Published in final edited form as:

J Biopharm Stat. 2020 November 01; 30(6): 1050–1059. doi:10.1080/10543406.2020.1818253.

Futility stopping in clinical trials, optimality and practical considerations

Yen Chang¹, Tianhao Song¹, Jane Monaco¹, Anastasia Ivanova^{1,*}

¹Department of Biostatistics, CB #7420, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-7420, USA

Abstract

Stopping for futility is a useful tool in a clinical trial. It is widely used in single-arm trials in oncology and in many two-arm trials. We review three stopping rules for futility. We give recommendations for the optimal timing of futility looks in two-stage trials in terms of the information fraction and the probability of stopping under the alternative hypothesis. We discuss futility stopping in trials with substantial uncertainty about the variability of the outcome and in crossover trials.

Keywords

Simon's design; Phase 2 trial; Futility stopping; Crossover trial

1. Introduction

Multi-stage designs with possibility to stop for futility at an interim analysis are popular. Two-stage designs with futility stopping are widely used in single arm phase 2 trials in oncology with a binary outcome where the probability of positive response is compared to the historical response rate (Simon 1989). A survey of phase 2 oncology trials (Ivanova et al. 2016) reported that 40% of phase 2 trials in oncology published in leading oncology journals in 2010–2015 used Simon's design. Futility stopping is a useful feature in other therapeutic areas, especially in phase 2 trials with the goal of evaluating efficacy of a novel therapy. Early stopping can help protect patients from receiving a treatment that is not effective (Snapinn 2006). It helps investigators make good use of resources such as time and money. For example, Herring et al. (2013) described a three-period crossover trial in patients with obstructive sleep apnea with excessive daytime sleepiness. An experimental treatment, MK-0249, was compared with placebo and an active comparator, modafinil, with respect to reduction of excessive daytime sleepiness. The trial used an adaptive design from Ivanova et al. (2009) and included an interim analysis to stop for futility. The trial was stopped for futility because the response in the MK-0249 arm was not much different from placebo and superior response was observed in the modafinil arm (Herring et al. 2013).

* aivanova@bios.unc.edu.

A number of futility stopping rules have been described in the literature (Wiener, Ivanova, and Koch 2020). A trial can be stopped for futility if the likelihood to achieve success, or conditional power, given the current trend is low (Lan, Simon, and Halperin 1982; Lan and Wittes 1988). One can set up a futility rule based on an interim test statistic to test for treatment effect (Fleming, Herrington, and O'Brien 1984). The trial is stopped if the value of the test statistic is negative or low, where negative values of the test statistic correspond to the placebo being better than the active treatment. He, Lai, and Liao (2012) proposed a rule where the trial is stopped if we have evidence that the alternative hypothesis is not true. The rule is based on testing the alternative hypothesis rather than testing the null hypothesis. We discuss these rules and give recommendations for which rule to use.

Several optimality criteria have been proposed for selecting interim futility and efficacy rules (Simon 1989; Jung et al. 2004; Mander and Thompson 2010; Mander et al. 2012; Wason and Mander 2012). Simon (1989) has proposed to select futility stopping rules that minimize the expected sample size under the null hypothesis. Jung et al. (2004) proposed to minimize the weighted average of the maximum total sample size and the expected sample size under the null hypothesis. Mander and Thompson (2010) and Mander et al. (2012) proposed optimality criteria for two-stage designs in single-arm trials that allow early stopping for either futility or efficacy. Wason and Mander (2012) applied the optimality criteria from Mander et al. (2012) to a two-arm trial with continuous outcomes.

As noted by Snappin (2006), “in practice, statistical stopping rules and stopping boundaries tend to be guidelines rather than hard rules, and the sponsor usually has some flexibility.” Even though futility stopping reduces the type I error rate, this reduction is often ignored when designing comparative studies and interim futility boundaries are determined separately from efficacy boundaries. “In this way, the sponsor will have full flexibility to follow or ignore the futility stopping boundaries, without inflating type I error.” (Snappin 2006). Since futility boundaries are non-binding in most trials, this is the approach we consider here. In this paper we describe how to set up an approximately optimal futility rule in a two-stage trial and briefly discuss trials with two futility interim analyses. Recommendations are given in terms of which decision rule to use, the timing of the futility looks and the probability to stop for futility under the alternative hypothesis.

