Published in final edited form as:

Clin Trials. 2013 April ; 10(2): 250–256. doi:10.1177/1740774512474374.

Monitoring futility in a two-by-two factorial design: The SPS3 experience

Leslie A McClure^a, Christopher S Coffey^b, and George Howard^a

^aDepartment of Biostatistics, University of Alabama at Birmingham, Birmingham, AL, USA

^bDepartment of Biostatistics, University of Iowa, Iowa City, IA, USA

Abstract

Background—For studies with two-by-two factorial designs, the complexity of determining an appropriate futility analysis plan is increased as compared to studies where patients are randomized to one treatment. Issues that must be addressed include the possibility of a significant interaction and the need to determine how to proceed given evidence of futility in one arm. Suggested approaches include a two-stage plan, which first assesses futility of the interaction term and proceeds to examine the main effects, given sufficient evidence that no interaction is present, and variations on one-stage plans, which assume the trial will not be stopped for futility in the interaction.

Purpose—To discuss different approaches to monitoring futility in two-by-two factorial clinical trials and compare their properties.

Methods—We utilized a simulation study, designed to mimic the Secondary Prevention of Small Subcortical Strokes (SPS3) Study, to determine which approach to monitoring futility in two-by-two factorial studies had the most desirable statistical properties.

Results—We found that in most scenarios typical of clinical trials, monitoring futility in each arm simultaneously was superior to or as good as monitoring the interaction and then assessing futility in each arm only when the interaction was deemed futile. Monitoring each arm simultaneously lead to early stopping more often when no treatment effect was present, and lower average sample numbers (ASNs). The exception to this was the unlikely case when a qualitative interaction was present.

Limitations—We assumed that one-sided tests were to be performed, and only assessed some of the possible methods for monitoring futility under the study design.

Conclusions—Futility monitoring in two-by-two factorial studies should proceed by assessing each arm simultaneously, rather than monitoring the interaction first. If sizeable interactions are anticipated, study design, rather than study monitoring, should account for this.

Introduction

Clinical trials may be stopped prior to the planned completion for several reasons. First, if at an interim time point, the experimental treatment is clearly better than the standard treatment, it would be unethical to continue to treat patients with the standard treatment

© The Author(s), 2013

Author for correspondence: Leslie A McClure, Department of Biostatistics, University of Alabama at Birmingham, RPHB 327G 1530, 3rd Ave S, Birmingham, AL 35294-0022, USA. lmcclure@uab.edu.

Conflict of interest

None declared.

while delaying access to the new (superior) treatment. Conversely, if the experimental therapy is established as clearly inferior to the standard treatment, it would similarly be unethical to continue randomization. Additionally, trials may be stopped for safety if a treatment under study produces an unacceptable rate of adverse events. Finally, a study could be stopped for futility if it becomes clear prior to the study's completion that the likelihood of finding a significant result is low.

Stochastic curtailment is a common approach to assess futility where the conditional power is calculated at an interim time point as the probability that the null hypothesis will be rejected at the end of the study, given the results obtained to an interim time point, and a fixed value for the future treatment difference [1,2]. Generally, the original treatment difference assumed during the design phase is taken as the assumption for the 'future' data; however, other possible expected treatment differences can be used that provide a range of values for the conditional power [3]. Typically, a study is stopped for futility if the conditional power falls below some predefined threshold, most commonly either 0.2 or 0.1 for main effects [4].

Clinical trial factorial designs are often used to assess multiple treatments simultaneously in the context of a single study, and to assess interactions between treatments. Under the assumption of no interactions between treatments, a factorial study in essence allows assessing the efficacy of two (or more) treatments for a cost similar to a single trial assessing a single treatment. In addition, fully factorial designs allow assessing and quantifying interactions; however, sufficient power to detect this interaction requires that the sample size be about four times greater than factorial studies assessing main effects alone [5]. Furthermore, if a significant interaction exists, issues arise in the examination of the main effects for the individual treatments. Although factorial designs permit separate inferences about more than one treatment within a single study, and allow the ability to assess potential interactions between treatments, these benefits are often at odds with each other. That is, the ability to make inference about more than one treatment assumes the absence of an interaction, while the ability to assess interactions relies on an increase in sample size beyond what is necessary to assess the treatment arms simultaneously. The design of factorial studies requires an assumption regarding the existence of an interaction, with profound implications regarding the study size. If one assumes an interaction is present, a rather large study would be required. However, if one assumes no interaction, the study may be underpowered should one actually exist. Thus, often, factorial studies are designed with only one of these two advantages in mind [5]. A common compromise is to design a study without assuming an interaction but declaring that the interaction will be assessed at the time of the final analysis.

