Published in final edited form as:

Clin Trials. 2017 February ; 14(1): 48–58. doi:10.1177/1740774516666502.

Comparison of futility monitoring guidelines using completed phase III oncology trials

Qiang Zhang^{1,2,*}, Boris Freidlin³, Edward L Korn³, Susan Halabi⁴, Sumithra Mandrekar⁵, and James J Dignam^{1,6}

¹Statistics and Data Management Center, NRG Oncology, Philadelphia, PA, USA.

²Department of Biostatistics and Epidemiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA USA.

³Biometric Research Program, National Cancer Institute, Bethesda, MD, USA.

⁴ Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, NC 27705, USA.

⁵ Division of Biomedical Statistics and Informatics, Alliance Statistics and Data Center, Mayo Clinic, Rochester, MN, USA.

⁶ Department of Public Health Sciences, University of Chicago, Chicago, IL, USA.

Abstract

Background—Futility (inefficacy) interim monitoring is an important component in the conduct of phase III clinical trials, especially in life threatening diseases. Desirable futility monitoring guidelines allow timely stopping if the new therapy is harmful or if it is unlikely to demonstrate to be sufficiently effective if the trial were to continue to its final analysis. There are a number of analytical approaches that are used to construct futility monitoring boundaries. The most common approaches are based on conditional power, sequential testing of the alternative hypothesis or sequential confidence intervals. The resulting futility boundaries vary considerably with respect to the level of evidence required for recommending stopping the study.

Purpose—We evaluate the performance of commonly used methods using event histories from completed phase III clinical trials of the Radiation Therapy Oncology Group, Cancer and Leukemia Group B and North Central Cancer Treatment Group.

Methods—We considered published superiority phase III trials with survival endpoints initiated after 1990. There are 52 studies available for this analysis from different disease sites. Total sample size and maximum number of events (statistical information) for each study were calculated using protocol specified effect size, type I and type II error rates. In addition to the common futility approaches we considered a recently proposed linear inefficacy boundary approach with an early harm look followed by several lack-of-efficacy analyses. For each futility approach, interim test statistics were generated for three schedules with different analysis frequency and early stopping was recommended if the interim result crossed a futility stopping

* **Corresponding author:** Qiang Zhang, Statistics and Data Management Center, NRG Oncology, Philadelphia, Pennsylvania, USA. Zhangq@NRGOncology.org, Phone: 215-574-3197.

boundary. For trials not demonstrating superiority, the impact of each rule is summarized as savings on sample size, study duration and information time scales.

Results—For negative studies, our results show that the futility approaches based on testing the alternative hypothesis and repeated confidence interval rules yielded less savings (compared to the other two rules). These boundaries are too conservative, especially during the first half of the study (<50% of information). The conditional power rules are too aggressive during the second half of the study (>50% of information) and may stop a trial even when there is a clinically meaningful treatment effect. The linear inefficacy boundary with three or more interim analyses provided the best results. For positive studies, we demonstrated that none of the futility rules would have stopped the trials.

Conclusions—The linear inefficacy boundary futility approach is attractive from statistical, clinical, and logistical standpoints in clinical trials evaluating new anti-cancer agents.

Keywords

Futility monitoring; oncology; clinical trials; conditional power; repeated confidence intervals; testing alternative hypothesis; linear inefficacy boundary

Introduction

Randomized clinical trials (RCT) usually incorporate interim monitoring of outcome data for early evidence of efficacy or inefficacy/futility. The latter has been an essential part of RCT design since the 1980s.¹⁻¹⁴ The broader concept of inefficacy includes both harm and absence of tangible benefit.¹¹ If the experimental arm is actually inferior to the control arm then the design should allow stopping the study as soon as possible to minimize patient exposure to a harmful treatment; in this case the motivation for interim monitoring for inefficacy is clear. In quantifying benefit of early stopping for situations where the new therapy does not improve outcome (but is not shown to cause direct harm either) it is useful to remember that in life-threatening diseases like cancer many experimental treatments are based on adding a new (usually toxic) therapy to an existing standard of care (A+C vs C alone comparison). Therefore, reducing patient exposure to the experimental regimen provides benefit by sparing patients unnecessary toxicity and by allowing them to consider alternative therapies from which they may benefit. Furthermore, in trials designed to establish superiority of a new investigational therapy over an existing active standard treatment (A vs C comparison), there is no need to demonstrate that the experimental therapy is inferior to the standard therapy; compelling evidence of lack of improvement is sufficient to address the scientific question of the study. Moreover, in this case, observing lack of improvement for the experimental arm would generally be consistent with a modest reduction in benefit suggesting the possibility that patients are being harmed by not being treated with a standard of care that has proven benefit.¹¹ Therefore, the design should allow stopping early if it becomes clear that the experimental treatment does not offer tangible benefit to optimize patient treatment and to save time and resources.¹¹ It should be noted that in this article the term experimental treatment is used to refer to both treatments containing investigational agents (agents that are not available outside clinical trials) and treatments that involve non-investigational therapies used in settings where they are not

considered to be standard of care (or in settings where their clinical benefit has not been demonstrated). Since the level of evidence required for early inefficacy stopping depends on the prior information about the experimental treatment and the support for the treatment in the clinical community, the interim monitoring rules should be selected accordingly.

Commonly used inefficacy rules are generally based on one of the following three approaches: sequential testing of the alternative hypothesis, repeated confidence intervals, and conditional power. The testing of the alternative hypothesis approach can be implemented either by using a constant nominal significance level α ¹² or by using a type II spending function.⁴ The repeated confidence interval rule recommends stopping a study if at any of the scheduled interim analyses the pre-specified level confidence interval excludes the alternative hypothesis.⁵ The conditional power approach recommends stopping a study if given the observed data, the conditional power of rejecting the null hypothesis under specific assumptions about the future treatment effect (usually the alternative hypothesis) is less than a pre-specified threshold (such as 10%).¹³

Inefficacy/futility boundaries obtained using these analytical approaches vary considerably with respect to the level of evidence that is required throughout the trial for recommending stopping the study. It has been noted that some of the commonly used inefficacy boundaries are too conservative in the beginning and middle of the trial.^{10,15} On the other hand, some rules become too aggressive in the second half of the study and recommend inefficacy/futility stopping even when the observed results are consistent with tangible clinical benefit.⁹ To address these suboptimal properties, an alternative linear inefficacy boundary approach has been introduced.¹⁰ This strategy combines an early harm look (at 25% of information) with an intuitive lack-of-benefit rule after 40-50% of information and recommends stopping the study if the observed treatment effect favors the control arm by any amount¹⁴ (see the Appendix for details).