2. Futility stopping rules

Consider a two-arm clinical trial with n subjects in each arm with one or two interim analyses for futility. The first interim analysis is performed when outcome data from n_1 subjects in each arm are obtained, the second interim analysis is performed when data from n_2 subjects in each arm are available, $0 < n_1 < n_2 < n$. Corresponding information fractions are $t_1 = n_1/n$ and $t_2 = n_2/n$, $0 < t_1 < t_2 < 1$. Let $\mathbf{Y} = (Y_1, \dots, Y_n)$ be the response data in the active treatment arm and $\mathbf{X} = (X_1, \dots, X_n)$ be the response data in the placebo arm, with $X_i \sim N(\mu_X, \sigma^2)$, $Y_i \sim N(\mu_Y, \sigma^2)$ for $i=1, \dots, n$. We assume that higher values of X_i and Y_i correspond to more favorable outcomes. At the end of the trial, the two arms are compared using a one-sided test with type I error of α . The sample size n (per group) is selected to yield power of $1 - \beta$ when the true treatment effect is $\delta = \mu_Y - \mu_X > 0$ and the variance of the

outcome is σ^2 . For a trial with no futility stopping, we have that $\delta = \sqrt{2/n_0}(Z_\beta + Z_\alpha)\sigma$. Here, we use n_0 to denote the required sample size per arm in a trial with no interim analyses. We define $Z_\tau = \Phi^{-1}(1 - \tau)$ where $\Phi^{-1}(\cdot)$ is the inverse of the cumulative distribution function of the standard normal distribution. Define the effect size, or a standardized effect size, as δ/σ . In a trial with n_0 subjects per arm and no interim analyses, the effect size that gives $1 - \beta$ power is equal to

$$\Delta = \delta/\sigma = \sqrt{2/n_0}(Z_\beta + Z_\alpha). \quad (1)$$

Let $\bar{Y}_t$ and $\bar{X}_t$ be the sample means of $(Y_1, \dots, Y_{tn})$ and $(X_1, \dots, X_{tn})$ respectively, where $t = t_1$ or t_2 . The variance σ^2 is estimated by the pooled sample variance

$$\hat{\sigma}^2 = (tn - 1)(S_{\bar{X},t}^2 + S_{\bar{Y},t}^2)/(2tn - 2) = (S_{\bar{X},t}^2 + S_{\bar{Y},t}^2)/2, \text{ where } S_{\bar{X},t}^2 \text{ and } S_{\bar{Y},t}^2 \text{ are the sample variances of the placebo arm and the active arm at the interim, respectively. Consider two } t \text{ statistics: } Z_t = \frac{\bar{Y}_t - \bar{X}_t}{\hat{\sigma}\sqrt{1/(tn) + 1/(tn)}} \text{ and } Z_t^F = \frac{\bar{Y}_t - \bar{X}_t - \delta}{\hat{\sigma}\sqrt{1/(tn) + 1/(tn)}}.$$

Below, we describe several futility rules. The cutoff for the test statistic used in a futility rule can be set based on the desired probability to stop for futility under the null hypothesis, H_0 , $\mu_Y - \mu_X = 0$, or under the alternative hypothesis, H_A , $\mu_Y - \mu_X = \delta$.

2.1. Rule 1, stopping for futility based on conditional power (Lan, Simon, and Halperin 1982; Lan and Wittes 1988)

The trial is stopped for futility when the conditional power at the interim analysis assuming the current trend, $\Phi\left\{\sqrt{\frac{n}{n-tn}}\left(\frac{\sqrt{n}}{\sqrt{2}\hat{\sigma}}(\bar{Y}_t - \bar{X}_t) - Z_\alpha\right)\right\}$, is below a cutoff c^{CP} . If we want the probability to stop under H_0 to be γ , then the cutoff value $c^{CP}(\gamma)$ is

$$\Phi\left\{\frac{1}{\sqrt{t(1-t)}}(Z_1 - \gamma - \sqrt{t}Z_\alpha)\right\}. \text{ If we want the probability to stop under } H_A \text{ to be } \xi, \text{ the cutoff value } c^{CP}(\xi) \text{ is } \Phi\left\{\frac{1}{\sqrt{t(1-t)}}(Z_1 - \xi - \sqrt{t}Z_\alpha + \frac{\delta/\sigma}{\sqrt{2/(tn)}})\right\}.$$

2.2. Rule 2, stopping for futility based on the test statistic testing $\mu_Y - \mu_X = 0$

The trial is stopped for futility when $Z_t < c(\gamma)$. Since Z_t is distributed $N(0,1)$ under H_0 , the probability to stop for futility under the H_0 is equal to γ if $c(\gamma) = \Phi^{-1}(\gamma)$. With $c(\gamma)$, the probability to stop for futility under H_A will be equal to $\Phi\left\{\Phi^{-1}(\gamma) - \frac{\delta/\sigma}{\sqrt{2/(tn)}}\right\}$. If we want to find the cutoff, $c(\xi)$, yielding a probability of ξ to stop under the H_A , then $c(\xi)$ is $\frac{\delta/\sigma}{\sqrt{2/(tn)}} + \Phi^{-1}(\xi)$ (Fleming, Herrington, O'Brien 1984).