Whitehead [6] not only describes a procedure for monitoring efficacy in factorial trials but also acknowledges work to be done. Whitehead's approach determines at an interim time point which treatment combination may be stopped early due to being significantly worse than another, with the interaction only assessed when the final treatment combination is stopped. Others have examined monitoring both efficacy and toxicity jointly (e.g., Refs. [7] and [8]), or monitoring a single composite endpoint, comprising multiple outcomes [9]. Additionally, methods exist to monitor efficacy for global tests (e.g., Refs. [10] and [11]); however, the explicit problem of monitoring futility in factorial studies has seemingly received little attention. Thus, this article attempts to fill a gap in the literature by focusing on our experience of monitoring futility in a factorial designed study, the Secondary Prevention of Small Subcortical Strokes (SPS3) Study.

SPS3 is a randomized clinical trial with a two-by-two factorial design, employed to examine the effect of both an antiplatelet intervention and blood pressure control on stroke recurrence

[12]. For the anti-platelet intervention, participants were randomized to receive either aspirin alone, or aspirin plus clopidogrel, while for the blood pressure control arm, participants were randomized to undergo either usual blood pressure control (systolic blood pressure (BP): 130–140 mmHg) or intensive blood pressure control (systolic BP: <130 mmHg). Thus, patients were randomized to one of four combinations of antiplatelet therapy and blood pressure control. Based on data from previous studies, SPS3 was designed assuming no interaction between the two treatments; however, the analysis plan called for assessing the interaction at each interim analysis and at end of the study. If the interaction is not significant, primary analyses will proceed by assessing the time to recurrent stroke in each treatment arm separately. Should a significant interaction be found, the primary analysis will assess appropriate subgroups. The initially proposed interim monitoring plan included three interim looks for efficacy based on the use of Haybittle–Peto bounds and no interim futility assessment. At an early Data and Safety Monitoring Board (DSMB) meeting, the DSMB requested the addition of an interim futility monitoring plan. It was unclear how ‘futility’ should be defined in this setting, and the literature offered little guidance. Even if one treatment arm is deemed futile at an interim time point, patients would still be recruited and followed in order to assess the other treatment. In order to assess the statistical properties of various approaches, a simulation study was undertaken.

While conditional power approaches for futility monitoring are relatively straightforward for single-factor designs, evaluating multiple factors simultaneously complicates approaches for early termination for futility, and is the focus of this report. If one treatment in a particular arm is shown to be more effective at an interim time point, all patients should receive the more effective treatment for the remainder of the study, and randomization to the second arm can continue as planned. However, if one treatment in a particular arm is shown to be futile (but not unsafe), there are several different options for assessing futility in the study, and several different decisions can be made based on the outcome of that monitoring. These decisions are influenced by many issues, including adverse event and safety concerns, economic issues, and issues of feasibility.

Motivated by the SPS3 trial, we examine approaches for assessing futility using conditional power in the context of a two-by-two factorial design, in the presence or absence of an interaction, and treatment effects in only one or both arms. While the methodology described herein focuses on conditional power in the SPS3 trial, the concepts can be applied more broadly to decisions regarding futility in factorial studies.

Monitoring futility

Monitoring futility in a two-by-two factorial setting can be achieved in different ways; however, each raises issues regarding making decisions concerning the continuation of the study. One possibility is to assess futility in each treatment arm separately. Different decisions could be made if futility is found in only one arm at an interim assessment. The decision may be made to stop randomizing to the treatment arm deemed futile and continue the study only in the second treatment arm. In subsequent interim analyses, if the second treatment arm is then deemed futile, the study would be stopped. This is akin to conducting futility analyses as if two separate clinical trials were being run, and is easy to implement both statistically and clinically. However, eliminating randomization to the futile arm curtails the ability to make more precise statements about the true difference between the two treatments in that arm at the end of the study, as well as the ability to assess the interaction at the end of the study, especially in studies designed to detect significant interactions.

