STOPPING RANDOMIZED TRIALS EARLY FOR FUTILITY: ERROR IN ESTIMATING THE TREATMENT EFFECT

BY

HEDAN HAN

A THESIS

SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

McMaster University

Master of Science (2014) McMaster University

(Statistics) Hamilton, Ontario

TITLE: Stopping Randomized Trials Early for Futility: Error in Estimating the Treatment Effect

AUTHOR: Hedan Han

SUPERVISORS: Dr. Stephen Walter

NUMBER OF PAGES: xviii, 108

Abstract

In the past few decades, interim analyses in long-term clinical trials have become increasingly popular. If a new treatment or intervention is truly ineffective or have lower effect than expected, it is better that the trial be stopped early for futility so as not to expose further patients to clinically inefficient or harmful treatment, and thus save time, resources and money by avoiding recruitment of more patients. Conditional power is one of the typical methods of futility monitoring.

In this paper, I develop expressions for the error in estimating the treatment effect in studies stopped early for futility, the over-estimations in completed studies and the overall bias for a study with a stopping rule for futility. I also discuss the probability of stopping and the weight of such stopped studies in a meta-analysis. Then I evaluate the expressions using commonly used parameters. The numerical results show that the error in estimating the treatment effect in studies stopped early for futility is usually substantial but fortunately the overall bias in single study is quite small. Corresponding biases are a little bit greater if two interim analyses are planned instead of only one interim analysis, thus it is expected that the under-estimation of treatment

effect in truncated studies and the overall bias of a single study would increase with the number of planned interim analyses.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Stephen Walter for his invaluable guidance, immense knowledge, constant encouragement, endless support and incredible patience. I appreciate very much everything he has done in the course of supervision of this thesis.

Special thanks to my committee members, Dr. Peter D. M. Macdonald and Dr. Román Viveros-Aguilera for serving on my thesis committee and offering valuable advice to me.

Finally, I would like to thank all of my friends for their love and help.

Thanks also to my parents, my husband and my kids for their understanding and support on this wonderful journey.

Contents

Appendix Tables **62**

List of Figures

List of Symbols and Abbreviations

Symbols

$CP_k(.)$	The conditional power at the k^{th} analysis
θ_F	The hypothesized true treatment effect for the future data
Z_k	The test statistic at the k^{th} analysis (k = 1,2, …, K)
$\hat{\theta}_k$	The observed treatment effect at the k^{th} interim analysis
$SE(\hat{\theta}_k)$	The standard error of $\hat{\theta}_k$
I_k	Information about the parameter $\hat{\theta}_k$, $I_k = \mathrm{SE}(\hat{\theta}_k)^{-2}$, ($k$ = 1,2, …, K)
t_k	The information time at the k^{th} analysis, $t_k = I_k/I_K$, $(0 < t_k \leq t_K = 1)$
$\mathrm{Z}(.)$	The probability density function of the standard normal distribution
$\Phi(.)$	The cumulative standard normal distribution function
α	The probability of type I error
β	The probability of type II error
γ	Threshold value for the conditional power
θ_A	The minimally clinically meaningful treatment effect
$c_{1-\alpha}$	$1-\alpha$ quantile of the standard normal distribution
$c_{1-\gamma}$	$1-\gamma$ quantile of the standard normal distribution

μ	The true mean difference in outcomes between the two groups
μ_k	The expected mean treatment effect for the accumulated data in the k^{th} analysis (k =1, 2)
δ_k	The treatment effect at the k^{th} interim analysis (k =1, 2)
$\hat{\delta}_k$	The treatment effect observed at the k^{th} interim analysis (k =1, 2)
δ_k^*	The threshold treatment effect at the k^{th} interim analysis (k =1, 2)
σ	Standard deviation of the outcome variable
$\sigma_{\delta k}$	The standard deviation of the mean response in each group at the k^{th} interim analysis (k =1, 2)
N	Planned sample size
n	Sample size per group per analysis
p	The probability of stopping at the interim analysis with one planned interim analysis
$E(.)$	Expected mean value
Z_{Hk}^*	The upper threshold for Z_k at the k^{th} interim analysis (k =1, 2)
Z_{Lk}^*	The lower threshold for Z_k at the k^{th} interim analysis (k =1, 2)
δ_{Hk}^*	The upper threshold for treatment effect at the k^{th} interim analysis (k =1, 2)
δ_{Lk}^*	The lower threshold for treatment effect at the k^{th} interim analysis (k =1, 2)
$\hat{\delta}_{interim}$	The treatment effect at the interim analysis

Abbreviations

CP	Conditional power
HR	Hazard ratio
HRPC	Hormone-refractory prostate cancer
RCTs	Randomized controlled trials
tRCTs	Truncated randomized controlled trials
ATG	Antithymocyte globulin
Cy	Cyclophosphamide
DSMB	Data safety and monitoring board
SPIRIT	Standard protocol items recommendations for interventional trials
CONSORT	Consolidated standards of reporting trials

Chapter 1

Introduction

1.1 Background

"Clinical trials are unique in that enrollment of patients is a continual process staggered in time. This means that if a treatment is particularly beneficial or harmful compared to the concurrent placebo group while the study is on-going, the investigators are ethically obliged to assess that difference using the data at hand and to make a deliberate consideration of terminating the study earlier than planned."[1] In the past few decades, interim analyses in long-term clinical trials have become increasingly popular.

Stopping trials early for futility has been discussed early in the 1980s (Halperin et al. 1982[2]; Lan et al. 1982[3]; Lan and Wittes,1988[4]), and the work in this area is still ongoing. It is often appropriate for the trial protocol to include a formal sequential monitoring plan. If a new treatment or intervention is truly ineffective or has lower effect than expected, it is worth planning an interim analysis and thinking about early stopping for futility. When early stopping for futility is taken into account, then if there is early extreme evidence of little probability for the treatment to show efficacy ultimately, early closure of the trial should be considered, and the investigators may claim that no statistically significant treatment benefit has been demonstrated and publish the result.

Early termination of a trial has many obvious appeals, including not exposing further patients to clinically inefficient or even harmful treatments, and potentially saving a significant amount of time, resources and money invested in the trials by avoiding the recruitment of more patients.

1.2 Problem of Interest

In the following chapters we discuss the question of estimating the magnitude of treatment effect in trials with a stopping rule for futility, and analyze the magnitude of bias inducing by adopting the rule. To the best of my knowledge, no one has focused on this question before, though there is much literature in this area concerning stopping rules for futility and conditional power.

Chapter 2 will first introduce some basic terminologies to be used, including endpoint, truncated normal distribution and conditional power.

Chapter 3 will develop expressions for under- and over-estimation in a study with a stopping rule for futility, and the overall bias for a single study (stopped or completed). This will require some theory about truncated normal distributions, which is introduced in Chapter 2. The probability of stopping and the weight of such stopped study in a meta-analysis are then discussed.

Chapter 4 analytically describes the results by substituting numeric parameters into the formulae developed in Chapter 3. Chapter 5 presents two real examples. And the final Chapter 6 summarizes the key findings from previous chapters and discusses their limitations.

Most of the following chapters are developed in a similar way to an unpublished paper by Walter et al.[5] (I have obtained oral permission from Walter through personal corresponding).

Chapter 2

Terminology

2.1 Clinical Endpoint (or Study Outcome)

"In a clinical research trial, a clinical endpoint generally refers to occurrence of a disease, symptom, sign or laboratory abnormality that constitutes one of the target outcomes of the trial, but may also refer to any such disease or sign that strongly motivates the withdrawal of that individual or entity from the trial, and then often termed humane (clinical) endpoint." [6]

The primary endpoint is important to the study design because it is used to calculate the sample size to provide the study enough power to detect the expected effect.

There are many types of endpoints. But statistically, there are three main types, binomial, survival and normal endpoints.

For analytic simplicity, this paper will assume the study outcome is normally distributed. Other types of endpoints can be easily considered approximately by converting to normal equivalent endpoints; for instance, see Chinn 2000[7] for details of converting binomial endpoints to normal ones.

2.2 Truncated normal distribution

A key analytic expression needed in this paper is the expected mean of a truncated normal distribution.

The following description of truncated normal distribution is excerpted from Wikipedia[8].

According to Wikipedia, "in probability and statistics, the truncated normal distribution is the probability distribution of a normally distributed random variable whose value is either bounded below or above (or both)."

Suppose a random variable X is normally distributed as $X \sim N(\mu, \sigma^2)$ where μ is the mean and σ^2 is the variance. Also suppose X lies in the interval $X \in (a, b), -\infty < a < b < +\infty$ where a and b are constants. Then X conditional on $a < X < b$ is said to have a truncated normal distribution. Its probability density function, f, for $a < X < b$, is given by

$$\mathrm{f}(X;\ \mu, \sigma, a, b) = \frac{\frac{1}{\sigma}\mathrm{Z}\left(\frac{X-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}$$

and by f=0 otherwise. Here $Z(.)$ is the probability density function of the standard normal distribution and $\Phi(.)$ is its cumulative distribution function.

For the two-sided truncation, the expected value is:

$$\mathrm{E}(X|a < X < b) = \mu + \frac{\mathrm{Z}\left(\frac{a-\mu}{\sigma}\right) - \mathrm{Z}\left(\frac{b-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}\sigma \qquad (1)$$

There are two cases of one-sided truncation. One is when X is truncated below, where a is a fixed value and b equals to $+\infty$:

$$\mathrm{E}(X|X>a)=\mu+\frac{\mathrm{Z}\left(\frac{a-\mu}{\sigma}\right)}{1-\Phi\left(\frac{a-\mu}{\sigma}\right)}\sigma \qquad (2)$$

The other case is when X is truncated above, where a equals to $-\infty$ and b is a fixed value:

$$\mathrm{E}(X|X<b)=\mu-\frac{\mathrm{Z}\left(\frac{b-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right)}\sigma \qquad (3)$$

2.3 Typical Method of Futility Monitoring: Conditional Power

As He, Lai and Liao[9] pointed out, currently the methods for futility stopping in the literature are mostly based on conditional power or predictive power, which are in conjunction with the theory of stochastic curtailment or group sequential design. Conditional Power (CP), which was introduced by Lan, Simon and Halperin[3] in 1982, is "the probability that the final study result will be statistically significant, given the data observed thus far and a specific assumption about the pattern of the data to be observed in the remainder of the study, such as assuming the original design effect, or the effect estimated from the current data, or under the null hypothesis"[9]. The approach of assuming the original design effect means calculating CP based on the original alternative hypothesis, i.e. the minimally clinically meaningful treatment effect, this is the most common implementation of stochastic curtailment indicated by

Freidlin[11]. The current trend approach calculates the CP under the current estimate of treatment effect, but as Freidlin said this approach is problematic and may lead to premature stopping [9]. The null effect approach computes CP assuming the null hypothesis for the remaining data, CP computed under this assumption is typically low since the null hypothesis is a very pessimistic assumption.

CP is "an appealing way to summarize the accumulated data to quantify the feasibility of continuing the study to the full accrual" [11] and very lower power means that the trial is futile, thus if CP falls below a fixed threshold (typically 0.10-0.15) indicated by Proschan et al[12], the trial should be considered to be stopped for futility.

The rest of this section will mostly follow the development of Chapter 12. Futility Analysis by Freidlin[11].

Define $CP_k(\theta_F)$ as the CP at the kth analysis as the probability of rejecting the null hypothesis at the final analysis given the currently observed data and a hypothesized true treatment effect θ_F for one-sided testing:

$$CP_k(\theta_F) = P(Z_K > c_{1-\alpha} | Z_k = z_k, \theta_F)$$

where θ_F denotes the assumed treatment effect for the future data, $c_{1-\alpha}$ denotes the 1-α quantile of the standard normal distribution and the quantity $Z_k = \frac{\hat{\theta}_k}{SE(\widehat{\hat{\theta}_k})}$, represents the test statistic at the k^{th} analysis (k=1,2, …, K) where $\hat{\theta}_k$ is the observed treatment effect at the k^{th} interim analysis and $SE(\widehat{\hat{\theta}_k})$ is the standard error of $\hat{\theta}_k$.

At the k^{th} interim analysis, the conditional distribution of Z_K, given $Z_k = z_k$, is normal with mean $z_k\sqrt{t_k} + \theta_F\sqrt{I_K}(1 - t_k)$ and variance $1 - t_k$, where the quantity

$t_k = I_k/I_K\ (0 < t_k \le t_K = 1)$ represents the proportion of the total information and is often referred to as the "information time", I_k represents the information at the k^{th} analysis (k=1,2, …, K), $I_k = \mathrm{SE}(\hat{\theta}_k)^{-2}$.

The conditional power is then

$$
\begin{aligned}
CP_k(\theta_F) &= P(Z_K > c_{1-\alpha} | Z_k = z_k,\ \theta_F) \\
&= \mathrm{P}\left(Z = \frac{Z_K - \left(z_k\sqrt{t_k} + \theta_F\sqrt{I_K}(1-t_k)\right)}{\sqrt{1-t_k}} > \frac{c_{1-\alpha} - \left(z_k\sqrt{t_k} + \theta_F\sqrt{I_K}(1-t_k)\right)}{\sqrt{1-t_k}} \right) \\
&= \mathrm{P}\left(Z > \frac{c_{1-\alpha} - \left(z_k\sqrt{t_k} + \theta_F\sqrt{I_K}(1-t_k)\right)}{\sqrt{1-t_k}} \right) \\
&= 1 - \mathrm{P}\left(Z < \frac{c_{1-\alpha} - \left(z_k\sqrt{t_k} + \theta_F\sqrt{I_K}(1-t_k)\right)}{\sqrt{1-t_k}} \right) \\
&= 1 - \Phi\left(\frac{c_{1-\alpha} - z_k\sqrt{t_k} - \theta_F\sqrt{I_K}(1-t_k)}{\sqrt{1-t_k}} \right)
\end{aligned}
$$

where Z has a standard normal distribution, $\Phi(.)$ is the cumulative standard normal distribution function.