2.3. Rule 3, stopping for futility based on the test statistic testing $\mu_Y - \mu_X - \delta \geq 0$

He, Lai and Liao (2012) described a test for futility with the one-sided test $H_{0,F}$: $\mu_Y - \mu_X - \delta \geq 0$. The trial is stopped for futility when $Z_t^F < c^F$, since it supports that the $H_{0,F}$ is not true. Since Z_t^F is $N(0,1)$ under H_A , with a cutoff $c^F(\xi) = \Phi^{-1}(\xi)$, the probability

to stop for futility under the H_A is always equal to ξ . The probability to stop for futility under H_0 is equal to $\Phi\left\{\Phi^{-1}(\xi) + \frac{\delta/\sigma}{\sqrt{2/(tn)}}\right\}$.

It is easy to show algebraically that Rule 1 and Rule 2 are equivalent. By construction, Rules 1 and 2 maintain a given probability of stopping for futility under H_0 , and Rule 3 maintains a given probability of stopping for futility under H_A .

The statistics and cutoffs for decisions rules for a two-stage design are tabulated in Table 1. For brevity, we dropped the subscript 1 indicating the first and only futility look.

In Section 3 we briefly discuss optimal three-stage designs. There is no closed form expression for cutoff values for the second and following interim analyses. We give the equations for the second stage cut-off in Appendix I. For a three-stage design, for Rule 2, the cutoffs are defined based on the probability to stop under the null hypothesis after stage 1, γ_1 , and the probability to stop after stage two while the trial was not stopped after stage 1, γ_2 . For Rule 3, the cutoffs are defined by similar probabilities to stop under the alternative hypothesis, ξ_1 and ξ_2 . The R code for obtaining the cutoff values for a three-stage trial using Rule 2 or Rule 3 is provided in the Appendix II.

3. Optimal stopping rules for futility

3.1 Optimal two-stage designs with futility stopping

Designing a futility rule for a multi-stage trial requires specification of both the timing of futility looks in terms of the information fractions and the desired probability to stop for futility under the alternative or the null hypothesis at each look. Additionally, one needs to specify the required sample size to achieve the desired power. Including a futility look adds an opportunity to fail to reject the null hypothesis, reducing the likelihood of showing significance at the end of the trial. As a result, both type I error rate and power are reduced. When futility stopping is non-binding, as we consider here, the primary analysis at the end of the trial is performed at the nominal α -level (instead of slightly larger α allowed due to futility stopping when futility stopping is binding). Because power is reduced, the total sample size in a trial with a futility rule needs to be higher compared to the trial without a futility rule to maintain the desired power. We denote the required sample size per arm without futility as n_0 and sample size per arm taking futility into consideration as n , such that $n_0 < n$. To specify a design with stopping for futility for given α , β , the effect size, δ , one needs to specify the parameters in the futility rule t , ξ , as well as the total maximum sample size per arm, n , for a two-stage design, and similar quantities, t_1 , t_2 , ξ_1 , ξ_2 , n , for a three-stage design.

For each set of input parameters that include the treatment effect, outcome variability, required power and type I error rate, one can run an optimization algorithm to find the optimal design. For a two-stage design, it is natural to minimize the expected sample size under H_0 , EN_0 , computed as $EN_0 = PET_1 n_1 + (1 - PET_1)n$, where PET_1 is the probability of stopping for futility at the interim analysis under the null hypothesis. Simon (1989) tabulated such design for single arm trials with binary outcomes. Simon's optimal designs can be

computed using the online software <http://cancer.unc.edu/biostatistics/program/ivanova/>. Alternatively, one can minimize a weighted average of EN_0 and EN_A , where EN_A is the expected sample size under the alternative hypothesis. This is similar to the approach of Jung et al. (2004) though the weighted average of EN_0 and n was minimized there. Mander et al. (2012) described the process of finding optimal designs for two-stage two-arm trials with continuous outcomes. Using a process similar to Mander et al. (2012), we found optimal designs for two-stage two-arm trials with continuous outcomes with non-binding futility stopping. The designs were obtained by running an optimization program in R (R Core Team 2013) for a given set of parameters. The optimal parameters were obtained for one-sided $\alpha = 0.025$ and $\alpha = 0.05$, and for power = 80% and 90% and are plotted in Figure 1 [Figure 1 near here] for a range of effect sizes. Everywhere in this paper we assume that the response is immediate or the trial is paused to wait for results of the interim analysis. Optimal designs will be different if this is not the case. Interestingly, all four optimal quantities are approximately constant with effect size. This allowed us to formulate futility stopping rules in a two-stage design for given α and β that are close to being optimal for any effect size and outcome standard deviation σ by computing the median values across the range of values (Table 2). [Table 2 near here].