Formal application of inefficacy stopping rules results in deflation of the type I error of the study. However, since these rules are generally considered nonbinding guidelines (rather than strict rules), adjusting the efficacy threshold to account for the inefficacy monitoring is not recommended.^{16,17}

We applied these rules to the individual-level patient data from a subset of NCI-sponsored phase III trials in order to evaluate the performance characteristics of the common futility rules in real trial settings. (Note that this post-hoc analysis was not intended to reflect on the conduct or results of specific studies.)

Methods

The following criteria were used to select studies for analysis: published phase III superiority RCTs activated since 1990, with sample size above 100 patients using a time-to-event primary endpoint, with only two-arm designs or multi-arm designs that compare each experimental arm to the control arm. For the multi-arm trials each experimental vs. control arm comparison was treated as a separate trial. Factorial and dose-response designs were excluded. An individual-level dataset containing the most up-to-date data was obtained for

each trial. For each study the target sample size and number of events were obtained from the study protocol (if the number of events was not explicitly stated, it was estimated based on the protocol-specified type I and II error rates for the log-rank test). The calendar duration was calculated based on the date of the final analysis (Note that for studies with <90% of the protocol-specified information in the acquired dataset, the duration was extrapolated based on the observed information vs. calendar date patterns). For trials in which the primary endpoint had competing risks, the cause-specific log-rank test was used. All p-values in this paper are one-sided. To isolate the effect of the futility rules, we considered interim monitoring procedures that included only futility monitoring, (although the original trials had guidelines for superiority monitoring).

Three interim analysis schedules were considered with analyses times defined on the statistical information scale, i.e., the observed number of events over the total protocol-specified number of events. Schedule 1 had a single interim analysis at 50% information and schedule 2 started at 25% of information followed by two additional interim analyses. Schedule 3 started at 25% of information followed by 3-6 additional interim analyses (The timing and number of the analyses depended on the design type I and II errors; see the Appendix for details.). We tabulated results for the following approaches: (1) testing the alternative hypothesis with one-sided level of 0.0025,¹⁸⁻²⁰ (2) conditional power with 10% and 30% cut-offs (CP10 and CP30, respectively), (3) the repeated confidence interval approach using an O'Brien-Fleming spending function with the overall one-sided design level alpha, for example, using the trial design type I error rate of 0.025 (overall one-sided coverage of 97.5%), and (4) the linear inefficacy boundary rule (with harm look). Technical details of the methods are given in the Appendix. Approaches (1)-(3) were selected because they represent the most commonly used futility methods in oncology RCTs. Most available futility boundaries can be derived using these approaches: for example the Pampallona-Tsiatis boundary⁴ can be represented as conditional power of 50% (CP50).¹⁰ Alternatively, by varying the significance level used to test the alternative hypothesis, one can adjust the operating characteristics of the boundary derived by approach (1).¹⁵

For each trial the acquired dataset, which contained patient entry and event dates, was used to reconstruct the accumulating data during the trial and tabulate the relevant test statistics at 5% statistical information increments. The test statistics were compared (at the designated monitoring times) with each of the five futility boundaries to determine whether the study would be stopped by that rule. It should be noted that this investigation used the event dates from the acquired mature datasets. Therefore, due to variability in the timing of reporting of study events, the interim data obtained through this approach may not be exactly identical to the way the interim data were accrued in the trial.

The performance of each of the futility rules was measured in terms of the rule's ability to stop the study earlier on the statistical information scale. The practical implications of the stopping rules were quantified by the reduction in patient accrual and study calendar duration. For each trial, savings were calculated as the ratio of the information/sample-size/calendar-duration at the time of the hypothetical stopping over the corresponding maximum information/sample-size/calendar-duration. The savings were then averaged over all trials.

Results

The search identified 58 superiority trials/comparisons carried out by the Radiation Therapy Oncology Group (RTOG)²¹⁻⁴³ and Cancer and Leukemia Group B (CALGB^a)⁴⁴⁻⁵⁶ and North Central Cancer Treatment Group (NCCTG^a).⁵⁷⁻⁷⁰ Out of 58 trials/comparisons that met the selection criteria six stopped with less than 50% of the design-specified information, three stopped early due to external information and three for slow accrual. The remaining 52 trials were used in the analysis.

Results of applying the five futility rules (using Schedule 3) to each of the 52 trials are presented in Figure 1 on the information scale. In Figure 1, the trials are ordered by the final p-value evaluated with data representing 100% information. For trials with more than 90% of information the maximum calendar duration was obtained based on the calendar time of the analysis closest to 100% of the statistical information; for the four trials that finished with less than 90% of information the maximum duration was obtained using a linear extrapolation of the observed calendar vs. information pattern. For the purpose of evaluating the performance of the futility rules it is instructive to separate the trials into three categories based on the final p-values: (1) trials that suggest benefit for the experimental treatment (p-values <0.05), (2) trials that showed no meaningful benefit for the experimental treatment (p-values >0.1), and (3) trials with potentially equivocal results (p-values in the 0.05 - 0.1 range). We consider the application of the futility rules for each of the three scenarios separately. Note that this p-value-based classification is a simplification intended for evaluation of futility rules and may not accurately reflect the clinical interpretation of any given study.