From the condition $CP_k(\theta_F) < \gamma$, we get:

$$CP_k(\theta_F) < \gamma = 1 - \Phi(c_{1-\gamma})$$

$$\Leftrightarrow 1 - \Phi\left(\frac{c_{1-\alpha} - z_k\sqrt{t_k} - \theta_F\sqrt{I_K}(1-t_k)}{\sqrt{1-t_k}} \right) < 1 - \Phi\left(c_{1-\gamma}\right)$$

$$\Leftrightarrow -\Phi\left(\frac{c_{1-\alpha} - z_k\sqrt{t_k} - \theta_F\sqrt{I_K}(1-t_k)}{\sqrt{1-t_k}} \right) < -\Phi\left(c_{1-\gamma}\right)$$

$$\Leftrightarrow \Phi\left(\frac{c_{1-\alpha} - z_k\sqrt{t_k} - \theta_F\sqrt{I_K}(1-t_k)}{\sqrt{1-t_k}} \right) > \Phi\left(c_{1-\gamma}\right)$$

$$\Leftrightarrow \frac{c_{1-\alpha} - z_k\sqrt{t_k} - \theta_F\sqrt{I_K}(1-t_k)}{\sqrt{1-t_k}} > c_{1-\gamma}$$

$$\Leftrightarrow Z_k < \frac{c_{1-\alpha}}{\sqrt{t_k}} - c_{1-\gamma}\frac{\sqrt{1-t_k}}{\sqrt{t_k}} - \theta_F\sqrt{I_K}\frac{(1-t_k)}{\sqrt{t_k}}$$

Hence for one-sided testing, the conditional $CP_k(\theta_F) < \gamma$ corresponds to stopping for futility at the k^{th} analysis if

$$Z_k < Z_k^* \qquad (4)$$

where $Z_k^* = \frac{c_{1-\alpha}}{\sqrt{t_k}} - c_{1-\gamma}\frac{\sqrt{1-t_k}}{\sqrt{t_k}} - \theta_F\sqrt{I_K}\frac{(1-t_k)}{\sqrt{t_k}}$.

For two sided testing, conditional power at the k^{th} analysis is:

$$CP_k(\theta_F) = P(|Z_K| > c_{1-\alpha/2}|Z_k = z_k, \theta_F)$$

$$= P(Z_K > c_{1-\alpha/2}, Z_K < c_{\alpha/2}|Z_k = z_k, \theta_F)$$

For the condition $CP_k(\theta_F) < \gamma$,

$$CP_k(\theta_F) = P(Z_K > c_{1-\alpha/2}, Z_K < c_{\alpha/2}|Z_k = z_k, \theta_F) < \gamma$$

$$\Leftrightarrow\ P\left(Z_K > c_{1-\alpha/2}\middle|Z_k = z_k, \theta_F\right) + \ P\left(Z_K < c_{\alpha/2}\middle|Z_k = z_k, \theta_F\right) < \gamma \qquad (5)$$

Unfortunately, there is no analytic solution to (5), but it can be solved by iteration to get a numeric results for the upper and lower threshold values for Z_k.

By definition, type I error involves rejection of the null hypothesis while type II involves rejection of the alternative hypothesis. Since futility monitoring does not involve rejecting the null hypothesis, it does not inflate type I error but does inflate the type II error. Proschan et al.[12] stated that though stochastic curtailment increases error rates, the magnitude of increase is relatively small even if the data is monitored continuously. Lan et al.[3] have demonstrated that the type II error rate is bounded by $\beta/(1-\gamma)$ irrespective of the number of interim analyses if stochastic curtailment is used to stop for futility when CP falls below γ and it is exactly $\beta/(1-\gamma)$ under continuous monitoring. Notice that if $\gamma > 0.15$, the increase will not be small.

Chapter 3

Statistical methods

We assume in this paper there are two groups in each trial, given an experimental treatment and control group. Group sizes are equally spaced, assuming 1:1 randomization, with a sample size *n* per group per analysis. That is, for the trial with one interim analysis, the available sample size at the interim analysis would be *2n*, which is half of the planned total sample size *N=4n*; for the trial with two planned interim analyses, the available sample size at 1st and 2nd interim analysis would then be one-third and two-thirds of the planned total sample size.

3.1 One Planned Interim Analysis, One-sided Testing

3.1.1 Threshold Treatment Effect and Probability of Stopping

For one interim analysis one-sided testing, according to inequality (4),

$$Z_1 < Z_1^* = \frac{c_{1-\alpha}}{\sqrt{t_1}} - c_{1-\gamma}\frac{\sqrt{1-t_1}}{\sqrt{t_1}} - \theta_F\sqrt{I_K}\frac{(1-t_1)}{\sqrt{t_1}},$$

where $c_{1-\alpha}$ and $c_{1-\gamma}$ are constants which are set in advance. The common value of α can be 5%, while the constant γ should be between 0 and 0.5, and values of 0.1 and 0.2 are recommended by Jennison and Turnbull[13], typical value of γ could be 0.10-0.15 indicated by Proschan et al[12]). The assumed effect size for the future data is defined by θ_F, common choices for its value are the original design effect, or the effect estimated from the current data, or under the null hypothesis. The information for $\hat{\theta}_1$ is $I_1 = var^{-1}(\hat{\theta}_1) = \left(\frac{\sigma^2}{n}+\frac{\sigma^2}{n}\right)^{-1} = \left(\frac{2\sigma^2}{n}\right)^{-1}$ and $I_K = var^{-1}(\hat{\theta}_K) = \left(\frac{\sigma^2}{2n}+\frac{\sigma^2}{2n}\right)^{-1} = \left(\frac{\sigma^2}{n}\right)^{-1}$, so $t_1 = I_1/I_K = \frac{1}{2}$.

Since $Z_1 = \frac{\hat{\delta}_1}{s.e.(\hat{\delta}_1)} \Rightarrow \hat{\delta}_1 = Z_1 \times s.e.(\hat{\delta}_1) = Z_1 \times \sigma_\delta = Z_1 \times \sigma\sqrt{\frac{2}{n}}$, where $\hat{\delta}_1$ is the treatment effect observed at the interim analysis, the mean difference in outcomes between the two groups, and $s.e.(\hat{\delta}_k)$ or $\sigma_{\delta k}$ (k=1,2) is the standard deviation of the mean response in each group at the kth interim analysis, it is a function of the standard deviation of the outcome variable (σ) and the current sample size per group, for example, $\sigma_{\delta 1} = \sqrt{\frac{\sigma^2}{n}+\frac{\sigma^2}{n}} = \sigma\sqrt{\frac{2}{n}}$ at the 1st interim analysis and $\sigma_{\delta 2} = \sqrt{\frac{\sigma^2}{2n}+\frac{\sigma^2}{2n}} = \sigma\sqrt{\frac{1}{n}}$ at the 2nd interim analysis. Without loss of generality, we take $\sigma = 1$ and hence the treatment effect will be expressed in standardized units.

$$\begin{aligned}
\hat{\delta}_1 &= Z_1 \times \sigma\sqrt{\frac{2}{n}} < Z_1^* \times \sigma\sqrt{\frac{2}{n}} \\
&= \left(\frac{c_{1-\alpha}}{\sqrt{t_1}} - c_{1-\gamma}\frac{\sqrt{1-t_1}}{\sqrt{t_1}} - \theta_F\sqrt{I_K}\frac{(1-t_1)}{\sqrt{t_1}}\right) \times \sigma\sqrt{\frac{2}{n}} \\
&= \left(\frac{c_{1-\alpha}}{\sqrt{1/2}} - c_{1-\gamma}\frac{\sqrt{1-1/2}}{\sqrt{1/2}} - \theta_F\sqrt{\frac{n}{\sigma^2}}\frac{(1-1/2)}{\sqrt{1/2}}\right) \times \sigma\sqrt{\frac{2}{n}}
\end{aligned}$$

$$= \sigma\sqrt{\tfrac{2}{n}}\left(\sqrt{2}c_{1-\alpha} - c_{1-\gamma}\right) - \theta_F$$

$$= \delta_1^*$$

Thus, the threshold treatment effect is given by

$$\delta_1^* = \sigma\sqrt{\frac{2}{n}}\left(\sqrt{2}c_{1-\alpha} - c_{1-\gamma}\right) - \theta_F$$

$CP_1(\theta_F) < \gamma$ is equivalent to $Z_1 < Z_1^*$, and also to $\widehat{\delta}_1 < \delta_1^*$. At the interim analysis, if the observed treatment effect $\widehat{\delta}_1$ is less than δ_1^*, the trial will stop, otherwise, it will continue. So the probability of stopping is $\Pr(\text{stop}) = \Pr(CP_1(\theta_F) < \gamma)$,

$$\Pr(CP_1(\theta_F) < \gamma) = \mathrm{P}(Z_1 < Z_1^*) = \mathrm{P}\left(\widehat{\delta}_1 < \delta_1^*\right)$$

$$= \mathrm{P}\left(\mathrm{Z} = \frac{\widehat{\delta}_1 - \mu}{se\left(\widehat{\delta}_1\right)} < \frac{\delta_1^* - \mu}{se\left(\widehat{\delta}_1\right)}\right)$$

$$= \mathrm{P}\left(\mathrm{Z} < \frac{\frac{\sigma}{\sqrt{n}}\left(2c_{1-\alpha} - \sqrt{2}c_{1-\gamma}\right) - \theta_F - \mu}{\sigma\sqrt{2/n},}\right)$$

$$= \mathrm{P}\left(\mathrm{Z} < \frac{\sigma}{\sqrt{n}} \cdot \frac{1}{\sigma}\sqrt{\frac{n}{2}}\left(2c_{1-\alpha} - \sqrt{2}c_{1-\gamma}\right) - \frac{\theta_F + \mu}{\sigma\sqrt{2/n}}\right)$$

$$= \Phi\left(\left(\sqrt{2}c_{1-\alpha} - c_{1-\gamma}\right) - \frac{\theta_F + \mu}{\sigma\sqrt{2/n}}\right)$$

$$= p$$

$$= \begin{cases} \Phi\left(\left(\sqrt{2}c_{1-\alpha} - c_{1-\gamma}\right) - \dfrac{\theta_A + \mu}{\sigma\sqrt{2/n}}\right), \text{under original design effect} \\ \Phi\left(\left(\sqrt{2}c_{1-\alpha} - c_{1-\gamma}\right) - \dfrac{\mu}{\sigma\sqrt{2/n}}\right), \text{under the null hypothesis} \\ \Phi\left(\left(\sqrt{2}c_{1-\alpha} - c_{1-\gamma}\right) - \dfrac{\hat{\delta}_I + \mu}{\sigma\sqrt{2/n}}\right), \text{effect estimated from} \\ \qquad\qquad\qquad\qquad\qquad\qquad\text{the current data} \end{cases}$$

The probability of not stopping is then 1-*p*.

3.1.2 Under-Estimation for the Studies That Stop for Futility

At the interim analysis, among studies whose treatment effect does not exceed the threshold value and hence stop early for futility, the sample treatment effect has a Normal distribution truncated above at δ_1^*, so from (3) the expected mean value of such studies is

$$\mathrm{E}\left(\hat{\delta}_1 \middle| \text{stop for futility}\right) = \mathrm{E}\left(\hat{\delta}_1 \middle| \hat{\delta}_1 < \delta_1^*\right) = \mu - \frac{\mathrm{Z}\left(\frac{\delta_1^* - \mu}{\sigma_{\delta 1}}\right)}{\Phi\left(\frac{\delta_1^* - \mu}{\sigma_{\delta 1}}\right)} \sigma_{\delta 1}$$

$$= \mu - \frac{\mathrm{Z}\left(\theta_{Futility1}\right)}{\Phi\left(\theta_{Futility1}\right)} \sigma_{\delta 1} \qquad (6)$$

where $\theta_{Futility1} = \frac{\delta_1^* - \mu}{se(\hat{\delta}_1)}$, μ is the true treatment effect and $\sigma_{\delta 1}$ denotes the standard error of the estimated treatment effect $s.e.(\hat{\delta}_1)$ at the interim analysis, $\sigma_{\delta 1} = \sigma\sqrt{2/n}$.

Notice that $p = \Phi\left(\frac{\delta_1^* - \mu}{\sigma_{\delta 1}}\right) = \Phi\left(\theta_{Futility1}\right)$ and $1 - p = 1 - \Phi\left(\theta_{Futility1}\right)$.

The second term of (5), $\frac{\mathrm{Z}(\theta_{Futility1})}{\Phi(\theta_{Futility1})}\sigma_{\delta 1}$ gives the expected under-estimation in a study that stops at the interim analysis.

3.1.3 Over-Estimation for the Studies That Do Not Stop for Futility

On the other hand, studies whose effect size exceed the threshold value and hence continue have outcomes at the interim analysis with a Normal distribution truncated below at δ_1^*. Hence at the interim analysis they have an expected mean value

$$\mathrm{E}\left(\hat{\delta}_1 \middle| \text{not stop for futility}\right) = \mathrm{E}\left(\hat{\delta}_1 \middle| \hat{\delta}_1 > \delta_1^*\right)$$

$$= \mu + \frac{\mathrm{Z}\left(\frac{\delta_1^* - \mu}{\sigma_{\delta 1}}\right)}{1 - \Phi\left(\frac{\delta_1^* - \mu}{\sigma_{\delta 1}}\right)} \sigma_{\delta 1}$$

$$= \mu + \frac{\mathrm{Z}\left(\theta_{Futility1}\right)}{1 - \Phi\left(\theta_{Futility1}\right)} \sigma_{\delta 1} \qquad (7)$$

Similarly, the second term of (7) shows these studies over-estimate the treatment effect.