From Table 2, the optimal timing for stopping for futility is between 0.30–0.39 of the total sample size when EN_0 is minimized and between 0.41–0.49 when $0.5EN_0 + 0.5EN_A$ is minimized. The probability of stopping for futility under H_A is rather high for futility rules that minimize EN_0 , 0.134 for trials with 80% power and 0.025 type I error rate. The corresponding probability is 0.073 for the rule minimizing $0.5EN_0 + 0.5EN_A$.

3.2 Optimal three-stage designs

Optimizing for more than one interim analysis in a two-arm trial with continuous outcomes is rather computationally intensive. To obtain approximately optimal three-stage designs we use an observation that optimal information fractions and probability of stopping under H_A for given α and power are very similar for a single-arm trial with binary outcome and a two-arm trial with continuous outcome. The optimal two-stage t and ξ for the design minimizing EN_0 in Table 2 are very similar to those in the Simon's optimal design with large enough sample size to eliminate the effect of discreteness in the optimal design. We analyzed optimal three-stage designs tabulated in Table II of Chen (1997). Chen's optimal three-stage designs minimize $EN_0 = PET_1n_1 + PET_2n_2 + (1 - PET_1 - PET_2)n$. Here PET_2 is the probability of not stopping for futility in the first interim analysis and stopping for futility at the second interim analysis. The optimal information fraction and the probability of stopping for futility under H_A did not vary substantially for different values of π_0 and π_1 . For one-sided $\alpha = 0.05$, the medians, across a range of π_0 and π_1 , of the optimal values of t_1 , t_2 , ξ_1 and ξ_2 for the three-stage design for 80% power were $t_1 = 0.22$, $t_2 = 0.51$, $\xi_1 = 0.11$ and $\xi_2 = 0.05$. For 90% power the approximately optimal values were $t_1 = 0.27$, $t_2 = 0.55$, $\xi_1 = 0.05$ and $\xi_2 = 0.03$.

To answer the question about the advantage of having two futility looks versus one look, we computed n_0 for a single stage design, as well as n , EN_0 and other quantities of interest for the optimal two- and three-stage designs for an effect size of 0.3, one-sided $\alpha = 0.05$, and

power 80% and 90% (Table 3). Comparing the two- and three- stage approximately optimal designs with a single stage design, adding a futility rule to a single stage design with 80% (90%) power, leads to a maximum total sample size increase of about 22% (18%). Adding the second futility look leads to a 43% (33%) increase in the sample size compared to a single stage trial to maintain the same power (Table 3). With a single futility look, the expected sample size under the null hypothesis, EN_0 , is 69% (70%) of the total sample size required for a single stage design. With two futility looks EN_0 is 62% (63%) of the total sample size of a single stage design. These values of EN_0 assume the response was observed immediately or that the enrollment is paused to wait for the futility analysis to be completed.

4. Simulation study

The goal of the simulation study is to illustrate the difference between futility Rule 2 and Rule 3. In a single stage design without futility stopping, sample size of $n_0 = 138$ in each arm yields 80% power with one-sided type I error rate of 0.05 if the effect size is 0.3. The optimal two-stage design parameters are given in Tables 2 and 3. In a two-stage design we applied futility stopping after $n_1 = 59$ subjects were enrolled in each arm, $t = 59/169 = 0.35$. For Rule 3 the probability of stopping for futility under the alternative hypothesis was set to 0.120 for the two-stage design. The probability of stopping for futility under the alternative hypothesis of 0.120 yields the probability of stopping for futility of 0.673 under H_0 . This is the probability that was used to compute the cutoff for Rule 2. Simulation results with 10000 runs for each design are presented in Table 4 [Table 4 near here].

The futility Rule 2 is based on the test statistic Z testing the null hypothesis $H_0: \mu_Y - \mu_X = 0$. Hence, it always yields a given probability of stopping for futility under H_0 , 0.67 for this two-stage design. Rule 3 will always yield a given probability of stopping for futility under the alternative hypothesis $\mu_Y - \mu_X = \delta$, 0.12 for this two-stage design. When both rules are simulated with the effect size δ/σ that gives required power of 80% with $n = 169$ subjects per arm, as in the column titled $\sigma = 1$ (Table 4), the two rules yield the same probability of stopping under the null and alternative hypothesis.