There are six trials with the final p-value of less than 0.05 (trials 1-6); none of them were recommended for early stopping by any of the futility rules. (These studies demonstrated promising treatment effects for which one would not want to stop early for futility.) There are 39 trials with the final p-values above 0.1 (trials 14-58). These are the trials that would have benefited from early stopping. Indeed, early stopping was recommended for these trials by at least one futility rule. Table 1 shows the savings that would have been obtained by applying the futility rules using each of schedules on the information, duration and sample size scales. The three interim analysis schedule (Schedule 2) yielded considerably more savings than one interim analysis schedule (Schedule 1). The difference between the three-analysis schedule and the 4-7 analyses schedule (Schedule 3) is smaller. The savings are most pronounced on the statistical information and calendar scales. However, the 30% conditional power and the linear inefficacy boundaries are also shown to provide tangible improvement on the sample size scale. To further illustrate the savings in patient enrollment, Table 2 presents the proportion of trials that were stopped before the expected time for accrual completion. It can be seen that the 30% conditional power and the linear inefficacy boundaries provide the greatest reduction in patient exposure. There were seven trials (trials 7-13) with the final p-value in the 0.05-0.1 range. (For these trials with ambivalent results the appropriate stopping decisions would depend on the clinical setting and relative toxicity profiles of the treatment arms.) Among the seven trials, five were stopped by at least one

^aBoth CALGB and NCCTG are now part of the Alliance for Clinical Trials in Oncology (Alliance).

rule: Three trials (8,9,13) were recommended for stopping only by the conditional power rules at or after 80% of information, with trials 8 and 9 ultimately reporting marginally promising results (p-value .058). Two trials (11,12) were recommended for stopping early by all the rules and substantially early (40%-50% of information) by several rules. In these trials the observed treatment effect was delayed (we will return to this in the Discussion). The remaining two trials (7,10) were not stopped by any of the rules.

Discussion

From a practical perspective, interim futility/inefficacy monitoring rules can result in three types of early RCT stopping: (1) if the study has not reached the target accrual then stopping means no enrollment of new patients, potential treatment change for the patients who are still receiving experimental therapy and release of the study data, (2) if the study has completed accrual but some patients are still receiving the experimental therapy then stopping means potential treatment change for patients receiving the experimental therapy and release of the study data, and (3) if the study completed accrual and all patients are off the experimental treatment then stopping only means release of the study data. The implications of these three types of stopping for the study and patients are different: while with the third type the final analysis can be performed when the data matures with further follow-up, some of the data is lost in case of the first two types. At the same time, the first two types of stopping reduce the study patient exposure to potentially toxic and ineffective therapy. It should be noted, however, that all three types of early stopping involve study data release and thus could benefit patients outside the trial. Investigational agents are often being evaluated simultaneously in multiple clinical trials, therefore early access to negative study results may benefit patients treated with the investigational treatment on other trials. More importantly, when the trial experimental treatment is an intervention no longer considered investigational in the community and thus already more widely used (for example, radiation dose or delivery modes or combinations of radiation and approved drugs), early futility stopping provides benefit by sparing toxicity for a potentially large number of patients. For example, among the 22 RTOG studies that had early stopping recommended by at least one of the rules about 10 trials had experimental arms that were already in use in the community during the time the study was conducted. In most cases, the experimental arms on those studies either contained an additional therapy element or were potentially more toxic than the corresponding control arms: early release of the data from these studies would have been beneficial for the patients treated in the community.

The type of stopping that could be expected in a given setting depends on the projected accrual and event patterns: In poor prognosis settings (e.g. metastatic disease with median overall survival in the 1-2 year range), the possibility of stopping the trial before completion of accrual is higher than in good prognosis settings (e.g., adjuvant trials with median overall survival of 10+ years). Therefore, interim monitoring rules are often designed to correspond to the type of stopping that is likely to occur and the proportion of information that is expected to be available at that time. For example, if at the time the study is projected to complete accrual the fraction of the total design-specified statistical information is expected to be low (e.g. <25%), then the investigators may want to ensure that a harm look is performed no later than when a certain fraction of the total accrual is reached (e.g., 80%) to

ensure that if a strong negative trend is observed then accrual could be stopped. Regardless of whether the monitoring guidelines are specified in this way, the Data and Safety Monitoring Boards often consider whether patients in the study or in the community are being treated with the experimental therapy in making their recommendations.

A key purpose of futility monitoring is to allow stopping the study when sufficiently convincing evidence for lack of benefit of the experimental arm becomes available. Therefore, futility approaches that have aggressive stopping boundaries (especially later in the study), such as the 30% conditional power rule, should be used with caution as they may compromise the study ability to address its scientific goal by stopping the study early with equivocal results. For example, in the published analysis with full information, trial 8 (RTOG 9704) evaluating the addition of gemcitabine to chemo-radiation in the treatment of resected pancreatic cancer demonstrated a 9% increase (31% vs 22%) in three-year overall survival rates (results that some considered promising). If the 30% conditional power boundary had been used in the trial, it would have recommended closing the study for futility at 80% of information (10% conditional power guideline recommended closing at 90%). This aggressiveness of the conditional power-based boundaries can be particularly pronounced in clinical trials underpowered for moderate but still potentially clinically meaningful treatment effects: RTOG 9704 was designed to detect an increase in median overall survival from 18 to 27 months (hazard ratio=0.67); in the published analysis the observed median overall survival increase was from 16.9 to 20.5 months (estimated hazard ratio=0.82), a moderate but still potentially clinically relevant difference.

In clinical settings with crossing hazards (e.g., where there is a possibility for the treatment effect to be delayed or for an aggressive treatment strategy to do worse initially) more conservative interim rules requiring longer follow-up may be needed. For example, in trial 11 (RTOG 9802) which evaluated the addition of chemotherapy to radiation therapy in brain cancer, the overall survival curves crossed with the experimental arm having better long-term survival but doing worse than the control arm in the first two years (Figure 2). If this trial evaluating an aggressive/toxic treatment strategy were using any of the typical futility rules then due to the early transient negative trend, all of the rules would have recommended stopping: most at or before 50% (e.g., at 40% of information (linear inefficacy boundary and testing alternative hypothesis), and at 50% information (30% conditional power and repeated confidence intervals)). However, at the scheduled final analysis an important improvement in the long term overall survival was observed. After two additional years of follow-up, the initial results were confirmed with practice-changing improvement in the long term overall survival.⁷¹

It has been pointed out^{11,15} that the repeated confidence interval and 10% conditional power rules are excessively conservative at the beginning of the trial (e.g. at 25% of information the 10% conditional power rule requires Z-value of 3.18 in favor of the control arm for a 90% power trial). This can compromise the rule's ability to stop a study evaluating a harmful therapy in a timely fashion. An example of a situation where these rules could have been problematic is trial 39 (NCCTG 914652) which stopped early for slow accrual and ultimately reported a better survival on the control arm. Both the repeated confidence interval and 10% conditional power rules would have recommended stopping the study at

60% information. In contrast, the other three rules would have recommended stopping at 25% information.