Another option, given futility is detected in only one treatment arm at an interim analysis, is to continue randomizing to both treatment arms, but only perform subsequent futility analyses on the other arm. Given that the first treatment arm has been deemed futile, the study would stop early concluding futility only if a subsequent interim analysis revealed futility in the second treatment arm. This option allows for ease in interpretation of the final trial results.

A third option, given the conclusion of futility in only one treatment arm at an interim analysis, is to continue to randomize to both arms, as well as to continue to assess futility in both arms, and to stop the study only if both arms are deemed futile *simultaneously* at a subsequent interim analysis. This option is the only one of the three that is consistent with the idea that the study should stop only if both arms simultaneously reveal that there is little chance of finding any treatment differences at the end of the study. However, this method is a bit more challenging to implement statistically, as this problem is two dimensional, and methodology, assuming each arm is monitored as a separate trial, implies two one-dimensional stopping bounds.

However, all of these previous approaches fail to acknowledge the possibility of an interaction between treatments, and a decision to stop a study could limit the potentially important ability to describe the presence or absence of clinically important effect modification present between the treatments. A final approach to monitoring futility in a two-by-two factorial study that takes this issue into consideration is to assess futility of detecting the interaction, even if the original study design did not call for an interaction. Then, futility will be assessed in the individual treatment arms only once the possibility of detecting a sizeable interaction has been deemed futile. Any of the approaches described above can then be used in evaluating the futility in the individual treatment arms. If an interaction is present, depending on whether it is qualitative or quantitative [13], the possibility of detecting a treatment difference in an individual arm may appear to be futile when in fact there is a treatment effect present. While this situation is not anticipated in most clinical trials, it also cannot be ruled out without evidence against it. In fact, not considering the interaction in the analysis may be inappropriate – if indeed an interaction exists, reporting the results absent the interaction could lead to erroneous clinical conclusions. However, this approach is difficult to implement, and the statistical properties of the method are not currently described. We will investigate the properties of this approach herein, and compare them to the properties of monitoring futility using one of the options described above.

Simulation study

We performed a simulation study to examine the properties of two of the different approaches to monitoring futility described above: (1) monitoring both arms simultaneously, and continuing to randomize to both arms, even if one arm is deemed futile and (2) assessing futility in each arm only if the possibility of detecting a sizable interaction is deemed futile. In the first approach, although randomization will continue to both arms, futility monitoring will continue only in the arm that has not been deemed futile, and early stopping will occur only if the second arm is deemed futile at a future interim analysis. We chose this approach from those described above after considering the strengths and weaknesses of the alternative approaches. Since we feel that monitoring futility only in a single arm more closely mirrors a single-arm study, continuing to randomize to both arms should be the ‘default’ approach, unless other factors such as safety concerns warrant stopping randomization to that arm at an interim assessment, in order to ease the interpretation of the final study results. This is different from simultaneous monitoring of both arms, in which the futility boundary is determined based on the joint distribution of the test statistics for each arm.

Although the outcomes from the SPS3 study are time until event data, we simulated 5000 datasets from a Gaussian distribution for ease of interpretation, under several different conditions:

1. The completely null case, in which there was no interaction nor were treatment effects present in either arm;
2. The case of a treatment effect in one arm only, so that no interaction was present, and main effects occurred only for treatment A;
3. The case for which treatment effects occur in both arms, so that no interaction was present, but main effects occurred for both treatments A and B; and
4. The case for which an interaction was present between treatments A and B.

For the fourth condition, we allowed the magnitude of the interaction to be driven by the subgroup analyses, so that the appropriate contrasts within the subgroup analyses had sufficient power to be detected. Furthermore, we assumed both quantitative interactions and qualitative (cross-over) interactions. While it is most often assumed that if an interaction does exist in a clinical trial, it is quantitative in nature [13], for completeness, we investigated both possibilities.