One of the two probabilities of stopping for futility, under either H_0 or H_A , changes when the true variability of outcome, σ^2 , is not correctly specified at the time the study is designed, and $n = 169$ subjects per arm no longer yield required power. If the true $\sigma = 2$ yielding the effect size is $\delta/\sigma = 0.15$, the proposed sample size is not large enough to yield a significant result at the end of the trial 80% of the time. When $\sigma = 0.5$ yielding the effect size of $\delta/\sigma = 0.6$, the resulting power is much higher than the nominal value of 80% with $n = 169$. Looking at columns for $\sigma = 2$ and $\sigma = 0.5$ in Table 4, the probability of stopping for futility in Rule 2 under $H_A: \mu_Y - \mu_X = \delta > 0$ is equal to 0.0023 when $\sigma = 0.5$, much lower than 0.12 when $\sigma = 1$. It is equal to 0.35, if the variance is larger than expected such as in the column where $\sigma = 2$. Higher probability of stopping for futility under H_A might be desirable when the observed effect size is not adequate and, hence, we are unlikely to show significance at the end of the trial. In other circumstances, we might want to maintain a given probability of stopping for futility for a certain treatment effect of the drug, as in Rule 3. That is, we will stop a trial for futility if the trial is unlikely that the true treatment effect is equal to or higher than δ .

Another consideration for choosing between Rule 2 and Rule 3 is that Rule 3 might be harder to implement. He et al. (2012) give recommendations on constructing Z_t^F for a time-to-event type outcome and for non-parametric tests. If it is desired to use the test statistic Z_t instead of Z_t^F in Rule 3, this can be achieved by using the estimated cutoff $c(\xi, \hat{\sigma}) = \frac{\delta}{\sqrt{2\hat{\sigma}^2/(tn)}} + \Phi^{-1}(\xi)$. This is because $\xi = \Pr(Z_t < c(\xi) \mid H_A) = \Pr\left(\frac{\bar{Y}_t - \bar{X}_t - \delta}{\sqrt{2\sigma^2/(tn)}} < c(\xi) - \frac{\delta}{\sqrt{2\sigma^2/(tn)}} \mid H_A\right) = \Phi\left(c(\xi) - \frac{\delta}{\sqrt{2\sigma^2/(tn)}}\right)$. Here δ is fixed (not estimated), and it represents the pre-specified treatment effect in H_A .

5. Futility testing in a crossover study

In a crossover trial, the required sample size and power depend not only on the treatment effect and the variability of the outcome, but also on the within-subject correlation, ρ . For example, in a two-period crossover where a subject receives an active treatment followed by placebo or vice versa, the variance of the treatment difference is equal to $2\sigma^2(1 - \rho)$. Hence, the relationship (1) changes to

$$\delta/(\sqrt{2(1 - \rho)}\sigma) = \sqrt{2/n_0}(Z_\beta + Z_\alpha), \quad (2)$$

where n_0 is the total number of subjects in a crossover study, each contributing two observations on placebo and active treatment. Crossover studies with high within-subject correlation require markedly less subjects compared to crossover studies with low correlation. If the study were powered for a given variance of the treatment effect of $2\sigma^2(1 - \rho)$ but the true within-subject correlation is much higher, the probability of stopping for futility will be affected because the sample size in the trial does not give us required power, e.g. we no longer have the relationship in (1) in a trial with no futility stopping. This, in turn, affects the probability to stop for futility under H_A if Rule 2 is used.

Recall that in a parallel group setting we can maintain a certain probability of stopping for futility given the effect size the trial was powered for, δ/σ , (Rule 2) or for a given treatment effect, δ , (Rule 3). In a crossover trial, power depends on the treatment effect, δ , effect size, δ/σ , and the within-subject correlation ρ . If we use Rule 2, we will maintain a certain probability of stopping for futility given δ/σ the effect size and ρ the trial is powered for. Rule 3 can be set-up to yield a certain probability of stopping for futility for a given treatment effect, a rule based on testing $\mu_Y - \mu_X = \delta > 0$. Alternatively, Rule 3 can be modified to test the hypothesis $(\mu_Y - \mu_X)\sigma = \delta > 0$.

This investigation was initiated when designing the PrecISE study ([ct.gov](https://clinicaltrials.gov/ct2/show/study/NCT04129931) number is [NCT04129931](https://clinicaltrials.gov/ct2/show/study/NCT04129931)). This clinical trial is a platform trial to test the efficacy of several novel interventions in patients with severe asthma against placebo. The study uses a multiple period crossover design. In the PrecISE study, there is an interim analysis for futility to potentially stop ineffective therapies early; this interim analysis uses Rule 3. Three equally important primary endpoints are considered. The study is stopped for futility if there is no effect with respect to all three endpoints. For two endpoints with well-established clinically

meaningful effects ($= \delta$), the futility decision rule is based on testing $\mu_Y - \mu_X = \delta$. For the third endpoint that is a novel endpoint without a well-established clinically meaningful effect, the futility decision rule is based on testing $(\mu_Y - \mu_X)\sigma = \delta$ (Ivanova et al., 2020). Futility stopping in the PrecISE study is not binding. The PrecISE Data and Safety Monitoring Board decides whether or not to discontinue an intervention for futility based on efficacy and safety of the intervention in question.