Since the probability of stopping at the interim analysis is p, and not stopping is 1-p, we can check that the weighted average of (6) and (7) is unbiased for μ as expected,

$$p \times (\text{equation } (6)) + (1 - p) \times (\text{equation } (7))$$

$$= p \times \left(\mu - \frac{\mathrm{Z}\left(\theta_{Futility1}\right)}{\Phi\left(\theta_{Futility1}\right)} \sigma_{\delta 1}\right) + (1 - p) \times \left(\mu + \frac{\mathrm{Z}\left(\theta_{Futility1}\right)}{1 - \Phi\left(\theta_{Futility1}\right)} \sigma_{\delta 1}\right)$$

$$= p \times \left(\mu - \frac{\mathrm{Z}\left(\theta_{Futility1}\right)}{p} s\sigma_{\delta 1}\right) + (1 - p) \times \left(\mu + \frac{\mathrm{Z}\left(\theta_{Futility1}\right)}{1 - p} \sigma_{\delta 1}\right)$$

$$= \mu.$$

3.1.4 Overall Bias for a Single Study

A study that does not stop at the interim analysis thus continues after the interim analysis accumulates a further $2n$ observations and complete its target sample size N=4n. The expected mean treatment effect from the completed study is a weighted combination of (7) and μ, since the data gathered after the interim analysis is unaffected by the stopping rule. The two parts of the data are equally weighted, so the expected mean in a study with a stopping rule but which continues to its final analysis is,

$$E\left(\widehat{\delta}_1 \middle| \text{completed}\right) = \frac{2n}{4n} \times \left(\mu + \frac{Z\left(\theta_{Futility1}\right)}{1 - \Phi\left(\theta_{Futility1}\right)} \sigma_{\delta 1}\right) + \frac{2n}{4n} \times \mu$$

$$= \frac{1}{2}\left(\left(\mu + \frac{Z\left(\theta_{Futility1}\right)}{1 - \Phi\left(\theta_{Futility1}\right)} \sigma_{\delta 1}\right) + \mu\right)$$

$$= \mu + \frac{Z\left(\theta_{Futility1}\right)}{2\left(1 - \Phi\left(\theta_{Futility1}\right)\right)} \sigma_{\delta 1} \qquad (8)$$

Take a weighted average of (6) and (8), with weight corresponding to the probability of stopping and not stopping, we get the overall mean for a single study,

$$\Pr(\text{stop}) \times E\left(\widehat{\delta}_1 \middle| \text{stop for futility}\right) + \Pr(\text{not stop}) \times E\left(\widehat{\delta}_1 \middle| \text{completed}\right)$$

$$= p \times \left(\mu - \frac{Z\left(\theta_{Futility1}\right)}{\Phi\left(\theta_{Futility1}\right)} \sigma_{\delta 1}\right)$$

$$+ (1 - p) \times \left(\mu + \frac{Z\left(\theta_{Futility1}\right)}{2\left(1 - \Phi\left(\theta_{Futility1}\right)\right)} \sigma_{\delta 1}\right)$$

$$= \mu - \frac{1}{2} \cdot Z\left(\theta_{Futility1}\right) \cdot \sigma_{\delta 1} \qquad (9)$$

This indicates a bias of $-\frac{1}{2} \cdot Z\left(\theta_{Futility1}\right) \cdot \sigma_{\delta 1}$, which is always negative, so the overall mean of a single study with a stopping rule for futility always underestimate the treatment effect on average.

This bias is maximized when $\theta_{Futility1} = 0$,

$$\theta_{Futility1} = \frac{\delta_1^* - \mu}{\sigma_{\delta 1}} = 0$$

$$\Rightarrow \delta_1^* - \mu = \sigma \sqrt{\frac{2}{n}} \cdot \left(\sqrt{2} c_\alpha - c_\gamma\right) - \theta_F - \mu = 0$$

$$\Rightarrow \theta_F + \mu = \left(\sqrt{2} c_\alpha - c_\gamma\right) \cdot \sigma \sqrt{2/n}$$

That is, when $\theta_F + \mu = \left(\sqrt{2}c_\alpha - c_\gamma\right) \cdot \sigma\sqrt{2/n}$, the bias attains its maximum. At that point, $Z(0) = 1/\sqrt{2\pi}$, so the maximum bias is $1/\sqrt{N\pi}$.

3.1.5 Relative Weight of Truncated Studies in a Meta-Analysis

The probability that a trial stops at the interim analysis is $\Phi\left(\theta_{Futility1}\right)$. At this point, the trial would have accumulated *2n* observations. The remaining proportion $1 - \Phi\left(\theta_{Futility1}\right)$ of studies do not stop and accumulate an additional 2n observations so have a total of 4n observations at the end. So if a meta-analysis consists of trials which all use the same protocol including the same stopping rule for futility, the relative weight of the truncated studies in the meta-analysis would be,

$$\begin{aligned}
\text{Relative weight} &= \frac{2n \times \Phi\left(\theta_{Futility1}\right)}{2n \times \Phi\left(\theta_{Futility1}\right) + 4n \times \left(1 - \Phi\left(\theta_{Futility1}\right)\right)} \\
&= \frac{\Phi\left(\theta_{Futility1}\right)}{2 - \Phi\left(\theta_{Futility1}\right)} \\
&= \frac{2}{2 - \Phi\left(\theta_{Futility1}\right)} - 1 \qquad (10)
\end{aligned}$$

Since

$$\begin{aligned}
\theta_{Futility1} &= \frac{\delta_1^* - \mu}{\sigma_{\delta 1}} \\
&= \left(\sqrt{2}c_{1-\alpha} - c_{1-\gamma}\right) - \frac{\theta_F + \mu}{\sigma\sqrt{2/n}} \\
&= \left(\sqrt{2}c_{1-\alpha} - c_{1-\gamma}\right) - \frac{\theta_F + \mu}{\sigma} \cdot \sqrt{\frac{n}{2}}
\end{aligned}$$

The relative weight is a decreasing function of μ and $\sqrt{n}$, i.e. more weight is associated with truncated studies if n is small or true treatment effect size is small.

3.2 One Planned Interim Analysis, Two-sided Testing

3.2.1 Threshold Treatment Effect and Probability of Stopping

For one interim analysis two-sided testing, again since $Z_1 = \frac{\hat{\delta}_1}{s.e.(\hat{\delta}_1)} \Rightarrow \hat{\delta}_1 = Z_1 \times s.e.(\hat{\delta}_1) = Z_1 \times \sigma\sqrt{\frac{2}{n}}$ with $\hat{\delta}_1$ as the treatment effect observed at the interim analysis, the mean difference in outcomes between the two groups, by numerically solving inequality (5), we can get the upper and lower threshold Z_{H1}^* and Z_{L1}^* respectively for Z_1, so we have

$$\delta_{L1}^* = Z_{L1}^* \times \sigma\sqrt{\frac{2}{n}} < \hat{\delta}_1 = Z_1 \times \sigma\sqrt{\frac{2}{n}} < Z_{H1}^* \times \sigma\sqrt{\frac{2}{n}} = \delta_{H1}^*$$

Thus, the lower and upper threshold treatment effects are given by δ_{L1}^* and δ_{H1}^* respectively.

Because $CP_1(\theta_F) < \gamma$ is equivalent to $\delta_{L1}^* < \hat{\delta}_1 < \delta_{H1}^*$, at the interim analysis, if the observed treatment effect $\hat{\delta}_1$ falls between δ_{L1}^* and δ_{H1}^*, the trial will stop, otherwise, it will continue. Thus the probability of stopping is $\Pr(\text{stop}) = \Pr(CP_1(\theta_F) < \gamma)$,

$$\Pr(CP_1(\theta) < \gamma) = \mathrm{P}\left(\delta_{L1}^* < \hat{\delta}_1 < \delta_{H1}^*\right)$$

$$= \mathrm{P}\left(\frac{\delta_{L1}^{*}-\mu}{\sigma_{\delta 1}} < \mathrm{Z} = \frac{\hat{\delta}_1-\mu}{\sigma_{\delta 1}} < \frac{\delta_{H1}^{*}-\mu}{\sigma_{\delta 1}}\right)$$

Denote $\frac{\delta_{H1}^{*}-\mu}{\sigma_{\delta 1}}$ as $\theta_{Futility1}$ and $\frac{\delta_{L1}^{*}-\mu}{\sigma_{\delta 1}}$ as $\theta'_{Futility1}$, then

$$\Pr(\text{stopping}) = \Phi(\theta_{Futility1}) - \Phi(\theta'_{Futility1}),$$

so the probability of not stopping is $1-\Phi(\theta_{Futility1})+\Phi(\theta'_{Futility1})$.

3.2.2 Expected Mean for the Studies That Stop for Futility

At the interim analysis, among studies whose treatment effect fall between the upper and lower threshold values and hence stop early for futility, the sample treatment effect has a Normal distribution truncated above and below respectively at δ_{H1}^{*} and δ_{L1}^{*}, so from (1) the expected mean value of such studies is

$$\mathrm{E}(\hat{\delta}_1 | \text{stop for futility}) = \mathrm{E}(\hat{\delta}_1 | \delta_{L1}^{*} < \hat{\delta}_1 < \delta_{H1}^{*})$$

$$= \mu + \frac{\mathrm{Z}\left(\frac{\delta_{L1}^{*}-\mu}{\sigma_{\delta 1}}\right) - \mathrm{Z}\left(\frac{\delta_{H1}^{*}-\mu}{\sigma_{\delta 1}}\right)}{\Phi\left(\frac{\delta_{H1}^{*}-\mu}{\sigma_{\delta 1}}\right) - \Phi\left(\frac{\delta_{L1}^{*}-\mu}{\sigma_{\delta 1}}\right)}\sigma_{\delta 1}$$

$$= \mu - \frac{\mathrm{Z}(\theta_{Futility1}) - \mathrm{Z}(\theta'_{Futility1})}{\Phi(\theta_{Futility1}) - \Phi(\theta'_{Futility1})}\sigma_{\delta 1} \qquad (11)$$

The second term of (11), $-\frac{\mathrm{Z}(\theta_{Futility1})-\mathrm{Z}(\theta'_{Futility1})}{\Phi(\theta_{Futility1})-\Phi(\theta'_{Futility1})}\sigma_{\delta 1}$ gives the expected error in estimating the treatment effect for a truncated study, i.e. a study that stops for futility at the interim analysis.

3.2.3 Expectation for the Studies That Do Not Stop for Futility

In contrast, studies whose effect exceed the upper threshold value δ_H^* (and hence continue with probability $1-\Phi(\theta_{Futility1})$) have interim outcomes with a Normal distribution truncated below at δ_H^*, with expectation

$$\mathrm{E}\left(\widehat{\delta}_1 \middle| \widehat{\delta}_1 > \delta_H^*\right) = \mu + \frac{\mathrm{Z}\left(\frac{\delta_H^* - \mu}{\sigma_{\delta 1}}\right)}{1-\Phi\left(\frac{\delta_H^* - \mu}{\sigma_{\delta 1}}\right)}\sigma_{\delta 1}$$

$$= \mu + \frac{\mathrm{Z}\left(\theta_{Futility1}\right)}{1-\Phi\left(\theta_{Futility1}\right)}\sigma_{\delta 1} \qquad (12)$$

On the other hand, studies whose effects fall below the lower threshold δ_L^* will also continue after the interim analysis, with probability $\Phi(\theta'_{Futility1})$. These studies have a Normal distribution truncated above at δ_L^* with means

$$\mathrm{E}\left(\widehat{\delta}_1 \middle| \delta_L^* < \widehat{\delta}_1\right) = \mu - \frac{\mathrm{Z}\left(\frac{\delta_L^* - \mu}{\sigma_{\delta 1}}\right)}{\Phi\left(\frac{\delta_L^* - \mu}{\sigma_{\delta 1}}\right)}\sigma_{\delta 1}$$

$$= \mu - \frac{\mathrm{Z}\left(\theta'_{Futility1}\right)}{\Phi\left(\theta'_{Futility1}\right)}\sigma_{\delta 1} \qquad (13)$$

Weighted-averaging (12) and (13) gives the expected mean for a non-stopped study at the interim analysis:

$$\left(\mu + \frac{\mathrm{Z}\left(\theta_{Futility1}\right)}{1-\Phi\left(\theta_{Futility1}\right)}\sigma_{\delta 1}\right) \cdot \left(\frac{1-\Phi\left(\theta_{Futility1}\right)}{1-\Phi\left(\theta_{Futility1}\right)+\Phi\left(\theta'_{Futility1}\right)}\right)$$

$$+\left(\mu - \frac{\mathrm{Z}\left(\theta'_{Futility1}\right)}{\Phi\left(\theta'_{Futility1}\right)}\sigma_{\delta 1}\right) \cdot \left(\frac{\Phi\left(\theta'_{Futility1}\right)}{1-\Phi\left(\theta_{Futility1}\right)+\Phi\left(\theta'_{Futility1}\right)}\right)$$

$$= \mu + \frac{\mathrm{Z}\left(\theta_{Futility1}\right) - \mathrm{Z}\left(\theta'_{Futility1}\right)}{1-\Phi\left(\theta_{Futility1}\right)+\Phi\left(\theta'_{Futility1}\right)}\sigma_{\delta 1} \qquad (14)$$