Our results illustrate that inefficacy/futility rules like the linear inefficacy boundary and 30% conditional power rules with three or more interim analyses will minimize the number of exposed patients and the duration of the study. However, given the aggressive nature of the 30% conditional power rule, it is not recommended. In addition, using only one interim futility analysis could lead to a delay in dissemination of important clinical evidence and exposure of patient on and outside the trial.⁷² The improvement from switching from three to a more frequent 4-7 analyses plan is relatively modest. However, we argue for a more frequent analysis schedules (when feasible), as this incurs little statistical disadvantage⁷³ and the increased patient protection is worth the marginal logistical burden associated with the additional analyses. One limitation to our study is that retrospective analysis of interim data based on a post-study dataset may not accurately reflect the data at each interim analysis during the study.

In summary, the choice of the futility rule should depend on the data required to provide the clinical community with appropriately convincing evidence that the experimental treatment should not be used. For trials evaluating new potentially toxic treatments or trials where the experimental arm does not include the current standard of care, we recommend using linear inefficacy boundary rules with Schedule 3. In settings where stronger evidence for lack of activity is required, for example, when treatments being compared are routinely used in the community or when there is potential for delayed treatment effect (e.g., Ref. 74), more conservative inefficacy/futility rules potentially requiring minimally mature data may be required.

Acknowledgments

Funding

This project was supported by RTOG grant U10 CA21661, and CCOP grant U10 CA37422 and NRG Oncology SDMC grant U10 CA180822 and NRG Oncology grant U10 CA180868 from the National Cancer Institute (NCI). This study was supported by CALGB grants CA 155296 and U10 CA33601 and United States Army Medical Research W81XWH-15-1-0467. This study was also supported, in part, by grants from the National Cancer Institute to the North Central Cancer Treatment Group (CA25224) and to the Alliance for Clinical Trials in Oncology (Monica M. Bertagnolli, MD, Chair, CA31946) and to the Alliance Statistics and Data Center (Daniel J. Sargent, PhD, CA33601). This project is funded, in part, under grant 4100057652 from the Pennsylvania Department of Health, which specifically disclaims responsibility for any analyses, interpretations or conclusions.

Appendix: Futility boundaries

Consider a trial with a survival endpoint designed using a proportional hazards model to detect the hazard ratio θ_A (experimental over control arm). The final analysis, using the log-rank test, is planned when the full information of D events is observed to provide power $1 - \beta$ at overall one-sided significance level α . It can be shown⁵ that the standardized log-rank statistics at the information fraction t and the final analysis ($Z(t)$ and $Z(1)$) are approximately bivariate normal with means $(\log \theta \sqrt{t \frac{D}{4}}, \log \theta \sqrt{\frac{D}{4}})$ and covariance $\sqrt{t}$. The CP10 and CP30 boundaries are defined using the conditional power under the alternative hypothesis,^{2,10} i.e., the probability that the null hypothesis will be rejected at the final analysis given the

observed data and assuming that the design-specified target effect θ_A holds:

$P[Z(1) \geq C_{1-\alpha}|Z(t), \theta_A] = 1 - \Phi\left\{\frac{|C_{1-\alpha} - Z(t)\sqrt{t} - \theta_A(1-t)|}{\sqrt{1-t}}\right\}$. The corresponding stopping boundaries can be found according to the pre-selected rejection probabilities: the study is stopped if conditional power is less than 10% or 30%, respectively.

For stopping rules derived from testing alternative hypothesis,^{10,12} the study is stopped if the alternative hypothesis is rejected based on the log-rank test. For example, with a small alpha, say 0.0025, the futility boundaries would be to stop if $Z(t)$ is greater than

$C_{1-0.0025} - \log(\theta_A) \sqrt{\frac{d}{4}}$, in which d is the number of observed events at time of interim analysis and C_γ is γ quantile of the standard normal distribution. Repeated confidence intervals are obtained based on a group sequential design using the O'Brien-Fleming alpha spending function.^{5,10} The study is stopped if a pre-specified level repeated confidence interval excludes the alternative hypothesis at any of the designated interim analyses. A comprehensive inefficacy/futility procedure, linear inefficacy boundary,¹⁰ is defined as follows: First, at 25% of the total information a harm look is conducted. This look is designed to allow stopping if the very early results demonstrate that the experimental arm is inferior to the control: stopping for harm recommended if the lower 95% confidence bound for the hazard ratio is above 1. After the harm look, the inefficacy monitoring starts at

information time $t_0 = \left(\frac{C_{0.975}}{C_{1-\alpha} + C_{1-\beta}}\right)$ corresponding to the earliest information time at which $\hat{\theta}_{>1}$ would imply that the two sided 95% confidence interval for $\log \theta$ would exclude the design alternative $\log \theta_A$: this means that after t_0 observing a hazard ratio that favors the control arm is inconsistent with the targeted benefit. (Note that t_0 is between 40% and 50% of the total information for a typical phase III trial.). Starting at t_0 the linear inefficacy boundary is constructed to have stopping cutoff (on the $\log \theta$ scale) satisfy the following two conditions: (1) the observed hazard ratio is inconsistent with the design alternative $\log \theta_A$, and (2) the observed hazard ratio does not exceed 20% of the targeted benefit $\log \theta_A$ (this condition is needed to prevent stopping for inefficacy when the experimental arm is doing moderately better).¹⁰ For the information time t in the interval $[t_0, 1]$ the stopping cutoff for the linear inefficacy boundary is:

$$0.20 * \log \theta_A \frac{(C_{1-\alpha} + C_{1-\beta})^2 t - C_{0.975}^2}{(C_{1-\alpha} + C_{1-\beta})^2 - C_{0.975}^2}.$$

For each of the rules above, three interim analysis schedules were considered. Schedule 1 had one interim analysis at $t=50\%$ information. Schedule 2 had the first interim analysis at $t=25\%$, followed by two equally spaced interim analyses starting at t_0 (where t_0 was determined using the formula above and the study-specified α and β). Schedule 3 had the first interim analysis at $t=25\%$, followed by interim analyses at 10% information increments starting at t_0 (where t_0 was determined using the formula above and the study-specified α and β).