6. Conclusions

We investigated three stopping rules for futility. Futility stopping based on conditional power (Rule 1) and futility based on testing $\mu_Y - \mu_X = 0$ (Rule 2) are equivalent. These two rules ensure a given probability of stopping for futility under H_0 . The third futility rule is based on testing $\mu_Y - \mu_X = \delta$ and yields a desired probability of stopping for futility under H_A . All three rules behave similarly when the sample size in the trial yields the required power, and either one can be used. If the variance of the outcome is substantially different from hypothesized, to select Rule 1 or 2 versus Rule 3 for stopping for futility, one needs to decide if they are looking to stop for futility when it is unlikely to see a significant result at the end of the trial, that is to see the effect size as specified, (Rules 1 and 2) or when it is unlikely that the treatment effect is as high as hypothesized (Rule 3). In a crossover trial, Rules 1 and 2 have a desired probability to stop for futility given a hypothesized effect size and the within-subject correlation. If one is looking to have a given probability of stopping for futility for a given effect size, Rule 3 needs to be used. Rule 3 can be set to stop for futility when it is unlikely that the effect size is as high as hypothesized by testing $\mu_Y - \mu_X = \delta$ or to stop for futility when the treatment effect is as high hypothesized by testing $(\mu_Y - \mu_X)\sigma = \delta$, correspondingly.

We describe approximately optimal futility rules. The design parameters for approximately optimal futility rules are given in Table 2 for four combinations of the type I error probability and power. These parameters are close to optimal for a wide variety of effect sizes and treatment effects. Note that this optimization was performed for trials where subject's response to treatment is observed immediately. When a subject's response to treatment is delayed, optimal design parameters will be different. Such an investigation is outside of the scope of this manuscript.

Acknowledgements

Ivanova's work was supported in part by the NIH grant P01 CA142538 and by the NIH grant U24 HL138998. The authors thank anonymous reviewers for their helpful comments.

Appendix I.: Cutoff values for Z-based and ZF-based three-stage trials

Under H_A , the Z-based test statistics at the first (Z_{t_1}) and second analyses (Z_{t_2}) jointly follow a bivariate normal distribution with mean of $(\Delta\sqrt{nt_1/2}, \Delta\sqrt{nt_2/2})^T$, variance of 1 and a covariance of $\sqrt{t_1/t_2}$ (Jennison and Turnbull 1999). If the desired probabilities to stop under the H_A are ξ_1 at the first interim analysis, and ξ_2 at the second interim analysis given the trial

does not stop at the first interim analysis, it can be shown that the cutoffs are given by $c_2(\xi_1) = \Phi^{-1}(\xi_1) + \frac{\delta/\sigma}{\sqrt{2/(t_1n)}} (H_A)$, and $c_2(\xi_2)$ satisfies

$$\begin{aligned} & \Pr(Z_{t_1} > c_2(\xi_1), Z_{t_2} < c_2(\xi_2) | H_A) \\ &= \Pr(Z_{t_2} < c_2(\xi_2) | H_A) - \Pr(Z_{t_1} < c_2(\xi_1), Z_{t_2} < c_2(\xi_2) | H_A) \\ &= \xi_2(1 - \xi_1). \end{aligned}$$

Similarly, the Z^F -based test statistics at the first ($Z_{t_1}^F$) and second analyses ($Z_{t_2}^F$) jointly follow a bivariate normal distribution with mean of $(0, 0)^T$, variance of 1 and covariance of $\sqrt{t_1/t_2}$. The cutoff values for Z^F -based three-stage trials are given by $c^F_2(\xi_1) = \Phi^{-1}(\xi_1)$, and $c^F_2(\xi_2)$ satisfies

$$\begin{aligned} & \Pr(Z_{t_1}^F > c^F_2(\xi_1), Z_{t_2}^F < c^F_2(\xi_2) | H_A) \\ &= \Phi_1(c^F_2(\xi_2)) - \Pr(Z_{t_1}^F < c^F_2(\xi_1), Z_{t_2}^F < c^F_2(\xi_2) | H_A) \\ &= \xi_2(1 - \xi_1). \end{aligned}$$