3.2.4 Overall Bias for a Single Study

A study that continues to its final analysis accumulated a further 2n observations and the 2n observations gathered after the 2nd interim analysis are unaffected by the stopping rule and thus give an unbiased estimate of treatment effect μ. Combining with (14), the overall mean for a completed study is given by:

$$\mathrm{E}\left(\hat{\delta}_1 \middle| \text{completed}\right) = \frac{2n}{4n} \times (equation\ (13)) + \frac{2n}{4n} \times \mu$$

$$= \frac{1}{2}\left(\left(\mu + \frac{Z(\theta_{Futility1}) - Z(\theta'_{Futility1})}{1 - \Phi(\theta_{Futility1}) + \Phi(\theta'_{Futility1})}\sigma_{\delta 1}\right) + \mu\right)$$

$$= \mu + \frac{Z(\theta_{Futility1}) - Z(\theta'_{Futility1})}{2\left(1 - \Phi(\theta_{Futility1}) + \Phi(\theta'_{Futility1})\right)}\sigma_{\delta 1} \qquad (15)$$

Take a weighted average of (11) and (15), with weights corresponding to the probability of stopping and not stopping, the overall mean for a single study is

$$\Pr(\text{stop}) \times \mathrm{E}\left(\hat{\delta}_1 \middle| \text{stop for futility}\right) + \Pr(\text{not stop}) \times \mathrm{E}\left(\hat{\delta}_1 \middle| \text{completed}\right)$$

$$= \left(\Phi(\theta_{Futility1}) - \Phi(\theta'_{Futility1})\right) \times \left(\mu - \frac{Z(\theta_{Futility1}) - Z(\theta'_{Futility1})}{\Phi(\theta_{Futility1}) - \Phi(\theta'_{Futility1})}\sigma_{\delta 1}\right)$$

$$+ \left(1 - \Phi(\theta_{Futility1}) + \Phi(\theta'_{Futility1})\right)$$

$$\times \left(\mu + \frac{Z(\theta_{Futility1}) - Z(\theta'_{Futility1})}{2\left(1 - \Phi(\theta_{Futility1}) + \Phi(\theta'_{Futility1})\right)}\sigma_{\delta 1}\right)$$

$$= \mu - \tfrac{1}{2} \cdot \left(Z(\theta_{Futility1}) - Z(\theta'_{Futility1})\right) \cdot \sigma_{\delta 1}$$

3.2.5 Relative Weight of Truncated Studies in a Meta-Analysis

The probability that a trial stops at the interim analysis is $\Pr(\text{stop}) = \Phi(\theta_{Futility}) - \Phi(\theta'_{Futility})$, at this point, the trial would have accumulated *2n* observations. The remaining proportion $\Pr(\text{not stop}) = 1 - \Phi(\theta_{Futility}) + \Phi(\theta'_{Futility})$ of studies do not stop and accumulate an additional 2n observations so have a total of 4n observations at the end. So if a meta-analysis consists of trials which all use the same protocol including the same stopping rule for futility, the relative weight of the truncated studies in the meta-analysis would be,

$$\begin{aligned}
\text{Relative weight} &= \frac{2n \times \Pr(\text{stop})}{2n \times \Pr(\text{stop}) + 4n \times \Pr(\text{not stop})} \\
&= \frac{\Pr(\text{stop})}{\Pr(\text{stop}) + 2\Pr(\text{not stop})} \\
&= \frac{\Phi(\theta_{Futility1}) - \Phi(\theta'_{Futility1})}{2 - [\Phi(\theta_{Futility1}) - \Phi(\theta'_{Futility1})]} \\
&= \frac{2}{2 - [\Phi(\theta_{Futility1}) - \Phi(\theta'_{Futility1})]} - 1
\end{aligned}$$

3.3 Two Planned Interim Analyses, One-sided Testing

3.3.1 Threshold Treatment Effects and Probability of Stopping

For two planned interim analyses with one-sided testing, according to inequality (4),

$$Z_1 < Z_1^* = \frac{c_{1-\alpha}}{\sqrt{t_1}} - c_{1-\gamma}\frac{\sqrt{1-t_1}}{\sqrt{t_1}} - \theta_F\sqrt{I_K}\frac{(1-t_1)}{\sqrt{t_1}}$$

$$Z_2 < Z_2^* = \frac{c_{1-\alpha}}{\sqrt{t_2}} - c_{1-\gamma}\frac{\sqrt{1-t_2}}{\sqrt{t_2}} - \theta_F\sqrt{I_K}\frac{(1-t_2)}{\sqrt{t_2}}$$

where Z_i denotes the test statistic at the *i*th analysis with *i=1, 2,* $\mathrm{I}_K = var^{-1}(\hat{\theta}_K) = (\frac{\sigma^2}{3n} + \frac{\sigma^2}{3n})^{-1} = (\frac{2\sigma^2}{3n})^{-1}$ and $t_1 = \frac{1}{3}$, $t_2 = \frac{2}{3}$.

Since

$$Z_1 = \frac{\hat{\delta}_1}{s.e.(\hat{\delta}_1)} \Rightarrow \hat{\delta}_1 = Z_1 \times s.e.(\hat{\delta}_1) = Z_1 \times \sqrt{(\frac{\sigma^2}{n} + \frac{\sigma^2}{n})} = Z_1 \times \sigma\sqrt{\frac{2}{n}}$$

and

$$Z_2 = \frac{\hat{\delta}_2}{s.e.(\hat{\delta}_2)} \Rightarrow \hat{\delta}_2 = Z_2 \times s.e.(\hat{\delta}_2) = Z_2 \times \sqrt{(\frac{\sigma^2}{2n} + \frac{\sigma^2}{2n})} = Z_2 \times \sigma\sqrt{\frac{1}{n}}$$

where $\hat{\delta}_1$ and $\hat{\delta}_2$ are the treatment effects observed at the 1st and 2nd interim analyses, the mean differences in outcomes between the two groups,

$$\begin{aligned}
\widehat{\delta}_1 &= Z_1 \times \sigma\sqrt{\frac{2}{n}} < Z_1^* \times \sigma\sqrt{\frac{2}{n}} \\
&= \left(\frac{c_{1-\alpha}}{\sqrt{t_1}} - c_{1-\gamma}\frac{\sqrt{1-t_1}}{\sqrt{t_1}} - \theta_F\sqrt{I_K}\frac{(1-t_1)}{\sqrt{t_1}}\right) \times \sigma\sqrt{\frac{2}{n}} \\
&= \left(\frac{c_{1-\alpha}}{\sqrt{1/3}} - c_{1-\gamma} \times \frac{\sqrt{1-\frac{1}{3}}}{\sqrt{1/3}} - \theta_F \times \sqrt{\frac{3n}{2\sigma^2}} \times \frac{\left(1-\frac{1}{3}\right)}{\sqrt{1/3}}\right) \times \sigma\sqrt{\frac{2}{n}} \\
&= \left(\sqrt{3}c_{1-\alpha} - \sqrt{2}c_{1-\gamma} - \theta_F \times \frac{\sqrt{2n}}{\sigma}\right) \times \sigma\sqrt{\frac{2}{n}} \\
&= \sigma\sqrt{\tfrac{2}{n}}\left(\sqrt{3}c_{1-\alpha} - \sqrt{2}c_{1-\gamma}\right) - 2\theta_F \\
&= \delta_1^*
\end{aligned}$$

and

$$\widehat{\delta}_2 = Z_2 \times \sigma\sqrt{\frac{1}{n}} < Z_2^* \times \sigma\sqrt{\frac{1}{n}}$$

$$= \left(\frac{c_{1-\alpha}}{\sqrt{t_2}} - c_{1-\gamma}\frac{\sqrt{1-t_2}}{\sqrt{t_2}} - \theta_F\sqrt{I_K}\frac{(1-t_2)}{\sqrt{t_2}}\right) \times \sigma\sqrt{\frac{1}{n}}$$

$$= \left(\frac{c_{1-\alpha}}{\sqrt{2/3}} - c_{1-\gamma} \times \frac{\sqrt{1-\frac{2}{3}}}{\sqrt{2/3}} - \theta_F \times \sqrt{\frac{3n}{2\sigma^2}} \times \frac{\left(1-\frac{2}{3}\right)}{\sqrt{2/3}}\right) \times \sigma\sqrt{\frac{1}{n}}$$

$$= \sigma\sqrt{\frac{2}{n}}\left(\frac{\sqrt{3}}{2}c_{1-\alpha} - \frac{\sqrt{2}}{2}c_{1-\gamma}\right) - \frac{\theta_F}{2}$$

$$= \frac{1}{2}\left(\sigma\sqrt{\frac{2}{n}}\left(\sqrt{3}c_{1-\alpha} - \sqrt{2}c_{1-\gamma}\right) - \theta_F\right)$$

$$= \delta_2^*$$

Thus, the stopping threshold at the 1st and 2nd interim analyses are given by, respectively,

$$\delta_1^* = \sigma\sqrt{\frac{2}{n}}\left(\sqrt{3}c_{1-\alpha} - \sqrt{2}c_{1-\gamma}\right) - 2\theta_F$$

$$\delta_2^* = \frac{1}{2}\left(\sigma\sqrt{\frac{2}{n}}\left(\sqrt{3}c_{1-\alpha} - \sqrt{2}c_{1-\gamma}\right) - \theta_F\right)$$

At the 1st (or 2nd) interim analysis, $CP_k(\theta_F) < \gamma$ is equivalent to $Z_1 < Z_1^*$ (or $Z_2 < Z_2^*$), and also to $\widehat{\delta}_1 < \delta_1^*$ (or $\widehat{\delta}_2 < \delta_2^*$). If the observed treatment effect $\widehat{\delta}_i$ is less than δ_i^* *i=1,2*, the trial will stop, otherwise, it will continue at that point. So the probability of stopping at the 1st interim analysis is $\Pr(\text{stop1}) = \Pr(CP_1(\theta_F) < \gamma)$,

$$\Pr(CP_1(\theta_F) < \gamma) = \mathrm{P}(Z_1 < Z_1^*) = \mathrm{P}\left(\widehat{\delta}_1 < \delta_1^*\right)$$

$$= \mathrm{P}\left(\mathrm{Z} = \frac{\widehat{\delta}_1 - \mu_1}{se\left(\widehat{\delta}_1\right)} < \frac{\delta_1^* - \mu_1}{se\left(\widehat{\delta}_1\right)}\right)$$

$$= \mathrm{P}\left(\mathrm{Z} < \frac{\sigma\sqrt{\frac{2}{n}}\left(\sqrt{3}c_{1-\alpha} - \sqrt{2}c_{1-\gamma}\right) - 2\theta_F - \mu_1}{\sigma\sqrt{2/n}\,,}\right)$$

$$= \Phi\left(\left(\sqrt{3}c_{1-\alpha} - \sqrt{2}c_{1-\gamma}\right) - \frac{2\theta_F + \mu_1}{\sigma\sqrt{2/n}}\right)$$

$$= \Phi\left(\theta_{Futility1}\right)$$

where $\theta_{Futility1} = \frac{\delta_1^* - \mu_1}{se(\hat{\delta}_1)}$ and μ_1 is the expected mean treatment effect for the accumulated data in the 1st analysis. Note that $\mu_1 = \mu$.

Thus the probability of continuing after the 1st interim analysis is $1 - \Phi(\theta_{Futility1})$.

The probability of stopping at the 2nd interim analysis is then

$$\Pr(\text{stop2}) = (1 - \Phi(\theta_{Futility1})) \cdot \Pr(CP_2(\theta_F) < \gamma)$$

where

$$\Pr(CP_2(\theta_F) < \gamma) = \mathrm{P}(Z_2 < Z_2^*) = \mathrm{P}\left(\hat{\delta}_2 < \delta_2^*\right)$$

$$= \mathrm{P}\left(\mathrm{Z} = \frac{\hat{\delta}_2 - \mu_2}{se(\hat{\delta}_2)} < \frac{\delta_2^* - \mu_2}{se(\hat{\delta}_2)}\right)$$

$$= \mathrm{P}\left(\mathrm{Z} < \frac{\frac{\sigma}{\sqrt{n}}\left(\sqrt{\frac{3}{2}}c_{1-\alpha} - \sqrt{\frac{1}{2}}\sqrt{2}c_{1-\gamma}\right) - \frac{\theta_F}{2} - \mu_2}{\sigma/\sqrt{n},}\right)$$

$$= \Phi\left(\left(\sqrt{\frac{3}{2}}c_{1-\alpha} - \sqrt{\frac{1}{2}}\sqrt{2}c_{1-\gamma}\right) - \frac{\frac{1}{2}\theta_F + \mu_2}{\sigma/\sqrt{n}}\right)$$

$$= \Phi\left(\theta_{Futility2}\right)$$

where $\theta_{Futility2} = \frac{\delta_2^* - \mu_2}{se(\hat{\delta}_2)}$, μ_2 is the expected mean treatment effect for the accumulated data in the 2nd analysis. Note that μ_2, the expected mean at the time of 2nd interim analysis for a study that continues after the 1st analysis, is given by (8).

Therefore the probability of continuing after the 2nd interim analysis is

$$\left(1-\Phi(\theta_{Futility1})\right)\cdot\left(1-\Phi(\theta_{Futility2})\right).$$

3.3.2 Expected Mean for Studies That Stop for Futility

The expected mean for a study that stops at the 1st interim analysis is given by (6).