References

1. Halperin M, Lan KK, Ware JH, et al. An aid to data monitoring in long-term clinical trials. *Control Clin Trials*. 1982; 3:311–323. [PubMed: 7160189]
2. Lan KK, Wittes J. The B-value: a tool for monitoring data. *Biometrics*. 1988; 44:579–585. [PubMed: 3390511]
3. Pocock SJ. When to stop a clinical trial. *BMJ*. 1992; 305:235–240. [PubMed: 1392832]
4. Pampallona S, Tsiatis AA. Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *J Stat Plan Inference*. 1994; 42:19–35.
5. Jennison, C., Turnbull, BW. Group sequential methods with applications to clinical trials. Chapman & Hall/CRC; Boca Raton: 2000.
6. Snappin S, Chen MG, Jiang Q, et al. Assessment of futility in clinical trials. *Pharm Stat*. 2006; 5:273–281. [PubMed: 17128426]
7. DeMets DL. Futility approaches to interim monitoring by data monitoring committees. *Clin Trials*. 2006; 3:522–529. [PubMed: 17170036]
8. Lachin JM. A review of methods for futility stopping based on conditional power. *Stat Med*. 2005; 24:2747–2764. [PubMed: 16134130]
9. Freidlin B, Korn EL. A comment on futility monitoring. *Control Clin Trials*. 2002; 23:355–366. [PubMed: 12161079]
10. Freidlin B, Korn EL, Gray R. A general inefficacy interim monitoring rule for randomized clinical trials. *Clin Trials*. 2010; 7:197–208. [PubMed: 20423925]
11. Freidlin B, Korn EL. Monitoring for lack of benefit: a critical component of a randomized clinical trial. *J Clin Oncol*. 2009; 27:629–633. [PubMed: 19064977]
12. Fleming TR, Harrington DP, O'Brien PC. Designs for group sequential tests. *Control Clin Trials*. 1984; 5:348–361. [PubMed: 6518769]
13. Lan KK, Simon R, Halperin M. Stochastically curtailed tests in long-term clinical trials. *Commun Stat C*. 1982; 1:207–219.
14. Ellenberg SS, Eisenberger MA. An efficient design for phase III studies of combination chemotherapies (with discussion). *Cancer Treat Rep*. 1985; 69:1147–1154. [PubMed: 4042093]
15. Anderson JR, High R. Alternatives to the standard Fleming, Harrington, and O'Brien futility boundary. *Clin Trials*. 2011; 8:270–276. [PubMed: 21441286]
16. European Medicines Agency. Reflection paper on methodological issues in confirmatory clinical trials with flexible design and analysis plan. 2007
17. Cook, TD., DeMets, DL. Introduction to statistical methods in clinical trials. Chapman and Hall/ CRC; Boca Raton: 2008.
18. Petrylak DP, Tangen CM, Hussain MH, et al. Docetaxel and estramustine compared with mitoxantrone and prednisone for advanced refractory prostate cancer. *N Engl J Med*. 2004; 351:1513–1520. [PubMed: 15470214]
19. Williamson SK, Crowley JJ, Lara PN, et al. Phase III trial of paclitaxel plus carboplatin with or without tirapazamine in advanced non-small-cell lung cancer: Southwest Oncology Group Trial S0003. *J Clin Oncol*. 2005; 23:9097–9104. [PubMed: 16361616]
20. Lara PN Jr, Natale R, Crowley J, et al. Phase III trial of irinotecan/cisplatin compared with etoposide/cisplatin in extensive-stage small-cell lung cancer: clinical and pharmacogenomic results from SWOG S0124. *J Clin Oncol*. 2009; 27:2530–2535. [PubMed: 19349543]
21. Fu KK, Pajak T, Trotti A, et al. A Radiation Therapy Oncology Group (RTOG) phase III randomized study to compare hyperfractionation and two variants of accelerated fractionation to standard fractionation radiotherapy for head and neck squamous cell carcinomas: First report of RTOG 90-03. *Int J Radiat Oncol Biol Phys*. 2000; 48:7–16. [PubMed: 10924966]
22. Forastiere AA, Goepfert H, Maor M, et al. Concurrent chemotherapy and radiotherapy for organ preservation in advanced laryngeal cancer. *N Engl J Med*. 2003; 349:2091–2098. [PubMed: 14645636]