Appendix II.: R code for obtaining cutoff values for Z-based and ZF-based three-stage trials

```
# c1 = cutoff for the first interim
# c2 = cutoff for the second interim
# t1 = information fraction at the first interim
# t2 = information fraction at the second interim
# ksi1 = target probability to stop under the H_A at the first interim
# ksi2 = target probability to stop under the H_A at the second interim
# alpha = type I error
# beta = type II error
## Z-based
library(mnormt)
c1 = qnorm(ksi1) + sqrt(t1)*(qnorm(1-beta) + qnorm(1-alpha))
func.2 = function(c1, c2){
```

```
pnorm(c2, mean=Delta*sqrt(n*t2/2)) - pnorm(c(c1,c2),
mean=Delta*sqrt(c(n*t1, n*t2)/2), varcov = matrix(c(1, sqrt(t1/t2), sqrt(t1/
t2), 1), 2,2))-ksi2*(1-ksi1)
}
c2 = uniroot(func.2, c(-1.5,1.5), c1 = c1)
## ZF-based
c1 = qnorm(ksi1)
func.3 = function(c1, c2){
pnorm(c2) - pnorm(c(c1,c2), varcov = matrix(c(1, sqrt(t1/t2), sqrt(t1/t2),
1), 2,2))-ksi2*(1-ksi1)
}
lower = -3; upper = 3; # initial values
c2 = uniroot(func.3, c(-3,3), c1 = c1)
```

References

- Chen TT 1997. Optimal three-stage designs for phase II cancer clinical trials. *Statistics in Medicine*. 16: 2701–2711. [PubMed: 9421870]
- Fleming T, Harrington D, and O'Brien PC. 1984. Designs for group sequential tests. *Controlled Clinical Trials* 5 (4):348–361. doi:10.1016/s0197-2456(84)80014-8. [PubMed: 6518769]
- He P, Lai TL, and Liao OY. 2012. Futility stopping in clinical trials. *Statistics and Its Interface* 5 (4):415–423. doi: 10.4310/SII.2012.v5.n4.a4.
- Herring WJ, Liu K, Hutzelmann J, Snaveley D, Snyder E, Ceesay P, Lines C, Michelson D, and Roth T. 2013. Alertness and psychomotor performance effects of the histamine-3 inverse agonist MK-0249 in obstructive sleep apnea patients on continuous positive airway pressure therapy with excessive daytime sleepiness: a randomized adaptive crossover study. *Sleep Medicine* 14 (10):955–963. doi: 10.1016/j.sleep.2013.04.010. [PubMed: 23920422]
- Ivanova A, Liu K, Snyder E, and Snaveley D. 2009. An adaptive design for identifying the dose with the best efficacy/tolerability profile with application to a crossover dose-finding study. *Statistics in Medicine* 28:2941–2951. doi:10.1002/sim.3684. [PubMed: 19731265]
- Ivanova A, Paul B, Marchenko O, Song G, Patel N, and Moschos SJ. 2016. Nine-year change in statistical design, profile, and success rates of phase II oncology trials. *Journal of Biopharmaceutical Statistics* 26 (1):141–149. doi:10.1080/10543406.2015.1092030. [PubMed: 26368744]
- Jennison C, Turnbull BW. 1999. *Group sequential methods with applications to clinical trials*. Chapman & Hall/CRC.
- Jung SH, Lee T, Kim K, and George SL. 2004. Admissible two-stage designs for phase II cancer clinical trials. *Statistics in Medicine* 23 (4):561–569. [PubMed: 14755389]
- Lan KKG, Simon R, and Halperin M. 1982. Stochastically curtailed tests in long-term clinical trials, *Communications in Statistics. Part C: Sequential Analysis* 1 (3):207–219. doi: 10.1080/07474948208836014.
- Lan KKG, and Wittes J. 1988. The B-value: A tool for monitoring data. *Biometrics* 44 (2):579–585. doi:10.2307/2531870. [PubMed: 3390511]
- Mander AP, and Thompson SG. 2010. Two-stage designs optimal under the alternative hypothesis for phase II cancer clinical trials. *Contemporary Clinical Trials* 31 (6):572–578. doi: 10.1016/j.cct.2010.07.008. [PubMed: 20678585]
- Mander AP, Wason JM, Sweeting MJ, and Thompson SG. 2012. Admissible two-stage designs for phase II cancer clinical trials that incorporate the expected sample size under the alternative hypothesis. *Pharmaceutical Statistics* 11:91–96. doi:10.1002/pst.501. [PubMed: 22232071]
- R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. <http://www.R-project.org/>.

- Simon R 1989. Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials* 10 (1):1–10. doi:10.1016/0197-2456(89)90015-9. [PubMed: 2702835]
- Snapinn S, Chen M, Jiang Q, and Koutsoukos T. 2006. Assessment of futility in clinical trials. *Pharmaceutical Statistics*. 5: 273–281. [PubMed: 17128426]
- Wason JMS and Mander AP. 2012. Minimizing the maximum expected sample size in two-stage phase II clinical trials with continuous outcomes. *Journal of Biopharmaceutical Statistics* 22 (4):836–852. doi: 10.1080/10543406.2010.528104. [PubMed: 22651118]
- Wiener LE, Ivanova A, and Koch G. Forthcoming. Methods for clarifying criteria for study continuation at interim analysis. *Pharmaceutical Statistics*.