A study that stops at the 2nd analysis has a Normal distribution of outcomes with mean given by (8)

$$\mu_2=\mu_1+\frac{Z(\theta_{Futility1})}{2\left(1-\Phi(\theta_{Futility1})\right)}\sigma_{\delta 1}=\mu+\frac{Z(\theta_{Futility1})}{2\left(1-\Phi(\theta_{Futility1})\right)}\sigma_{\delta 1}$$

truncated above at δ_2^*, thus giving an expected mean for such a study as

$$\mathrm{E}\left(\hat{\delta}_2\middle|\text{stop for futility}\right)=\mathrm{E}\left(\hat{\delta}_2\middle|\hat{\delta}_2<\delta_2^*\right)=\mu_2-\frac{Z\left(\frac{\delta_2^*-\mu_2}{\sigma_{\delta 2}}\right)}{\Phi\left(\frac{\delta_2^*-\mu_2}{\sigma_{\delta 2}}\right)}\sigma_{\delta 2}$$

$$=\mu_2-\frac{Z(\theta_{Futility2})}{\Phi(\theta_{Futility2})}\sigma_{\delta 2}$$

$$=\mu+\frac{Z(\theta_{Futility1})}{2\left(1-\Phi(\theta_{Futility1})\right)}\sigma_{\delta 1}-\frac{Z(\theta_{Futility2})}{\Phi(\theta_{Futility2})}\sigma_{\delta 2} \qquad (15)$$

3.3.3 Expected Mean for Studies That Do Not Stop for Futility

The expected mean for a study that does not stop at the 1st interim analysis is given by (7).

The corresponding mean for a study that does not stop at the 2nd analysis, having a Normal distribution truncated below at δ_2^* with mean μ_2, is

$$\mathrm{E}\left(\widehat{\delta}_2 \mid \widehat{\delta}_2 > \delta_2^*\right) = \mu_2 + \frac{\mathrm{Z}\left(\frac{\delta_2^* - \mu_2}{\sigma_{\delta 2}}\right)}{1 - \Phi\left(\frac{\delta_2^* - \mu_2}{\sigma_{\delta 2}}\right)} \sigma_{\delta 2}$$

$$= \mu_2 + \frac{\mathrm{Z}(\theta_{Futility2})}{1 - \Phi(\theta_{Futility2})} \sigma_{\delta 2}$$

$$= \mu + \frac{\mathrm{Z}(\theta_{Futility1})}{2\left(1 - \Phi(\theta_{Futility1})\right)} \sigma_{\delta 1} + \frac{\mathrm{Z}(\theta_{Futility2})}{1 - \Phi(\theta_{Futility2})} \sigma_{\delta 2} \qquad (16)$$

3.3.4 Overall Bias for a Single Study

The expected mean treatment effect for a study that does not stop at the 1st interim analysis and thus continues to the 2nd interim analysis is given by (8).

A study that continues to its final analysis accumulates the last third of its data unaffected by the stopping rule, and hence its mean is a weighted average of (16) and μ:

$$\frac{4n}{6n} \times (equation\ (16)) + \frac{2n}{6n} \times \mu$$

$$= \frac{2}{3} \times \left(\mu + \frac{\mathrm{Z}(\theta_{Futility1})}{2\left(1 - \Phi(\theta_{Futility1})\right)} \sigma_{\delta 1} + \frac{\mathrm{Z}(\theta_{Futility2})}{1 - \Phi(\theta_{Futility2})} \sigma_{\delta 2} \right) + \frac{1}{3} \times \mu$$

$$= \mu + \frac{1}{3} \cdot \frac{\mathrm{Z}(\theta_{Futility1})}{1 - \Phi(\theta_{Futility1})} \sigma_{\delta 1} + \frac{2}{3} \cdot \frac{\mathrm{Z}(\theta_{Futility2})}{1 - \Phi(\theta_{Futility2})} \sigma_{\delta 2} \qquad (17)$$

Combining (6), (15) and (17) with probabilities of $\Phi(\theta_{Futility1})$, $\left(1 - \Phi(\theta_{Futility1})\right) \cdot \Phi(\theta_{Futility2})$ and $\left(1 - \Phi(\theta_{Futility1})\right) \cdot \left(1 - \Phi(\theta_{Futility2})\right)$ as their weights respectively, gives the overall mean as

$$\mu - \frac{1}{3}\,\mathrm{Z}(\theta_{Futility2}) \cdot \left(1 - \Phi(\theta_{Futility1})\right) \cdot \sigma_{\delta 2}$$
$$+ \mathrm{Z}(\theta_{Futility1}) \cdot \left(\frac{1}{6}\Phi(\theta_{Futility2}) - \frac{2}{3}\right) \cdot \sigma_{\delta 1}.$$

3.3.5 Relative Weight of Truncated Studies in a Meta-Analysis

The relative weight for truncated studies in a meta-analysis is

$$\frac{2n\Phi_1 + 4n(1-\Phi_1)\Phi_2}{2n\Phi_1 + 4n(1-\Phi_1)\Phi_2 + 6n(1-\Phi_1)(1-\Phi_2)} = \frac{\Phi_1 + 2\Phi_2 - 2\Phi_1\Phi_2}{3 - 2\Phi_1 - \Phi_2 + \Phi_1\Phi_2}$$

where Φ_1 denotes $\Phi(\theta_{Futility1})$ and Φ_2 denotes $\Phi(\theta_{Futility2})$.

3.4 Two Planned Interim Analyses, Two-sided Testing

3.4.1 Threshold Treatment Effects and Probabilities of Stopping

For two planned interim analysis two-sided testing, we can solve inequality (5) numerically to get a solution of the upper and lower thresholds at the 1st interim analysis and the 2nd interim analysis, namely Z_{H1}^*, Z_{L1}^*, Z_{H2}^* and Z_{L2}^* respectively for $Z_1\ and\ Z_2$. Since $Z_i = \frac{\hat{\delta}_i}{s.e.(\hat{\delta}_i)} \Rightarrow \hat{\delta}_i = Z_i \times \sigma_{\delta i}$, $i = 1, 2$ with Z_i as the test statistic at the i^{th} interim analysis, $\hat{\delta}_i$ as the observed treatment effect, and $\sigma_{\delta i}$as the standard error of $\hat{\delta}_i$, we have:

$$\delta_{Li}^* = Z_{Li}^* \times \sigma_{\delta i} < \hat{\delta}_1 = Z_1 \times \sigma_{\delta i} < Z_{Hi}^* \times \sigma_{\delta i} = \delta_{Hi}^*, \qquad i = 1, 2$$

where $\sigma_{\delta 1} = \sigma\sqrt{\frac{2}{n}}$ and $\sigma_{\delta 2} = \sigma\sqrt{\frac{1}{n}}$.

Hence the lower and upper threshold treatment effects at 1st and 2nd interim analysis are given by, respectively, δ_{L1}^{*}, δ_{H1}^{*}, δ_{L2}^{*} and δ_{H2}^{*}.

Similarly, at the 1st (or 2nd) interim analysis, $CP_k(\theta_F) < \gamma$ is equivalent to $\delta_{L1}^{*} < \widehat{\delta}_1 < \delta_{H1}^{*}$ (or $\delta_{L2}^{*} < \widehat{\delta}_2 < \delta_{H2}^{*}$). If the observed treatment effect $\widehat{\delta}_i$ falls between δ_{Li}^{*} and δ_{Hi}^{*} with i=1, 2, the trial will stop, otherwise, it will continue at that point. So the probability of stopping at the 1st interim analysis is

$$
\begin{aligned}
\Pr(\text{stop1}) &= \Pr(CP_1(\theta_F) < \gamma) \\
&= \mathrm{P}\left(\delta_{L1}^{*} < \widehat{\delta}_1 < \delta_{H1}^{*}\right) \\
&= \mathrm{P}\left(\frac{\delta_{L1}^{*} - \mu_1}{\sigma_{\delta 1}} < \mathrm{Z} = \frac{\widehat{\delta}_1 - \mu_1}{\sigma_{\delta 1}} < \frac{\delta_{H1}^{*} - \mu_1}{\sigma_{\delta 1}}\right) \\
&= \Phi\left(\theta_{Futility1}\right) - \Phi(\theta'_{Futility1})
\end{aligned}
$$

where $\theta_{Futility1} = \frac{\delta_{H1}^{*} - \mu_1}{se(\widehat{\delta}_1)}$, $\theta'_{Futility1} = \frac{\delta_{L1}^{*} - \mu_1}{se(\widehat{\delta}_1)}$ and μ_1 is the expected mean treatment effect for the accumulated data in the 1st analysis. Note that $\mu_1 = \mu$.

Thus the probability of continuing at 1st interim analysis is $1 - \left(\Phi\left(\theta_{Futility1}\right) - \Phi\left(\theta'_{Futility1}\right)\right) = 1 - \Phi\left(\theta_{Futility1}\right) + \Phi(\theta'_{Futility1})$.

The probability of stopping at the 2nd interim analysis is then

$$
\Pr(\text{stop2}) = (\, 1 - \Phi\left(\theta_{Futility1}\right) + \Phi(\theta'_{Futility1})) \cdot \Pr(CP_2(\theta_F) < \gamma)
$$

where

$$
\begin{aligned}
\Pr(CP_2(\theta_F) < \gamma) &= \mathrm{P}\left(\delta_{L2}^{*} < \widehat{\delta}_2 < \delta_{H2}^{*}\right) \\
&= \mathrm{P}\left(\frac{\delta_{L2}^{*} - \mu_2}{\sigma_{\delta 2}} < \mathrm{Z} = \frac{\widehat{\delta}_2 - \mu_2}{\sigma_{\delta 2}} < \frac{\delta_{H2}^{*} - \mu_2}{\sigma_{\delta 2}}\right) \\
&= \Phi\left(\theta_{Futility2}\right) - \Phi\left(\theta'_{Futility2}\right)
\end{aligned}
$$

where $\theta_{Futility2} = \frac{\delta^*_{H2}-\mu_2}{se(\hat{\delta}_2)}$, $\theta'_{Futility2} = \frac{\delta^*_{L2}-\mu_2}{se(\hat{\delta}_2)}$ and μ_2 is the expected mean treatment effect for the accumulated data in the 2nd analysis. Note that μ_2, the expected mean at the time of 2nd interim analysis for a study that continues after the 1st analysis, is given by (14).

Therefore the probability of continuing at the 2nd interim analysis is

$$\Big(1-\Phi\big(\theta_{Futility1}\big)+\Phi\big(\theta'_{Futility1}\big)\Big)\cdot\big(1-\Phi\big(\theta_{Futility2}\big)+\Phi\big(\theta'_{Futility2}\big)\big).$$

3.4.2 Expected Mean for Studies That Stop for Futility

The expected mean for a study that stopped at the 1st interim analysis is given by (10).

A study that stops at the 2nd analysis has a Normal distribution of outcomes with mean given by (14)

$$\mu_2 = \mu + \frac{Z\big(\theta_{Futility1}\big)-Z\big(\theta'_{Futility1}\big)}{2\Big(1-\Phi\big(\theta_{Futility1}\big)+\Phi\big(\theta'_{Futility1}\big)\Big)}\sigma_{\delta 1}$$

truncated below and above at δ^*_{L2} and δ^*_{H2}, thus giving an expected mean for such a study as

$$E\big(\hat{\delta}_2|\delta^*_{L2} < \hat{\delta}_2 < \delta^*_{H2}\big) = \mu_2 - \frac{Z\left(\frac{\delta^*_{H2}-\mu_2}{\sigma_{\delta 2}}\right)-Z\left(\frac{\delta^*_{L2}-\mu_2}{\sigma_{\delta 2}}\right)}{\Phi\left(\frac{\delta^*_{H2}-\mu_2}{\sigma_{\delta 2}}\right)-\Phi\left(\frac{\delta^*_{L2}-\mu_2}{\sigma_{\delta 2}}\right)}\sigma_{\delta 2}$$

$$= \mu_2 - \frac{Z\big(\theta_{Futility2}\big)-Z\big(\theta'_{Futility2}\big)}{\Phi\big(\theta_{Futility2}\big)-\Phi\big(\theta'_{Futility2}\big)}\sigma_{\delta 2}$$

$$= \mu + \frac{Z(\theta_{Futility1}) - Z(\theta'_{Futility1})}{2\left(1 - \Phi(\theta_{Futility1}) + \Phi(\theta'_{Futility1})\right)}\sigma_{\delta 1}$$

$$- \frac{Z(\theta_{Futility2}) - Z(\theta'_{Futility2})}{\Phi(\theta_{Futility2}) - \Phi(\theta'_{Futility2})}\sigma_{\delta 2} \qquad (18)$$

3.4.3 Overall Mean for the Studies That Do Not Stop for Futility

The expected mean for a study that does not stop at the 1st interim analysis is given by (13).