23. Cooper JS, Pajak T, Forastiere AA, et al. Postoperative concurrent radiotherapy and chemotherapy for high-risk squamous-cell carcinoma of the head and neck. *N Engl J Med*. 2004; 350:1937–1944. [PubMed: 15128893]
24. Machtay M, Pajak T, Suntharalingam M, et al. Radiotherapy with or without erythropoietin for anemic patients with head and neck cancer: a randomized trial of the radiation therapy oncology group (RTOG 99-03). *Int J Radiat Oncol Biol Phys*. 2007; 69:1008–1017. [PubMed: 17716826]
25. Ang KK, Harris J, Wheeler R, et al. Human papilloma virus and survival of patients with oropharyngeal cancer. *N Engl J Med*. 2010; 363:24–35. [PubMed: 20530316]
26. Trotti A, Zhang Q, Bentzen SM, et al. Randomized trial of hyperfractionation versus conventional fractionation in T2 squamous cell carcinoma of the vocal cord (RTOG 9512). *Int J Radiat Oncol Biol Phys*. 2014; 89:958–963. [PubMed: 25035199]
27. Hanks GE, Pajak T, Porter A, et al. Phase III trial of long-term adjuvant androgen deprivation after neoadjuvant hormonal cyoreduction and radiotherapy in locally advanced carcinoma of the prostate: The Radiation Therapy Oncology Group protocol 92-02. *J Clin Oncol*. 2003; 21:3972–3978. [PubMed: 14581419]
28. Roach M, DeSilvio M, Lawton C, et al. Phase III trial comparing whole-pelvic versus prostate-only radiotherapy and neoadjuvant versus adjuvant combined androgen suppression: Radiation Therapy Oncology Group 94-13. *J Clin Oncol*. 2003; 21:1904–1911. [PubMed: 12743142]
29. Albain KS, Swann SR, Rusch VW, et al. Phase III comparison of concurrent chemotherapy plus radiotherapy (CT/RT) and CT/RT followed by surgical resection for stage IIIA (Pn2) non-small cell lung cancer (NSCLC): Initial results from intergroup trial 0139 (RTOG 93-09). *Lancet*. 2009; 374:379–386. [PubMed: 19632716]
30. Movsas B, Scott C, Langer C, et al. Randomized trial of amifostine in locally advanced non-small-cell lung cancer patients receiving chemotherapy and hyperfractionated radiation: Radiation Therapy Oncology Group trial 98-01. *J Clin Oncol*. 2005; 23:2145–2154. [PubMed: 15800308]
31. Cairncross G, Berkey B, Shaw E, et al. Phase III trial of chemotherapy plus radiotherapy compared with radiotherapy alone for pure and mixed anaplastic oligodendroglioma: Intergroup Radiation Therapy Oncology Group trial 9402. *J Clin Oncol*. 2006; 24:2707–2714. [PubMed: 16782910]
32. Knisely J, Berkey B, Chakravarti A, et al. A phase III study of conventional radiation therapy plus thalidomide vs. conventional radiation therapy for multiple brain metastases (RTOG 0118). *Int J Radiat Oncol Biol Phys*. 2008; 71:79–86. [PubMed: 18164847]
33. Gore EM, Bae K, Wong SJ, et al. Phase III Comparison of prophylactic cranial irradiation versus observation in patients with locally advanced non-small-cell lung cancer: primary analysis of Radiation Therapy Oncology Group study RTOG 0214. *J Clin Oncol*. 2011; 29:272–278.
34. Ajani JA, Winter KA, Gunderson LL, et al. Fluorouracil, mitomycin, and radiotherapy vs fluorouracil, cisplatin, and radiotherapy for carcinoma of the anal canal. *JAMA*. 2008; 299:1914–1921. [PubMed: 18430910]
35. Regine W, Winter KA, Abrams RA, et al. Fluorouracil vs gemcitabine chemotherapy before and after fluorouracil-based chemotherapy following resection of pancreatic adenocarcinoma: A randomized controlled trial. *JAMA*. 2008; 299:1019–1026. [PubMed: 18319412]
36. Scott, WJ., Curran, WK., et al. Proc ASCO, # 1546. Los Angeles, CA: 1998. Long term results of RTOG-9006: A randomized trial of hyperfractionated radiotherapy to 72 Gy and carmustine vs. standard RT and carmustine for malignant glioma patients with emphasis on anaplastic astrocytoma patients..
37. Murray KJ, Scott C, Greenberg HM, et al. A randomized phase III study of accelerated hyperfractionation versus standard in patients with unresected brain metastases: a report of the radiation therapy oncology group (RTOG) 9104. *Int J Radiat Oncol Biol Phys*. 1997; 39:571–574. [PubMed: 9336134]
38. Bradley JD, Scott BC, Paris KJ, et al. A phase III comparison of radiation therapy with or without recombinant β -interferon for poor-risk patients with locally advanced non-small-cell lung cancer (RTOG 93-04). *Int J Radiat Oncol Biol Phys*. 2002; 52:1173–1179. [PubMed: 11955727]
39. Souhami L, Seiferheld W, Brachman D, et al. Randomized comparison of stereotactic radiosurgery followed by conventional radiotherapy with carmustine to conventional radiotherapy with