Figure 1.

Optimal two-stage designs with an interim analysis for futility to minimize EN_0 or $0.5EN_0 + 0.5EN_A$. Maximum sample size with futility over sample size without futility stopping (n/n_0), information fraction (t) for the futility look, probability to stop for futility under H_A (ξ) and probability to stop for futility under H_0 (γ) for different combinations of one-sided type I error and power, (0.025, 90%) shown by solid lines, (0.05, 90%) by dashed lines, (0.025, 80%) by dotted lines and (0.05, 80%) by dot-dash lines).

Table 1:

The three decision rules and corresponding two-stage cutoff values yielding probability of γ to stop under the null hypothesis or probability of ξ to stop under the alternative hypothesis $\mu_Y - \mu_X = \delta > 0$. A trial is stopped for futility at information fraction t if the value of the statistic is less than a corresponding cutoff value.

Rule	Statistic	Cutoff value
Rule 1, conditional power	$\Phi \left\{ \sqrt{\frac{1}{1-t}} \left(\frac{\sqrt{n}}{\sqrt{2}\hat{\sigma}} (\bar{Y}_t - \bar{X}_t) - Z_\alpha \right) \right\}$	$\Phi \left\{ \frac{1}{\sqrt{t(1-t)}} (Z_1 - \gamma - \sqrt{t} Z_\alpha) \right\}$ (for H_0) $\Phi \left\{ \frac{1}{\sqrt{t(1-t)}} (Z_1 - \xi - \sqrt{t} Z_\alpha + \frac{\delta/\sigma}{\sqrt{2/(tn)}}) \right\}$ (for H_A)
Rule 2, Z -based	$\frac{\bar{Y}_t - \bar{X}_t}{\hat{\sigma}\sqrt{2/(tn)}}$	$\Phi^{-1}(\gamma)$ (for H_0) $\Phi^{-1}(\xi) + \frac{\delta/\sigma}{\sqrt{2/(tn)}}$ (for H_A)
Rule 3, Z^F -based	$\frac{\bar{Y}_t - \bar{X}_t - \delta}{\hat{\sigma}\sqrt{2/(tn)}}$	$\Phi^{-1}(\xi)$ (for H_A)

Table 2.

Approximately optimal two-stage designs with an interim analysis for futility to minimize EN_θ or $0.5EN_\theta + 0.5EN_A$ for four combinations of one-sided type I error, α , and power.

Minimizing	EN_θ				$0.5EN_\theta + 0.5EN_A$			
Type I error	$\alpha = 0.025$		$\alpha = 0.05$		$\alpha = 0.025$		$\alpha = 0.05$	
Power	80%	90%	80%	90%	80%	90%	80%	90%
n/n_0	1.277	1.217	1.228	1.182	1.071	1.050	1.058	1.043
Information fraction (t)	0.305	0.349	0.347	0.388	0.410	0.449	0.455	0.489
Probability to stop under H_A (ξ)	0.134	0.066	0.120	0.059	0.073	0.032	0.066	0.030
Probability to stop under H_0 (γ)	0.741	0.729	0.673	0.664	0.658	0.647	0.588	0.581

Table 3:

The sample size per arm in a single stage trial without consideration of futility stopping (n_0), the maximum sample size per arm considering futility (n), and the expected sample size per arm under H_0 , EN_0 , to achieve power of 80% and 90% when the effect size is 0.3 in a two- and three-stage design with a futility stopping.

Power	Single stage design	Two-stage design	Three-stage design
80%	$n_0 = 138$ $EN_0 = 138$	$n = 169$ $EN_0 = 94.8$	$n = 198$ $EN_0 = 85.8$
90%	$n_0 = 191$ $EN_0 = 191$	$n = 226$ $EN_0 = 133.9$	$n = 254$ $EN_0 = 120.4$

Table 4:

The probabilities to stop for futility under the null ($\delta = 0$) or alternative hypotheses ($\delta = 0.3$) for various values of σ yielding various effect sizes () for a two-stage design. The value $\sigma = 1$ is as hypothesized when planning the study and yields the hypothesized effect size of $\delta = 0.3$. The second column shows cutoffs for decision rules.

Futility Rule	Cutoff	$\sigma = 1$		$\sigma = 2$		$\sigma = 0.5$	
		Alternative	Null	Alternative	Null	Alternative	Null
		$\delta = 0.3$ $= 0.3$	$\delta = 0$ $= 0$	$\delta = 0.3$ $= 0.15$	$\delta = 0$ $= 0$	$\delta = 0.3$ $= 0.6$	$\delta = 0$ $= 0$
Rule 2	0.454	0.12	0.67	0.35	0.67	0.0023	0.67
Rule 3	-1.174	0.12	0.67	0.12	0.35	0.12	0.98