At the 2nd interim analysis, studies whose effect exceed the upper threshold value δ^*_{H2} (and hence continue with probability $\left(1 - \Phi(\theta_{Futility1}) + \Phi(\theta'_{Futility1})\right)\left(1 - \Phi(\theta_{Futility2})\right)$) have interim outcomes with a Normal distribution truncated below at δ^*_{H2}, with expectation

$$\mathrm{E}\left(\hat{\delta}_2 | \hat{\delta}_2 > \delta^*_{H2}\right) = \mu_2 + \frac{Z\left(\frac{\delta^*_{H2} - \mu}{\sigma_{\delta 2}}\right)}{1 - \Phi\left(\frac{\delta^*_{H2} - \mu}{\sigma_{\delta 2}}\right)}\sigma_{\delta 2}$$

$$= \mu_2 + \frac{Z(\theta_{Futility2})}{1 - \Phi(\theta_{Futility2})}\sigma_{\delta 2}$$

$$= \mu + \frac{Z(\theta_{Futility1}) - Z(\theta'_{Futility1})}{2\left(1 - \Phi(\theta_{Futility1}) + \Phi(\theta'_{Futility1})\right)}\sigma_{\delta 1} + \frac{Z(\theta_{Futility2})}{1 - \Phi(\theta_{Futility2})}\sigma_{\delta 2} \qquad (19)$$

On the other hand, studies whose effects fall below the lower threshold δ^*_{L2} will also continue after the 2nd interim analysis, with probability of $\left(1 - \Phi(\theta_{Futility1}) + \Phi(\theta'_{Futility1})\right) \cdot \Phi(\theta'_{Futility2})$. These studies have a Normal distribution truncated above at δ^*_{L2} with means

$$\mathrm{E}\left(\hat{\delta}_2 \mid \hat{\delta}_2 < \delta^*_{L2}\right) = \mu_2 - \frac{\mathrm{Z}\left(\frac{\delta^*_{L2} - \mu}{\sigma_{\delta 2}}\right)}{\Phi\left(\frac{\delta^*_{L2} - \mu}{\sigma_{\delta 2}}\right)} \sigma_{\delta 2}$$

$$= \mu_2 - \frac{\mathrm{Z}\left(\theta'_{Futility2}\right)}{\Phi\left(\theta'_{Futility2}\right)} \sigma_{\delta 2}$$

$$= \mu + \frac{\mathrm{Z}\left(\theta_{Futility1}\right) - \mathrm{Z}\left(\theta'_{Futility1}\right)}{2\left(1 - \Phi\left(\theta_{Futility1}\right) + \Phi\left(\theta'_{Futility1}\right)\right)} \sigma_{\delta 1} - \frac{\mathrm{Z}\left(\theta'_{Futility2}\right)}{\Phi\left(\theta'_{Futility2}\right)} \sigma_{\delta 2} \quad (20)$$

Weightedly averaging (19) and (20) gives the expected mean for a non-stopped study at the 2nd interim analysis:

$$(\text{equation } (18)) \cdot \left(\frac{\left(1 - \Phi\left(\theta_{Futility1}\right) + \Phi\left(\theta'_{Futility1}\right)\right) \cdot \left(1 - \Phi\left(\theta_{Futility2}\right)\right)}{\left(1 - \Phi\left(\theta_{Futility1}\right) + \Phi\left(\theta'_{Futility1}\right)\right) \cdot \left(1 - \Phi\left(\theta_{Futility2}\right) + \Phi\left(\theta'_{Futility2}\right)\right)}\right)$$

$$+ (\text{equaion } (19)) \cdot \left(\frac{\left(1 - \Phi\left(\theta_{Futility1}\right) + \Phi\left(\theta'_{Futility1}\right)\right) \cdot \Phi\left(\theta'_{Futility2}\right)}{\left(1 - \Phi\left(\theta_{Futility1}\right) + \Phi\left(\theta'_{Futility1}\right)\right) \cdot \left(1 - \Phi\left(\theta_{Futility2}\right) + \Phi\left(\theta'_{Futility2}\right)\right)}\right)$$

$$= \mu + \frac{\mathrm{Z}\left(\theta_{Futility1}\right) - \mathrm{Z}\left(\theta'_{Futility1}\right)}{2\left(1 - \Phi\left(\theta_{Futility1}\right) + \Phi\left(\theta'_{Futility1}\right)\right)} \sigma_{\delta 1} + \frac{\mathrm{Z}\left(\theta_{Futility2}\right) - \mathrm{Z}\left(\theta'_{Futility2}\right)}{1 - \Phi\left(\theta_{Futility2}\right) + \Phi\left(\theta'_{Futility2}\right)} \sigma_{\delta 2} \quad (21)$$

3.4.4 Overall Bias for a Single Study

A study that continues to its final analysis accumulated a further 2n observations and the 2n observations gathered after the 2nd interim analysis are unaffected by the stopping rule thus have an unbiased estimate of treatment effect μ. Combining with (21), the overall mean for a complete study is given by:

$$\frac{4n}{6n}\times(equation\ (21))+\frac{2n}{6n}\times\mu$$
$$=\frac{2}{3}\left(\mu+\frac{Z(\theta_{Futility1})-Z(\theta'_{Futility1})}{2\left(1-\Phi(\theta_{Futility1})+\Phi(\theta'_{Futility1})\right)}\sigma_{\delta 1}\right.$$
$$\left.+\frac{Z(\theta_{Futility2})-Z(\theta'_{Futility2})}{1-\Phi(\theta_{Futility2})+\Phi(\theta'_{Futility2})}\sigma_{\delta 2}\right)+\frac{1}{3}\mu$$
$$=\mu+\frac{1}{3}\cdot\frac{Z(\theta_{Futility1})-Z(\theta'_{Futility1})}{1-\Phi(\theta_{Futility1})+\Phi(\theta'_{Futility1})}\sigma_{\delta 1}$$
$$+\frac{2}{3}\cdot\frac{Z(\theta_{Futility2})-Z(\theta'_{Futility2})}{1-\Phi(\theta_{Futility2})+\Phi(\theta'_{Futility2})}\sigma_{\delta 2} \qquad (22)$$

Combining (10), (18) and (22) with probabilities of $(\theta_{Futility1})-\Phi(\theta'_{Futility1})$, $\left(1-\Phi(\theta_{Futility1})+\Phi(\theta'_{Futility1})\right)\cdot\left(\Phi(\theta_{Futility2})-\Phi(\theta'_{Futility2})\right)$

and

$\left(1-\Phi(\theta_{Futility1})+\Phi(\theta'_{Futility1})\right)\cdot\left(1-\Phi(\theta_{Futility2})+\Phi(\theta'_{Futility2})\right)$

respectively, gives the overall mean as

$$\mu-\frac{1}{3}\sigma_{\delta 2}\cdot\left(Z(\theta_{Futility2})-Z(\theta'_{Futility2})\right)\cdot\left(1-\Phi(\theta_{Futility1})+\Phi(\theta'_{Futility1})\right)$$
$$+\sigma_{\delta 1}\cdot\left(Z(\theta_{Futility1})-Z(\theta'_{Futility1})\right)\cdot\left(\frac{1}{6}\left(\Phi(\theta_{Futility2})-\Phi(\theta'_{Futility2})\right)-\frac{2}{3}\right).$$

3.4.5 Relative Weight of Truncated Studies in a Meta-Analysis

The relative weight for truncated studies in a meta-analysis is

$$\frac{2n\cdot(\Phi_1-\Phi_1')+4n\cdot(1-\Phi_1+\Phi_1')(\Phi_2-\Phi_2')}{2n\cdot(\Phi_1-\Phi_1')+4n\cdot(1-\Phi_1+\Phi_1')(\Phi_2-\Phi_2')+6n\cdot(1-\Phi_1+\Phi_1')(1-\Phi_2+\Phi_2')}$$
$$=\frac{(\Phi_1-\Phi_1')+2(\Phi_2-\Phi_2')-2(\Phi_1-\Phi_1')(\Phi_2-\Phi_2')}{3-2(\Phi_1-\Phi_1')-(\Phi_2-\Phi_2')+(\Phi_1-\Phi_1')(\Phi_2-\Phi_2')}$$

where Φ_1 denotes $\Phi\left(\theta_{Futility1}\right)$, Φ_1' denotes $\Phi\left(\theta'_{Futility1}\right)$, Φ_2 denotes $\Phi\left(\theta_{Futility2}\right)$ and Φ_2' denotes $\Phi\left(\theta'_{Futility2}\right)$.

3.5 Summary of the Formulae

	1-interim analysis, 1-sided	1-interim analysis, 2-sided
Expected treatment effect size for tRCTs	$\mu - \frac{Z(\theta_{Futility1})}{\Phi(\theta_{Futility1})}\sigma_{\delta 1}$ ①	$\mu - \frac{Z(\theta_{Futility1}) - Z(\theta'_{Futility1})}{\Phi(\theta_{Futility1}) - \Phi(\theta'_{Futility1})}\sigma_{\delta 1}$ ②
Expected effect size for a completed study	$\mu + \frac{Z(\theta_{Futility1})}{2[1-\Phi(\theta_{Futility1})]}\sigma_{\delta 1}$ ①	$\mu + \frac{Z(\theta_{Futility1}) - Z(\theta'_{Futility1})}{2\{1-[\Phi(\theta_{Futility1}) - \Phi(\theta'_{Futility1})]\}}\sigma_{\delta 1}$ ②
Overall mean	$\mu - \frac{1}{2}\, Z(\theta_{Futility1})\, \sigma_{\delta 1}$ ①	$\mu - \frac{1}{2}\, [Z(\theta_{Futility1}) - Z(\theta'_{Futility1})]\, \sigma_{\delta 1}$ ②
Relative weight of tRCTs in a meta-analysis	$\frac{\Phi(\theta_{Futility1})}{2-\Phi(\theta_{Futility1})}$ ①	$\frac{\Phi(\theta_{Futility1}) - \Phi(\theta'_{Futility1})}{2-[\Phi(\theta_{Futility1}) - \Phi(\theta'_{Futility1})]}$ ②
Prob. of stopping	$\Phi(\theta_{Futility1})$ ①	$\Phi(\theta_{Futility1}) - \Phi(\theta'_{Futility1})$ ②

	2-interim analysis, 1-sided	2-interim analysis, 2-sided
μ_2	$\mu + \frac{Z_1}{2(1-\Phi_1)}\sigma_{\delta 1}$	$\mu + \frac{Z_1 - Z_1'}{2(1-\Phi_1+\Phi_1')}\sigma_{\delta 1}$
Expected effect size for tRCTs	$\mu_2 - \frac{Z_2}{\Phi_2}\sigma_{\delta 2}$	$\mu_2 - \frac{Z_2 - Z_2'}{\Phi_2 - \Phi_2'}\sigma_{\delta 2}$
For studies that continue at 2nd analysis	$\mu_2 + \frac{Z_2}{1-\Phi_2}\sigma_{\delta 2}$	$\mu_2 + \frac{Z_2 - Z_2'}{1-\Phi_2+\Phi_2'}\sigma_{\delta 2}$
Weighted mean for a study that does not stop at 2nd analysis	$\mu + \frac{Z_1}{3(1-\Phi_1)}\sigma_{\delta 1}$ $+\frac{2Z_2}{3(1-\Phi_2)}\sigma_{\delta 2}$	$\mu + \frac{Z_1 - Z_1'}{3(1-\Phi_1+\Phi_1')}\sigma_{\delta 1} + \frac{2(Z_2 - Z_2')}{3(1-\Phi_2+\Phi_2')}\sigma_{\delta 2}$
Overall mean for a study with a stopping rule	$\mu - \frac{1}{2}Z_2 \cdot (1-\Phi_1)\sigma_{\delta 2}$ $+Z_1 \cdot \sigma_{\delta 1}(\frac{1}{6}\Phi_2 - \frac{2}{3})$	$\mu - \frac{1}{2}(Z_2 - Z_2') \cdot (1-\Phi_1+\Phi_1')\sigma_{\delta 2}$ $+(Z_1 - Z_1') \cdot \sigma_{\delta 1}[\frac{1}{6}(\Phi_2 - \Phi_2') - \frac{2}{3})]$
Relative weight for truncated studies in a meta-analysis	$\frac{\Phi_1 + 2\Phi_2 - 2\Phi_1\Phi_2}{3 - 2\Phi_1 - \Phi_2 + \Phi_1\Phi_2}$	$\frac{(\Phi_1 - \Phi_1') + 2(\Phi_2 - \Phi_2') - 2(\Phi_1 - \Phi_1')(\Phi_2 - \Phi_2')}{3 - 2(\Phi_1 - \Phi_1') - (\Phi_2 - \Phi_2') + (\Phi_1 - \Phi_1')(\Phi_2 - \Phi_2')}$
Prob. of stopping	$(1-\Phi_1) \cdot \Phi_2$	$(1-\Phi_1+\Phi_1') \cdot (\Phi_2 - \Phi_2')$

μ – true treatment effect size

$\theta_{Futilityi} = \frac{\delta^*_{Hi} - \mu}{\sigma_{\delta i}}$, $\theta'_{Futilityi} = \frac{\delta^*_{Li} - \mu}{\sigma_{\delta i}}$ (i=1, 2)

$Z_i = Z(\theta_{Futilityi})$, $Z_i{}' = Z(\theta'{}_{Futilityi})$, $\Phi_i = \Phi(\theta_{Futilityi})$, $\Phi_i{}' = \Phi(\theta'{}_{Futilityi})$ i=1, 2

① $\theta_{Futility1} = \frac{\delta^*_{1} - \mu}{\sigma_{\delta 1}}$

② $\theta_{Futility1} = \frac{\delta^*_{H1} - \mu}{\sigma_{\delta 1}}$, $\theta'_{Futility1} = \frac{\delta^*_{L1} - \mu}{\sigma_{\delta 1}}$

Chapter 4

Results

To avoid discrepancies caused by rounding, we use exact non-integer values as our sample sizes, rather than the rounded numbers.

We first examine numerical results for trials with one planned interim analysis, then the ones with two planned interim analyses. We consider and describe each of the parameters of interest in turn.

The range of the effect size in our tables is from 0 to 1.0, which covers most of the possible treatment outcome values in practice. The values of the power used to calculate the planned sample sizes are 80%, 85%, 90% and 95% which are also commonly used values, we fix the type I error level α as 0.05 and the standard deviation σ as 1 for analytical simplicity. We use 0.15 as the threshold value for CP.

4.1 Trials With One Planned Interim Analysis, One-sided Testing

4.1.1 Under-estimation for Truncated RCTs

Based on equation (6), we see from Table 1a, 1b-1, 1b-2 and 1c that the trend varies for the under-estimation of the treatment effect in a study that stops early for futility, i.e. truncated randomized controlled trials (truncated RCTs), depending on what we assume about the future data.

The magnitude of the under-estimation gets greater when the true effect size μ is greater. This is true for each of the three assumptions, design, current trend and null assumption.