- carmustine for patients with glioblastoma multiforme: report of radiation therapy oncology group 93-05 protocol. *Int J Radiat Oncol Biol Phys.* 2004; 60:853–860. [PubMed: 15465203]
40. Prados MD, Scott C, Sandler H, et al. A phase III randomized study of radiotherapy plus procarbazine, CCNU and vincristine (PCV) with or without DUDR for the treatment of anaplastic astrocytoma: a preliminary report of RTOG 9404. *Int J Radiat Oncol Biol Phys.* 1999; 45:1109–1115. [PubMed: 10613302]
 41. Jones CU, Hunt D, McGowan DG, et al. Radiotherapy and short-term androgen deprivation for localized prostate cancer. *N Engl J Med.* 2011; 365:107–118. [PubMed: 21751904]
 42. Andrews DW, Scott BC, Sperduto PW, et al. Whole brain radiation therapy with or without stereotactic radiosurgery boost for patients with one to three brain metastases: phase III results of the RTOG 9508 randomised trial. *Lancet.* 2004; 363:1665–1672. [PubMed: 15158627]
 43. Shaw EG, Wang M, Coons SW, et al. Randomized trial of radiation therapy plus procarbazine, lomustine, and vincristine chemotherapy for supratentorial adult low-grade glioma: Initial results of RTOG 9802. *J Clin Oncol.* 2012; 30:3065–3070. [PubMed: 22851558]
 44. Miller AA, Herndon JE, Hollis DR, et al. Schedule dependency of 21-day oral versus 3-day intravenous etoposide in combination with intravenous cisplatin in extensive-stage small-cell lung cancer: A randomized phase III study of the cancer and leukemia group B. *J Clin Oncol.* 1995; 13:1871–1879. [PubMed: 7636529]
 45. Clamon G, Herndon JE, Cooper R, et al. Radiosensitization with carboplatin for patients with unresectable stage III non-small-cell lung cancer: A phase III trial of the cancer and leukemia group B and the eastern cooperative oncology group. *J Clin Oncol.* 1999; 17:4–11. [PubMed: 10458211]
 46. Kantoff PW, Halabi S, Conaway M, et al. Hydrocortisone with or without mitoxantrone in men with hormone-refractory prostate cancer: Results of the cancer and leukemia group B 9182 study. *J Clin Oncol.* 1999; 17:2506–2513. [PubMed: 10561316]
 47. McClay EF, Bogart J, Herndon JE, et al. A phase III trial evaluating the combination of cisplatin, etoposide, and radiation therapy with or without tamoxifen in patients with limited-stage small cell lung cancer: Cancer and leukemia group B study (9235). *Am J Clin Oncol.* 2005; 28:81–90. [PubMed: 15685040]
 48. Colacchio T, Niedzwiecki D, Compton CC, et al. Phase III study of adjuvant immunotherapy with MoAb 17-1A following resection for stage II adenocarcinoma of the colon (CALGB 9581). *J Clin Oncol.* 2004; 22(Suppl):3522.
 49. Strauss GM, Herndon JE, Maddaus MA, et al. Adjuvant paclitaxel plus carboplatin compared with observation in stage IB non-small-cell lung cancer: CALGB 9633 with the Cancer and Leukemia Group B, Radiation Therapy Oncology Group, and North Central Cancer Treatment Group Study Groups. *J Clin Oncol.* 2008; 26:5043–5051. [PubMed: 18809614]
 50. Lilenbaum RC, Herndon JE, List MA, et al. Single-agent versus combination chemotherapy in advanced non-small-cell lung cancer: The cancer and leukemia group B (study 9730). *J Clin Oncol.* 2005; 23:190–196. [PubMed: 15625373]
 51. Niell HB, Herndon JE, Miller AA, et al. Randomized phase III intergroup trial of etoposide and cisplatin with or without paclitaxel and granulocyte colony-stimulating factor in patients with extensive-stage small-cell lung cancer: Cancer and Leukemia Group B trial 9732. *J Clin Oncol.* 2005; 23:3752–3759. [PubMed: 15923572]
 52. Marcucci G, Moser B, Blum W, et al. A phase III randomized trial of intensive induction and consolidation chemotherapy ± oblimersen, a pro-apoptotic Bcl-2 antisense oligonucleotide in untreated acute myeloid leukemia patients >60 years old. *J Clin Oncol.* 2007; 25(Suppl):7012.
 53. Vokes EE, Herndon JE, Kelley MJ, et al. Induction chemotherapy followed by chemoradiotherapy compared with chemoradiotherapy alone for regionally advanced unresectable stage III non-small-cell lung cancer: Cancer and Leukemia Group B. *J Clin Oncol.* 2004; 25:1698–1704.
 54. Kindler HL, Niedzwiecki D, Hollis D, et al. Gemcitabine plus bevacizumab compared with gemcitabine plus placebo in patients with advanced pancreatic cancer: Phase III trial of the Cancer and Leukemia Group B (CALGB 80303). *J Clin Oncol.* 2010; 28:3617–3622. [PubMed: 20606091]

55. Saltz LB, Niedzwiecki D, Hollis D, et al. Irinotecan fluorouracil plus leucovorin is not superior to fluorouracil plus leucovorin alone as adjuvant treatment for stage III colon cancer: Results of CALGB 89803. *J Clin Oncol.* 2007; 25:3456–3461. [PubMed: 17687149]
56. Kelly WK, Halabi S, Carducci M, et al. Randomized, double-blind, placebo-controlled phase III trial comparing docetaxel and prednisone with or without bevacizumab in men with metastatic castration-resistant prostate cancer: CALGB 90401. *J Clin Oncol.* 2012; 30:1534–1540. [PubMed: 22454414]
57. Buckner JC, Schomberg PJ, McGinnis WL, et al. A phase III study of radiation therapy plus carmustine with or without recombinant interferon-alpha in the treatment of patients with newly diagnosed high-grade glioma. *Cancer.* 2001; 92:420–433. [PubMed: 11466698]
58. Bonner JA, Sloan JA, Shanahan TG, et al. Phase III comparison of twice-daily split-course irradiation versus once-daily irradiation for patients with limited stage small-cell lung carcinoma. *J Clin Oncol.* 1999; 17:2681–2691. [PubMed: 10561342]
59. Ingle JN, Suman VJ, Mailliard JA, et al. Randomized trial of tamoxifen alone or combined with fluoxymesterone as adjuvant therapy in postmenopausal women with resected estrogen receptor positive breast cancer. North Central Cancer Treatment Group trial 89-30-52. *Breast Cancer Res Treat.* 2006; 98:217–222. [PubMed: 16538529]
60. Goldberg RM, Hatfield AK, Kahn M, et al. Prospectively randomized north central cancer treatment group trial of intensive-course fluorouracil combined with the l-isomer of intravenous leucovorin, oral leucovorin, or intravenous leucovorin for the treatment of advanced colorectal cancer. *J Clin Oncol.* 1997; 15:3320–3329. [PubMed: 9363861]
61. Markovic S, Suman VJ, Dalton RJ, et al. Randomized, placebo-controlled, phase III surgical adjuvant clinical trial of megestrol acetate (megace) in selected patients with malignant melanoma. *Am J Clin Oncol.* 2002; 25:552–556. [PubMed: 12477996]
62. Bonner JA, McGinnis WL, Stella PJ, et al. The possible advantage of hyperfractionated thoracic radiotherapy in the treatment of locally advanced non-small cell lung carcinoma: Results of a North Central Cancer Treatment Group phase III study. *Cancer.* 1998; 82:1037–1048. [PubMed: 9506347]
63. Goldberg RM, Langdon R, Sargent DJ, et al. Phase III trial of 5-fluorouracil (5FU), levamisole (LEV), +/- leucovorin (LV) as adjuvant treatment (ADJ) of resected metastatic colorectal cancer (M-CRC) an NCCTG/MD Anderson/SWOG study. *Proc ASCO.* 2001; 20:133a.
64. Edmonson JH, Suman VJ, Dalton RJ, et al. Comparison of conventional dose and double dose carboplatin in patients receiving cyclophosphamide plus carboplatin for advanced ovarian carcinoma: A North Central Cancer Treatment Group study. *Cancer Invest.* 2001; 19:597–602. [PubMed: 11486702]
65. Creagan ET, Suman VJ, Dalton RJ, et al. Phase III clinical trial of the combination of cisplatin, dacarbazine, and carmustine with or without tamoxifen in patients with advanced malignant melanoma. *J Clin Oncol.* 1999; 17:1884–1890. [PubMed: 10561229]
66. Martenson JA, Willett CG, Sargent DJ, et al. Phase III study of adjuvant chemotherapy and radiation therapy compared with chemotherapy alone in the surgical adjuvant treatment of colon cancer: Results of intergroup protocol 0130. *J Clin Oncol.* 2004; 22:3277–3283. [PubMed: 15249584]
67. Schild SE, Stella PJ, Geyer SM, et al. Phase III trial comparing chemotherapy plus once-daily or twice-daily radiotherapy in stage III non-small-cell lung cancer. *Int J Radiat Oncol Biol Phys.* 2002; 54:370–378. [PubMed: 12243810]
68. Johnson EA, Marks RS, Mandrekar SJ, et al. Phase III randomized, double-blind study of maintenance CAI or placebo in patients with advanced non-small cell lung cancer (NSCLC) after completion of initial therapy (NCCTG 97-24-51). *Lung Cancer.* 2008; 60:200–207. [PubMed: 18045731]
69. Goldberg M, Sargent DJ, Morton RF, et al. Randomized controlled trial of reduced-dose bolus fluorouracil plus leucovorin and irinotecan or infused fluorouracil plus leucovorin and oxaliplatin in patients with previously untreated metastatic colorectal cancer: A north american intergroup trial. *J Clin Oncol.* 2004; 22:23–30. [PubMed: 14665611]