Treatment effects were specified so that the simulation study had 80% power to detect the treatment effect of interest at the end of the study, with a type I error rate of 5% for each test. We considered tests for main effects to be one-sided and the test for the interaction to be two-sided. We examined a sample size of 150 patients, with futility analyses performed after one-third and two-third of the patients were enrolled in the study. Assuming a standard deviation of 1, the conditions above led to a detectable main effect of 0.41, and a detectable interaction (difference in the differences of means, i.e., differences of means for group 'A' vs. 'a' among those randomized to group 'B' between the differences in means for group 'A' vs. 'a' among those randomized to group 'b') of 0.46.

We computed conditional power considering two plausible effect sizes for the 'future' data: the hypothesized effect size and the observed effect size. These were chosen to represent values that include both conservative and anticonservative possibilities. We utilized a futility index of 0.80 for main effects, and 0.50 for interactions. This implies that if the conditional probability that a treatment difference will be observed in the main effect by the end of the study is 0.20 or less, the study would stop early for futility. Similarly, if the conditional probability that a significant interaction will be observed by the end of the study is 0.50 or less, then futility in the interaction would be declared [3]. We used a futility index of 0.50 for the interaction to mimic that agreed upon by the investigators and the DSMB for SPS3, in order to have the ability to detect an 'unexpected' interaction, since the study was not powered to detect one. This allows us to be sufficiently 'liberal' in this decision, so as to avoid the concern with interpreting main effects if an interaction *may* be present.

Results

Tables 1 and 2 provide the percentage of studies out of the 5000 simulations that were stopped early by the two different monitoring approaches considered: (1) monitoring both arms simultaneously, and continuing to randomize to both arms, even if one arm is deemed futile (referred to as: two-arm monitoring); and (2) assessing futility in each arm only if a sizable interaction is deemed futile (referred to as: interaction monitoring) for each of the different effect sizes utilized in the conditional power calculation. For each of the five different scenarios considered, the percentages of simulated studies that were stopped at each of the first and second interim assessments are presented. The percent of studies that

then went to completion can be inferred by subtracting the sum of those stopped at each of the first and second look from 100%.

Tables 3 and 4 provide the percentage of studies out of the 5000 simulations for which a correct decision was made in each treatment arm, as well as the average sample number (ASN) required for that combination of each of the five scenarios and the monitoring plan, again for each of the different effect sizes utilized in the conditional power calculation. These results include whether the appropriate result (main effect or subgroup test in the case of a significant interaction) was found at the end of the study.

Table 1 indicates that it is not likely to stop a study early and declare that detection of a significant treatment effect is futile after only one-third of the data are collected, regardless of how the study is monitored, when utilizing the hypothesized effect size for the conditional power calculation. However, under the completely null scenario, the two-arm monitoring method would lead to early stopping for futility after two-thirds of the data are collected 24% of the time, while assessing the interaction first will lead to early stopping for futility 14% of the time. Under this scenario, it appears that the two-arm monitoring plan performs better. This is confirmed by the results in Table 3, which indicates that the ASN is smaller for the two-arm monitoring plan, under the completely null scenario. The two-arm monitoring plan leads to a correct decision 93% of the time in each arm, while assessing the interaction initially leads to a correct decision approximately as frequently. Neither method leads to a correct decision 95% of the time, likely because we tested our hypotheses in both treatment arms but made no adjustment for the multiple testing.

When the observed effect size is used to compute the conditional power, the likelihood of stopping after one-third of the data are collected increases, as indicated in Table 2. In fact, with the two-arm monitoring method, and using the observed effect size in the conditional power calculation, almost 50% of the studies would be stopped early for futility. And while the proportion of correct decisions does not differ depending on the effect size used for the conditional power calculation, the ASN decreases as the effect size becomes more conservative (Table 4).

When a treatment effect is present in either one or both arms, or when there is a quantitative interaction present, the two methods of monitoring futility differ little in terms of the frequency of stopping early, the likelihood of making a correct decision, and the ASN, regardless of the effect size used in the conditional power calculation. When a treatment effect is not present in one of the arms, the proportion of times that a correct decision is made for that arm is lower than 95%, again presumably due to the fact that multiple testing was not accounted for.

