis
	Mean	Median	Mean	Median	
Uniform(−1, 1)	0.007	0.005	4.7%	1.5%	50.0
Truncated Normal					
N(0, 1)	0.008	0.007	6.3%	2.2%	51.1
N(0, 0.5 ²)	0.009	0.007	9.1%	2.3%	48.0
N(0, 0.25 ²)	0.013	0.011	11.5%	4.6%	39.7
N(0, 0.1 ²)	0.022	0.020	23.3%	16.5%	23.7
N(0, 0.05 ²)	0.041	0.040	43.1%	55.4%	9.7

^aOver-estimate is computed as the estimate minus the true value.

^bRelative over-estimate is the ratio of the estimate versus the true value.

Ratio of observed relative risk (RR) versus the true relative risk in simulation studies. RR is sampled from $\ln(RR) \sim N(\ln(0.82), 0.19^2)$.

Table 2

Control group risk	Interim look when trial stopped				Completed	
	1 st	2 nd	3 rd	4 th	5 th (final)	
0.4	0.58 ^a	0.88	0.96	0.99	1.00	1.02
	0.50, 0.67 ^b	0.76, 1.01	0.86, 1.09	0.91, 1.10	0.92, 1.09	0.94, 1.12
	0.1% ^c	13.6%	21.2%	14.5%	9.6%	50.6%
0.3	0.31	0.79	0.92	0.97	0.99	1.03
	0.26, 0.35	0.67, 0.93	0.80, 1.05	0.86, 1.10	0.89, 1.11	0.92, 1.15
	0.04%	6.8%	16.9%	15.4%	10.7%	60.9%
0.2	0.06 ^d	0.67	0.84	0.92	0.96	1.03
	0.06, 0.06	0.57, 0.78	0.72, 0.98	0.81, 1.05	0.85, 1.09	0.90, 1.19
	0.02%	2.7%	11.3%	14.8%	10.9%	71.2%

^aMedian

^b₁₀th and 90th percentile.

^cPercentage of trials that cross early stopping boundary at a particular stage.

^dThis is a single trial among all the simulated ones when one of the cells of the two-by-two table is 0.

Table 3

Ratio of Bayesian posterior estimate of relative risk (RR) versus the true relative risk. RR is sampled from $\ln(\text{RR}) \sim N(\ln(0.82), 0.19^2)$, with risk in control group=0.4.

Prior distribution	Interim analysis where trial stopped				Completed	Early stopped
	1 st	2 nd	3 rd	4 th		
$N(\ln(0.82), 0.19^2)$	1.13 ^a 0.95, 1.31 ^b	1.00 0.87, 1.16	1.00 0.90, 1.13	1.00 0.91, 1.12	1.00 0.92, 1.10	1.00 0.89, 1.14
$N(\ln(0.82), 0.25^2)$	0.98 0.83, 1.23	0.96 0.84, 1.10	0.98 0.88, 1.12	1.00 0.91, 1.11	1.01 0.93, 1.11	0.98 0.87, 1.11
$N(\ln(1), 0.25^2)$	1.09 0.91, 1.25	1.00 0.88, 1.16	1.02 0.91, 1.16	1.02 0.93, 1.14	1.03 0.94, 1.13	1.02 0.90, 1.15
$N(\ln(1), 0.35^2)$	0.90 0.75, 1.02	0.95 0.83, 1.10	0.99 0.89, 1.13	1.01 0.92, 1.12	1.03 0.94, 1.12	0.99 0.87, 1.12

^aMedian

^b10th and 90th percentile.

Table 4

“Freezing effect”. Simulation results of the meta-analysis for up to 10 randomized trials. Trials with a fixed value of true relative risk (RR), and event risk in control = 0.6.