70. Alberts SR, Sargent DJ, Nair S, et al. Effect of oxaliplatin, fluorouracil, and leucovorin with or without cetuximab on survival among patients with resected stage III colon cancer: A randomized trial. *JAMA*. 2012; 307:1383–1393. [PubMed: 22474202]
71. Van der Bent MJ. Practice changing mature results of RTOG study 9802: another positive PCV trial makes adjuvant chemotherapy part of standard of care in low-grade glioma. *Neuro Oncol*. 2014; 16:1570–1574. [PubMed: 25355680]
72. Freidlin B, Korn EL. Erythropoietin to treat anaemia in patients with head and neck cancer. Letters to the Editor. *Lancet*. 2014; 363:81.
73. Freidlin B, Korn EL, George SL. Data monitoring committees and interim monitoring guidelines. *Control Clin Trials*. 1999; 20:395–407. [PubMed: 10503800]
74. Dignam JJ, Bryant J, Wieand HS, et al. Early stopping of a clinical trial when there is evidence of no treatment benefit: protocol B-14 of the National Surgical Adjuvant Breast and Bowel Project. *Control Clin Trials*. 1998; 19:575–588. [PubMed: 9875837]

Figure 1.

Information scale: stopping times using Schedule 3. Plus signs denote stopping for futility; circles denote the protocol-specified final analysis (or analysis at the maximum information available for trials that did not reach the protocol-specified information). THA – orange, LIB20 – blue, CP10 – green, CP30 – red, RCI - yellow.

Figure 2.
RTOG 9802 overall survival curves

Table 1
Inefficacy/Futility monitoring results for negative trials (n=39) (Average, median, interquartile range)

Boundary	Savings for One interim analysis			Savings for Three interim analyses			Savings for 6-7 interim analyses		
	Information scale (%)	Calendar time scale (%)	Sample size scale (%)	Information scale (%)	Calendar time scale (%)	Sample size scale (%)	Information scale (%)	Calendar time scale (%)	Sample size scale (%)
LIB20	22.4 1.9, 0-50	17.2 0, 0-38.3	5.3 0, 0-0	30.3 35, 0-50	24.0 22.2, 0-39.2	8.3 0, 0-14.9	33.2 40, 4.6-50	26.3 26.0, 0-39.2	9.5 0, 0-16.3
RCI	8.7 0, 0-2	6.2 0, 0-0	0 0, 0-0	18.4 20, 0-30	14.5 14.5, 0-22.9	2.7 0, 0-0	21.6 25, 1-37	17.1 17.8, 0-26.3	3.4 0, 0-5.8
THA	10.4 0, 0-4.6	6.0 0, 0-0	0.4 0, 0-0	16.4 0, 0-30	11.4 0, 0-17.3	2.8 0, 0-0	17.6 1.9, 0-32.5	12.0 0, 0-18.6	3 0, 0-0
CP10	11.3 0, 0-18.4	8.3 0, 0-10.0	0.8 0, 0-0	23.5 25, 2-35	18.9 18.7, 0-26.2	4.5 0, 0-5.7	29 30, 17.5-40	21.9 23.0, 13.3-28.4	5.4 0, 0-7.8
CP30	22.6 1.9, 0-50	18.6 0, 0-37.9	4.4 0, 0-0	35.1 30, 25-50	29.3 29.7, 16.4-38.6	7.3 0, 0-12.8	40.3 40, 30-50	31.8 31.7, 20.2-38.6	8.8 1.9, 0-12.9

For information and sample-size scales the percentages are standardized to the protocol-specified maximum information and sample size, respectively. For the calendar duration scale the maximum duration was obtained based on the time of the final analysis; for trials that finished with less than 90% of information the maximum duration was obtained using linear extrapolation of the observed calendar vs. information pattern. LIB: linear inefficacy boundary, RCI: repeated confidence interval, THA: testing alternative hypothesis, CP: conditional power

Table 2

Percent of trials stopped before the expected time of accrual completion among the negative trials (n=39)

Boundary	Schedule 1 (One interim analysis)	Schedule 2 (Three interim analyses)	Schedule 3 (4-7 interim analyses)
LIB20	23%	36%	41%
RCI	0%	23%	33%
THA	2.6%	18%	21%
CP10	5.1%	28%	38%
CP30	21%	41%	54%

LIB: linear inefficacy boundary, RCI: repeated confidence interval, THA: testing alternative hypothesis, CP: conditional power