When a qualitative interaction is present, we find that the two-arm monitoring procedure stops early after two-thirds of the data are collected 23% of the time, while monitoring the interaction initially leads to early stopping for futility after two-thirds of the data are collected only 1% of the time when the hypothesized effect size is used in the conditional power calculation. Neither method stops early for futility after only one-third of the data are collected. As the method for computing conditional power becomes more conservative, the two-arm monitoring procedure leads to early stopping in over 50% of the studies, while monitoring the interaction leads to early stopping in less than 5% of the studies. Monitoring the interaction initially makes a correct decision approximately 60% of the time in each arm, while the two-arm monitoring plan makes a correct decision about one-third of the time in each arm. The ASN is uniformly lower for the two-arm monitoring scheme (Table 4), a reflection of that method's tendency to stop early more often than the method in which the interaction is first assessed.

Discussion

In this article, we examined monitoring futility in a two-by-two factorial clinical trial: (1) monitoring both arms simultaneously, and continuing to randomize to both arms, even if one arm is deemed futile and (2) assessing futility in each arm only if a sizable interaction is deemed futile. We simulated several different scenarios in which treatment effects may or may not occur, assessed different methods for computing the conditional power, and compared the frequency of stopping early after one-third and two-thirds of the data are collected. For each, the frequency of making a correct conclusion in each treatment arm, and the ASN between the two monitoring plans were reported.

We chose to assess the approach under which if one arm is deemed futile at the first interim assessment, randomization would continue to occur in that arm even though futility would be assessed only in the other arm, which allows for the ability to make a valid comparison in the arm deemed futile at the end of the study. We did so after weighing the strengths and weaknesses of the different options described above but made some assumptions in choosing, such as the futile arm incurs little to no additional cost to the study, there is a non-threatening safety profile for the futile arm (i.e., it is ethical to continue to randomize to that arm), and that the conditional power does not increase at the second interim analysis. However, this may not always be the case, and the decision to continue to randomize to a futile arm, as well as the decision to discontinue futility assessments in that arm, should be made after careful consideration from both the study personnel and the safety monitoring board.

In only one scenario was there gain monitoring the interaction prior to assessing the individual treatment arms. However, this scenario was that of a qualitative interaction, one that does not occur frequently in clinical trials settings. In fact, adjustments in the sample size and monitoring plan made during the design phase are appropriate if a strong qualitative interaction is assumed a priori. In terms of monitoring, a study for which a qualitative interaction was assumed would not assess futility or efficacy in each arm in the absence of consideration of the interaction. In addition, we benefited from designing our simulation study with a particular effect size for the interactions, allowing us to assess conditional power in the interaction utilizing the planned effect size. However, realistically, factorial studies are often not powered to detect a sizeable interaction, but rather to minimize necessary sample size to find significant results in each arm. Although most factorial studies assume that there is not an interaction and are not designed to detect a potentially present interaction, most also recognize the importance of testing this assumption during the analysis phase of the study, underscoring the importance of both monitoring and subsequently testing the interaction. This further complicates monitoring the interaction, as it is unclear what effect size to use for the conditional power calculations. Thus, the two-arm monitoring plan provides the most benefit in a variety of realistic clinical trials settings.

Acknowledgments

The authors would like to acknowledge the members of the SPS3 Study DSMB for their lively discussions regarding the topic addressed in this manuscript: K Michael Welch, MD (Chairperson); William Clarke, PhD; Jeffrey Cutler, MD; Karen Furie, MD; Karen Johnston, MD; and Matthew Mayo, PhD.

Funding

This work was supported in part by a grant from the National Institutes of Health/National Institute for Neurological Disorders and Stroke (grant number: 5U01NS038529).