True effect (RR)	N= 400 per arm (80% power to detect a RR=0.85 at onesided alpha 0.05)				N= 230 per arm (80% power to detect a RR=0.8 at one-sided alpha 0.05)				N=106 per arm (80% power to detect a RR=0.7 at with one-sided alpha 0.05)			
	Pooled RR estimate	Ratio of RR ^a	N of trials per series that had <10 trials ^b	% series with < 10 trials conducted	Pooled RR estimate	Ratio of RR ^a	N of trials per series that had <10 trials ^b	% series with <10 trials conducted	Pooled RR estimate	Ratio of RR ^a	N of trials per series that had <10 trials ^b	% series with <10 trials conducted
0.6	0.60 ^c 0.51, 0.67 ^d	1.00 ^c 0.84, 1.12 ^d	1 ^c 1, 1 ^d	100	0.59 0.48, 0.69	0.98 0.80, 1.15	1 1, 1	100	0.58 0.43, 0.67	0.97 0.72, 1.12	1 1, 2	100
0.7	0.69 0.60, 0.77	0.99 0.86, 1.10	1 1, 1	100	0.69 0.57, 0.77	0.98 0.81, 1.09	1 1, 2	100	0.67 0.54, 0.75	0.96 0.78, 1.07	2 1, 4	99.6
0.75	0.75 0.65, 0.81	0.99 0.87, 1.08	1 1, 1	100	0.73 0.62, 0.80	0.98 0.83, 1.06	1 1, 2	100	0.72 0.59, 0.80	0.96 0.78, 1.06	2 1, 6	97.0
0.8	0.79 0.70, 0.84	0.99 0.87, 1.05	1 1, 2	100	0.78 0.68, 0.84	0.97 0.85, 1.05	2 1, 4	99.7	0.77 0.63, 0.84	0.96 0.79, 1.04	3 1, 7	86.4
0.85	0.83 0.76, 0.88	0.98 0.89, 1.04	2 1, 4	99.7	0.83 0.73, 0.88	0.98 0.86, 1.03	3 1, 6	94.0	0.81 0.67, 0.88	0.96 0.79, 1.03	4 1, 8	64.5
0.9	0.88 0.81, 0.92	0.98 0.90, 1.02	3 1, 7	88.2	0.87 0.77, 0.92	0.97 0.86, 1.02	4 (4.2) 1, 8	67.5	0.86 0.71, 0.92	0.95 0.78, 1.02	5 1, 8	38.6
0.95	0.93 0.84, 0.96	0.98 0.89, 1.01	4 1, 8	42.0	0.92 0.80, 0.96	0.96 0.84, 1.01	5 1, 8	29.3	0.90 0.73, 0.96	0.94 0.77, 1.01	5 1, 9	18.3
1.0	0.97 0.86, 1.00	0.97 0.86, 1.00	5 1.2, 9	8.3	0.96 0.81, 1.00	0.96 0.81, 1.00	5 1, 9	8.1	0.94 0.75, 1.00	0.94 0.75, 1.00	5 1, 9	7.1

^aRatio of estimated RR from meta-analysis versus true RR

^bN of trials per series that had <10 trials is among the series experiencing “freezing effect”, the number of trials conducted in a series when a trial stops early for efficacy, thereafter no additional trials are conducted. These trials are included in a series for meta-analysis.

^c, ^dMedian, and 10th and 90th percentile.

Table 5

“Freezing effect”. Simulation results of the meta-analysis for up to 10 randomized trials, relative risk (RR) is sampled from $\ln(RR) \sim 0.5 * I_0(\ln(RR)) + 0.5 * N(\ln(0.82), 0.19^2)$, with event risk in control = 0.6.

N= 400 per arm (80% power to detect a RR=0.85 at one-sided alpha 0.05)				N= 230 per arm (80% power to detect a RR=0.8 at one-sided alpha 0.05)				N=106 per arm (80% power to detect a RR=0.7 at with one-sided alpha 0.05)			
Pooled RR estimate	Ratio of RR ^a	N of trials per series that had <10 trials ^b	% series with < 10 trials conducted	Pooled RR estimate	Ratio of RR ^a	N of trials per series that had <10 trials ^b	% series with <10 trials conducted	Pooled RR estimate	Ratio of RR ^a	N of trials per series that had <10 trials ^b	% series with <10 trials conducted
0.78 ^c 0.60, 0.93 ^d	0.98 0.87, 1.07	1 1, 5	42.3	0.76 0.58, 0.92	0.97 0.83, 1.06	1 1, 6	40.4	0.72 0.55, 0.90	0.95 0.77, 1.058	2 1, 7	34.9
0.77 (mean) 0.76 (geometric mean)	0.98 (mean) 0.97 (geometric mean)	2.0 (mean)		0.75 (mean) 0.74 (geometric mean)	0.96 (mean) 0.95 (geometric mean)	2.3 (mean)		0.72 (mean) 0.71 (geometric mean)	0.93 (mean) 0.92 (geometric mean)	3.0 (mean)	

^aRatio of estimated RR from meta-analysis versus true RR.

^bN of trials per series that had <10 trials is among the series experiencing “freezing effect”, the number of trials conducted in a series when a trial stops early for efficacy, thereafter no additional trials are conducted. These trials are included in a series for meta-analysis.

^c, ^dMedian, and 10th and 90th percentile.