Under the design and current trend assumptions, the under-estimation becomes greater as the power to detect the minimally clinically meaningful treatment effect increases. This is not always true for the null assumption, because here under-estimation increases with the power only when the minimally clinically meaningful treatment effect θ_A is small, but it declines with the power when θ_A is not small.

The under-estimation increases as θ_A increases under the design assumption, while it declines with θ_A under the current trend assumption. It increases with θ_A only when θ_A is not small for the null assumption.

Also under the current trend assumption, the under-estimation increases with the effect estimated from the current data.

Relative to μ, the under-estimation is always substantial no matter what assumption we make about the future data.

4.1.2 Over-estimation for Non-Truncated RCTs

In order for a study with a stopping rule for futility to continue its full sample size, it must show results that exceed the stopping boundaries at any of its interim analyses, and hence over-estimation of the treatment effect occurs. Equation (7) gives the over-estimation for non-truncated RCTs, i.e. trials with stopping rules but which do not actually stop for futility. From Table 2a, 2b1, 2b2, 2c, there is a clear trend that the over-estimation is smaller in several cases: if μ is large, the power to detect the minimally clinically meaningful treatment effect or the effect estimated from the current data is large, or when θ_A is small. These are true for all three assumptions.

Compared to the under-estimation for truncated RCTs, the over-estimation is small but can be non-trivial for the design assumption and substantial for the null assumption. And relative to the true effect sizes, these deviations should not be ignored.

Serious over-estimation can occur under the current trend assumption if the effect estimated from the current data is particularly negatively great or μ is small. For instance, under the current trend assumption, when $\hat{\delta}_{interim} = \mu - 2.0 \times se(\hat{\delta}_{interim})$, if power=80% and θ_A=1.0, the over estimation is 0.8969 and 0.8037 when μ=0 or 0.1 respectively. But this is quite unlikely to occur because there is a very strong probability of the stopping boundary not being crossed, for instance, for the previous example, the probability of stopping is 99.95% and 99.81% respectively; - hence the study will quite unlikely to continue in these cases.

4.1.3 Overall Bias

Table 3a, 3b1, 3b2 and 3c show the overall bias in studies with a one-sided stopping rule calculated from equation (9). The overall bias is typically quite small and the overall treatment effect is under-estimated for all three assumptions.

The magnitude of the overall bias decreases as μ increases, or as the effect estimated from the current data decreases. When $\mu=0$, all three assumptions have a little bit greater but still small bias. For the design and null assumptions, the bias declines with the power, and this is usually also true for the current trend assumption as long as μ or θ_A is not small. The bias increases with θ_A under the design and null assumption, but there is no clear trend with the current trend assumption.

The maximum bias occurs when $\theta_F + \mu = \left(\sqrt{2}c_\alpha - c_\gamma\right) \cdot \sigma\sqrt{2/n}$ and the maximum is $1/\sqrt{N\pi}$ for all three assumptions.

4.1.4 Probability of Stopping

Table 4a, 4b1, 4b2 and 4c shows the probability of stopping at the interim analysis. As can be anticipated from their corresponding conditional power that among the three assumptions, usually the CP under the design assumption is the largest and the CP under the null assumption is the smallest; hence with the design assumption there is the lowest probability of stopping, while with the null assumption there is the highest probability of stopping. These findings are reflected in their stopping thresholds, specifically that the design assumption has a very low stopping threshold, -while the null assumption has relatively high threshold, and more studies would stop at the interim analysis under the null assumption.

Generally, the probability of an early stop increases with θ_A and decreases as μ and the power to detect θ_A increases. If the true treatment effect μ is zero, the probability of an early stop is greater for all three assumptions.

4.1.5 Relative Weight

Table 5a, 5b1, 5b2 and 5c shows the expected relative weight of truncated studies in a meta-analysis calculated from equation (10).

Under the design assumption, this weight is smaller than under the null or current trend assumption. This is intuitively reasonable since under the design assumption usually a study has the highest conditional power and the highest stopping threshold compared to the other two assumptions, thus most of the time the study will continue to the final stage so that truncated studies provide a small proportion of the information in a meta-analysis.

The weight increases with θ_A and decreases as μ, the power to detect θ_A or the effect estimated from the current data gets great. When the effect size is actually zero, the truncated studies will have even more weight.

4.1.6 Summary

From Figure 1, we can see that the under-estimation for the studies that stop for futility with one planned interim analysis always becomes serious when the true treatment effect μ is great under all three assumptions.

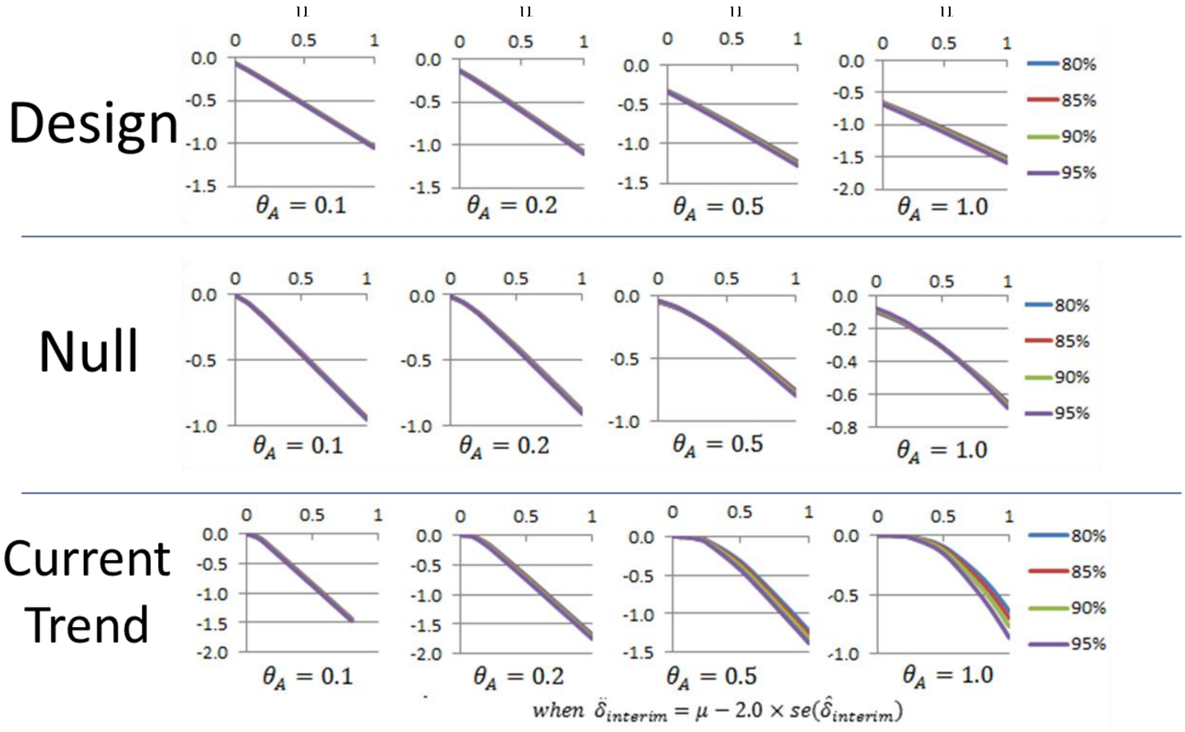

Figure 1: Under-estimation for the studies that stop for futility under three assumptions with one planned interim analysis

Note that when we change the power, most of the time (except when θ_A is large under the current design assumption) there is very little difference between different lines of powers, indicating that the effect of μ is much more important than the power and the power is not an important determinant here.

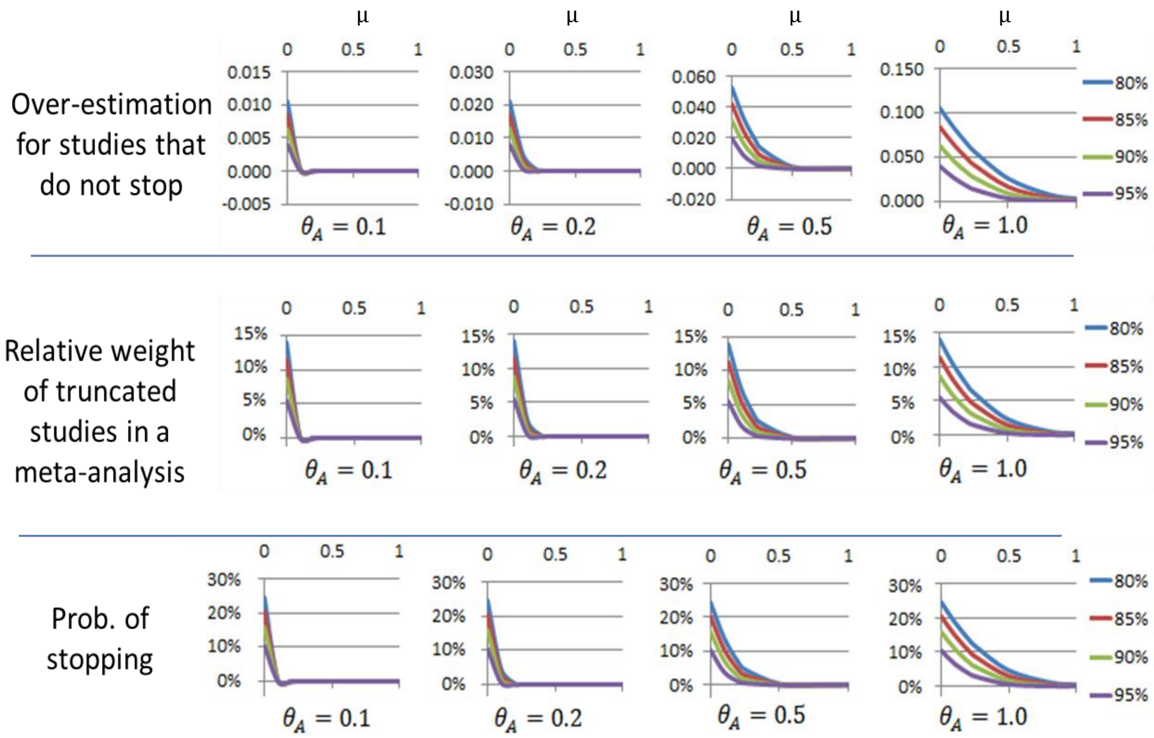

Figure 2: Over-estimation for non-tRCTs, relative weight of tRCTs and probability of stopping under the original design assumption with one planned interim analysis

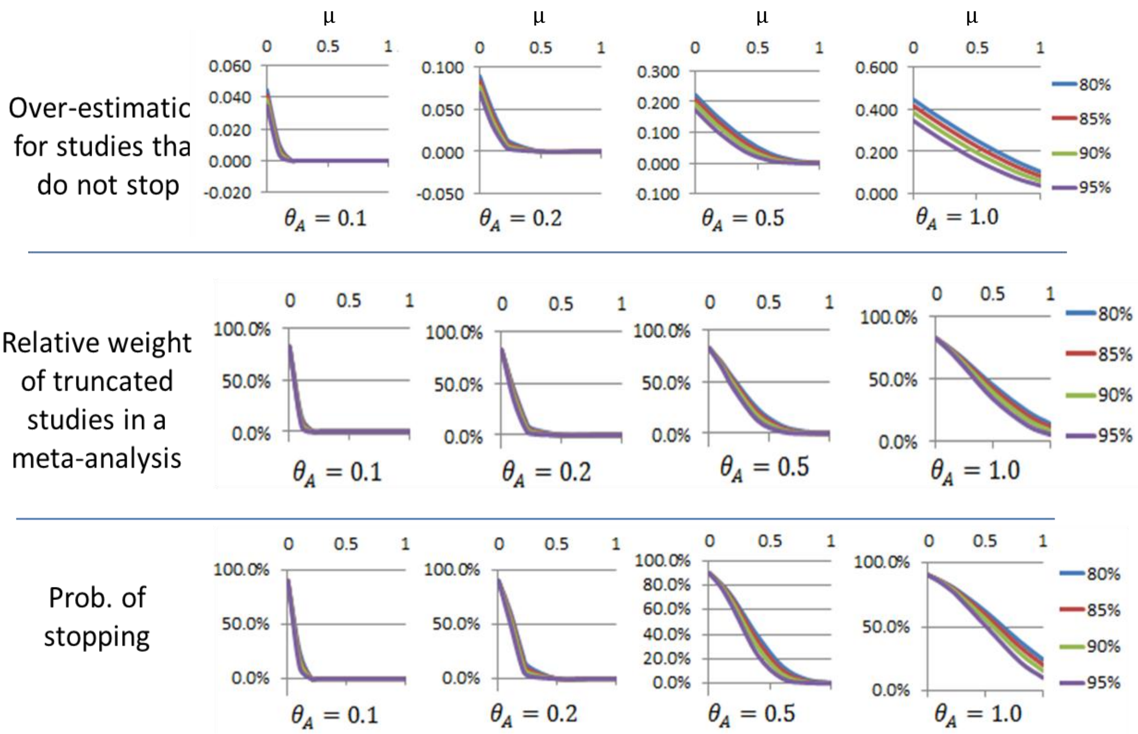

Figure 3: Over-estimation for non-tRCTs, relative weight of tRCTs and probability of stopping under the null assumption with one planned interim analysis