References

1. Halperin M, Lan KKG, Ware JH, Johnson NJ, DeMets DL. An aid to data monitoring in long-term clinical trials. *Control Clin Trials*. 1982; 3:311–323. [PubMed: 7160189]
2. Lan KKG, Simon R, Halperin M. Stochastically curtailed tests in long-term clinical trials. *Seq Anal*. 1982; 1:207–219.
3. Dmitrienko, A.; Molenberghs, G.; Chuang-Stein, C.; Offen, W. *Analysis of Clinical Trials Using SAS: A Practical Guide*. Cary, NC: SAS Institute Inc; 2005.
4. Jennison, C.; Turnbull, BW. *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton, FL: Chapman & Hall/CRC; 2000.
5. Piantadosi, S. *Clinical Trials: A Methodologic Perspective*. Hoboken, NJ: John Wiley & Sons; 2005.
6. Whitehead, J. *The Design and Analysis of Sequential Clinical Trials*. 2nd edn.. Chichester: John Wiley & Sons; 1997. revised
7. Cook RJ, Farewell VT. Guidelines for monitoring efficacy and toxicity responses in clinical trials. *Biometrics*. 1994; 50:1146–1152. [PubMed: 7786995]
8. Thall PF, Simon RM, Estey EH. Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Stat Med*. 1995; 14:357–379. [PubMed: 7746977]
9. Green SB, Freedman LS. Early stopping of prevention trials when multiple outcomes are of interest: A discussion. *Stat Med*. 1982; 3:311–323.
10. Jennison C, Turnbull BW. Exact calculations for sequential $t\chi^2$ and F tests. *Biometrika*. 1991; 78:133–141.
11. Tang D, Geller NL, Pocock SJ. On the design and analysis of randomized clinical trials with multiple endpoints. *Biometrika*. 1993; 49:23–30.
12. Benavente O, White CL, Pearce L, et al. The Secondary Prevention of Small Subcortical Strokes (SPS3) Study. *Int J Stroke*. 2011; 6:164–175. [PubMed: 21371282]
13. Gail M, Simon R. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*. 1985; 41:361–372. [PubMed: 4027319]

Table 1

Percentage of studies stopped early for futility by two different monitoring approaches, out of 5000 replicated datasets ($N = 150$), using the hypothesized effect size for the CP calculation

Scenario	Method of monitoring futility			
	Two-arm monitoring		Interaction monitoring	
	First look (n = 50)	Second look (n = 100)	First look (n = 50)	Second look (n = 100)
Completely null	<0.1	24	0	14
Treatment effect n arm A	0	0.9	0	0.6
Treatment effects in both arms	0	0	0	0
Quantitative interaction	0	<0.1	0	0
Qualitative interaction	0	23	0	1

CP: conditional power.

Table 2

Percentage of studies stopped early for futility by two different monitoring approaches, out of 5000 replicated datasets ($N = 150$), using the observed effect size for the CP calculation

Scenario	Method of monitoring futility			
	Two-arm monitoring		Interaction monitoring	
	First look (n = 50)	Second look (n = 100)	First look (n = 50)	Second look (n = 100)
Completely null	30	20	7	16
Treatment effect in arm A	6	3	1	1
Treatment effects in both arms	1	0.5	0.2	0.2
Quantitative interaction	1	0.4	0.2	0.2
Qualitative Interaction	31	20	2	2

CP: conditional power.

Table 3

Percentage of studies where a correct decision was made in each arm, and the ASN, out of 5000 replicated datasets (N = 150), using the hypothesized effect size for the CP calculation

Scenario	Method of monitoring futility					
	Two-arm monitoring			Interaction monitoring		
	Arm A (%)	Arm B (%)	ASN	Arm A (%)	Arm B (%)	ASN
Completely null	93	93	133.8	93	92	143.6
Treatment effect in arm A	79	90	149.6	79	89	149.7
Treatment effects in both arms	81	82	150	81	82	150
Quantitative interaction	86	87	150	86	87	150
Qualitative interaction	51	49	138.3	61	60	149.3

CP: conditional power; ASN: average sample number.

Percentage of studies where a correct decision was made in each arm, and the ASN, out of 5000 replicated datasets (N = 150), using the observed effect size for the CP calculation

Table 4

Scenario	Method of monitoring futility					
	Two-arm monitoring			Interaction monitoring		
	Arm A (%)	Arm B (%)	ASN	Arm A (%)	Arm B (%)	ASN
Completely null	94	94	109.5	93	93	134.6
Treatment effect in arm A	76	90	143.0	78	89	147.9
Treatment effects in both arms	80	81	148.7	81	82	149.7
Quantitative interaction	85	86	148.7	86	87	149.7
Qualitative interaction	33	32	108.5	60	58	146.7

CP: conditional power; ASN: average sample number.