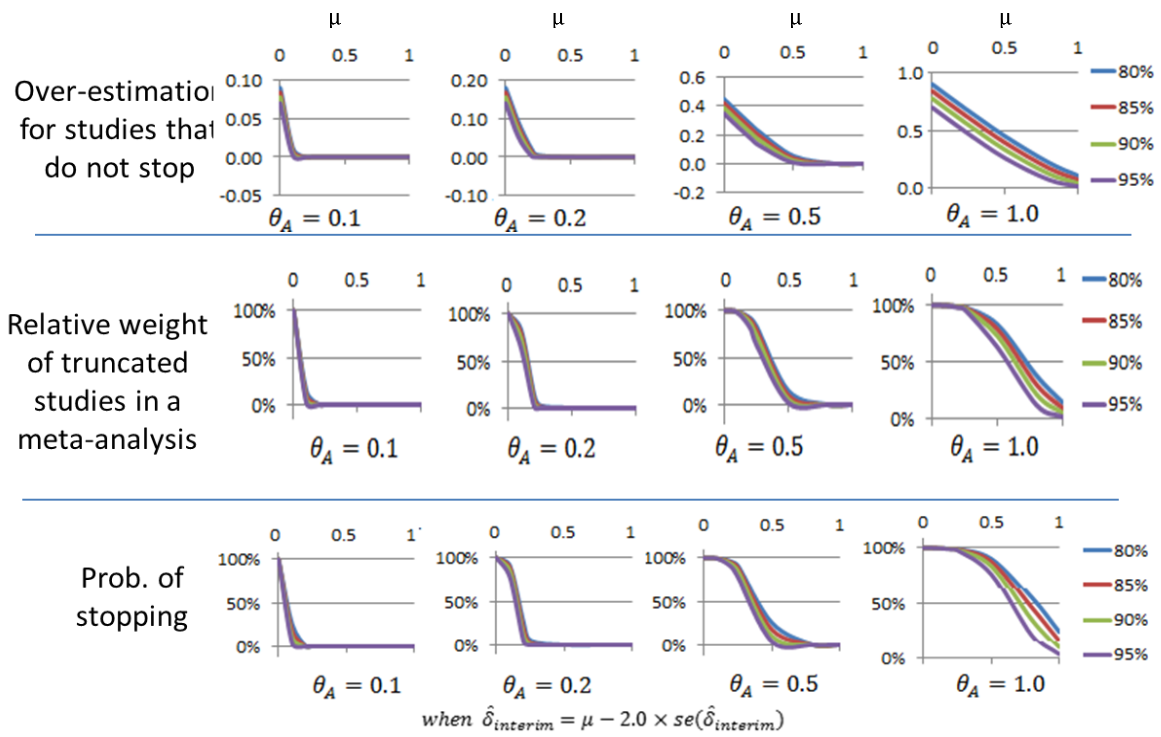

Figure 4: Over-estimation for non-tRCTs, relative weight of tRCTs and probability of stopping under the current trend assumption with one planned interim

From Figure 2, 3 and 4, we can see that the graphs of over-estimation for studies that do not stop, relative weight of truncated studies in a meta-analysis and probability of stopping have similar shapes, indicating that larger over-estimation for studies that do not stop corresponds to higher possibility of stopping, and thus truncated studies give more weight in a meta-analysis. Generally, over-estimation for studies that do not stop, relative weight of truncated studies in a meta-analysis and probability of stopping are all declines with the true treatment effect μ no matter what assumption we make for the future data.

Figure 5 shows that the maximum overall bias for a single study is always somewhat less than 0.1 for all three assumptions.

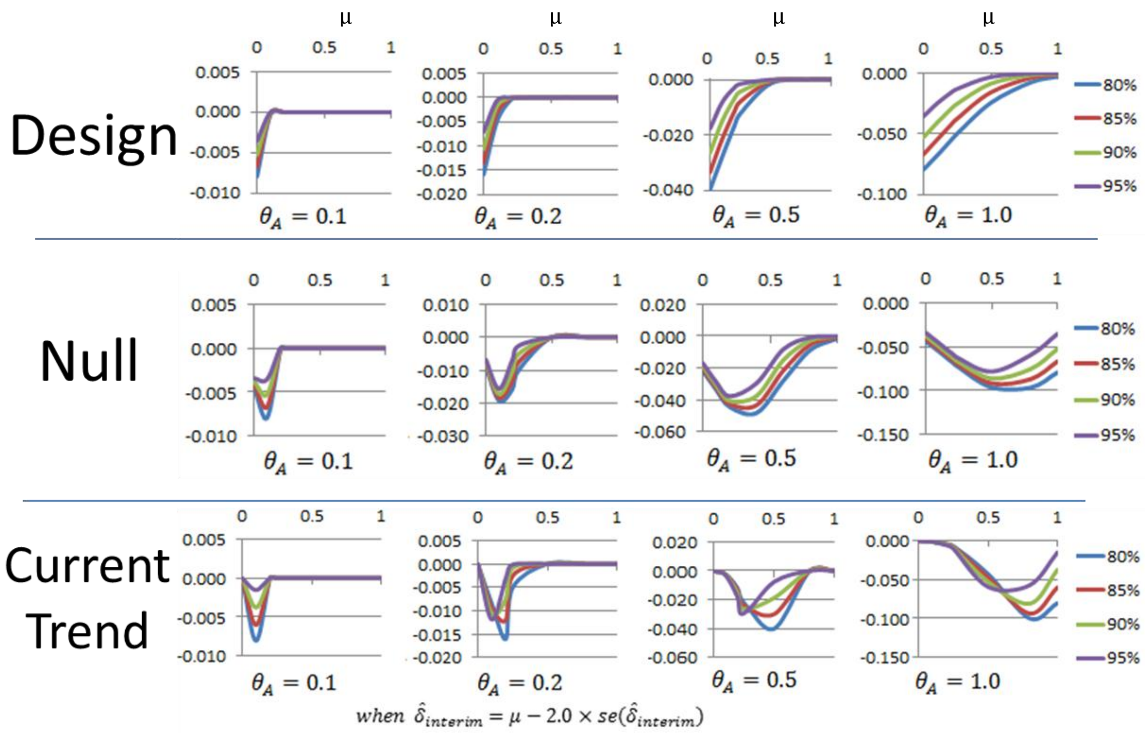

Figure 5: Overall bias for a single study under the three assumptions

4.2 Trials With Two Planned Interim Analysis, One-sided Testing

4.2.1 Error in Estimating the Treatment Effect in tRCTs

As expected the pattern is almost the same at the 1st interim analysis if two interim analyses had been planned. However, the degree of under-estimation is large at the 1st interim analysis if two interim analyses had been planned compared to studies with only one planned interim analysis if assuming the original design effect. This is mostly true if assuming the remaining data has the same effect trend as the current observed data as long as the true effect size is not small and the effect estimated from the current observed data is not negatively large. It is the opposite under the null assumption when the true effect size is not zero. For an example, if θ_A=0.1and μ=0.1, assuming the original design effect, with only one planned interim analysis the under-estimation at the 1st interim analysis is 0.1504, versus 0.2301 if two interim analyses are planned. While under the null assumption, the under-estimation is 0.0648 with one planned interim analysis versus 0.0589 with two planned interim analyses.

Under the null and current trend assumptions, with two interim analyses planned, at the 2nd interim analysis, the estimate error of the truncated study changes from positive value to negative value, i.e. from over-estimation to under-estimation, as μ, θ_A, the power to detect the minimally clinically meaningful treatment effect or the effect estimated from the current data increases. Under the design assumption, studies that stopped at the 2nd interim analysis always under-estimate the treatment effect and the under-estimation increases with μ, θ_A and the power to detect θ_A.

We also notice that the under-estimation of the truncated study at the 2nd interim analysis is smaller than the one at the 1st interim analysis if assuming the original design effect, corresponding to stringent (higher) stopping threshold at the 2nd interim analysis than at the 1st interim analysis.

4.2.2 Over-estimation of Non-Truncated RCTs and Overall Bias

With two planned interim analyses, similar to the 1st interim analysis, the non-truncated studies at the 2nd interim analysis, i.e. whose effect size exceed the 2nd interim stopping threshold thus continue to the final analysis, over-estimate the treatment effect, while the overall mean of a single study with a stopping rule for futility tends to under-estimate the treatment effect. These hold for all three assumptions.

The degree of the overall under-estimation is somewhat larger if two interim analyses are planned, compared to trials with only one. For instance, assuming the original effect, if θ_A=0.1 and μ=0.1, the overall under-estimation is 0.0003 with only one planned interim analysis, versus 0.0006 with two planned interim analyses (refer to Table 6).

4.2.3 Probability of Stopping

Assuming the original design effect, the probability of stopping early at the 1st interim analysis is quite low for trials with two planned interim analyses. This probability increases somewhat at the 2nd interim analysis if the true effect is non-zero. If the true effect is zero, the probability of stopping increases greatly at the 2nd interim analysis. For an example, if θ_A=0.1 and the power to detect this effect is 80%, the probabilities of stopping at the 1st interim analysis are 24.47% and 0.38% (Table 7) if the true effect size is 0 and 0.1 respectively with only one planned interim analysis. If two interim analyses are planned, the corresponding values are 3.20% and 0.03% at the 1st interim analysis, and 51.69% and 1.58% at the 2nd interim analysis respectively.

Under the null assumption, the probability of stopping early is quite high at the 1st interim analysis and surprisingly low at the 2nd interim analysis if two interim analyses are planned. Trials with two planned interim analysis have a little bit higher possibility to stop at the 1st interim analysis than trails with only one planned interim analysis. For instance, under the null assumption, if θ_A=0.1 and the power to detect this effect is 80%, the probabilities of stopping at the 1st interim analysis are 90.14% and 24.47% (Table 8) if μ=0 and 0.1 respectively with only one planned interim analysis. If two interim analyses are planned, the corresponding values are 91.67% and 40.74% at the 1st interim analysis, and 4.10% and 4.20% at the 2nd interim analysis respectively.

These facts indicate that the design assumption is a less stringent criterion at first and a more stringent one later, while it is the opposite for the null assumption. Hence with two planned interim analyses, under the original design assumption, most of the trials will continue after the 1st interim analysis but they are more likely to stop at the 2nd interim analysis; while under the null assumption, most of the trials will possibly stop at the 1st interim analysis, however once they survive from the 1st interim analysis, they are more likely to continue to the final analysis.

4.2.4 Relative Weight

If two interim analyses are planned, in meta-analyses, truncated studies occupy somewhat more weight than those with one planned interim analysis (Table 9A, 9B and 9C). This holds for all three assumptions.

4.2.5 Summary

From Figure 6-1 and 6-2, similar to trials with one planned interim analysis, at the 1^{st} and 2^{nd} interim analysis, the error in estimating the treatment effect in studies that stop for futility with two planned interim analyses always becomes serious when the true treatment effect μ is great under all three assumptions. And when we change the power, there is little difference between different lines of powers (except when θ_A is large under the current design assumption), indicating that the effect of μ is a much more important factor than the power.

From Figure 7-1 and 7-2, we can see that the over-estimation for studies that do not stop, relative weight of truncated studies in a meta-analysis and probability of stopping always decline with the true treatment effect μ under the original design assumption. The plots of these three outcomes again exhibit similar trend and patterns, i.e., larger over-estimation for non-truncated studies corresponds to higher possibility of stopping, and thus truncated studies give more weight in a meta-analysis.

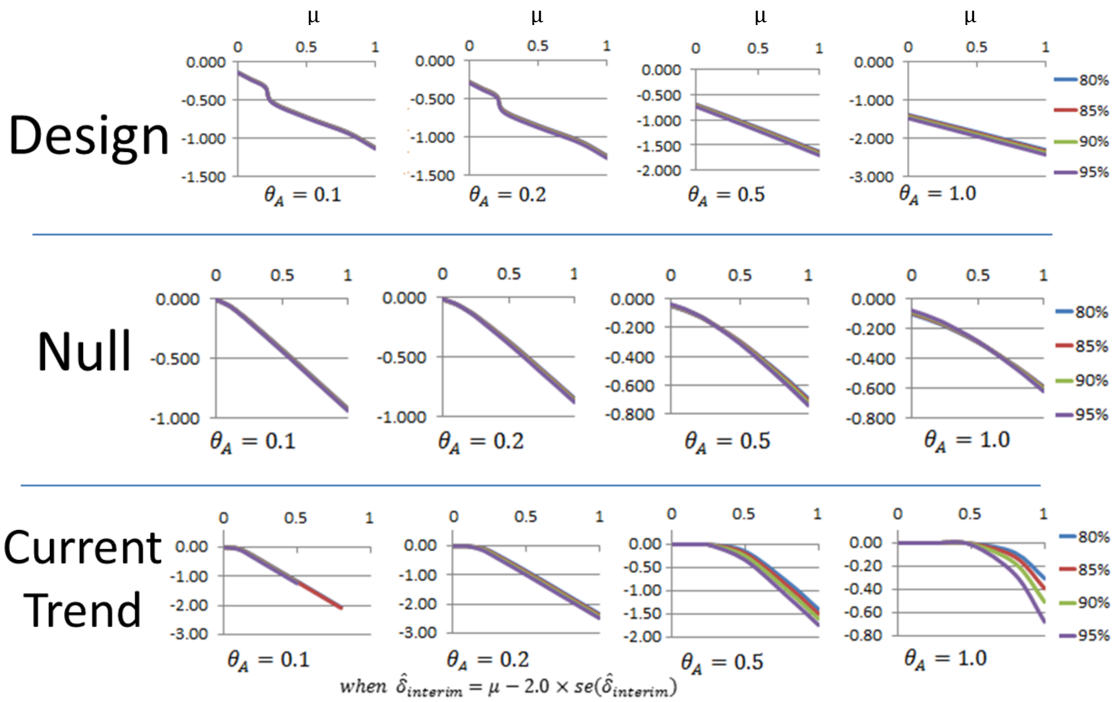

Figure 6-1: Errors in estimating the treatment effect in tRCTs under the three assumptions with two planned interim analysis, at the 1st interim analysis

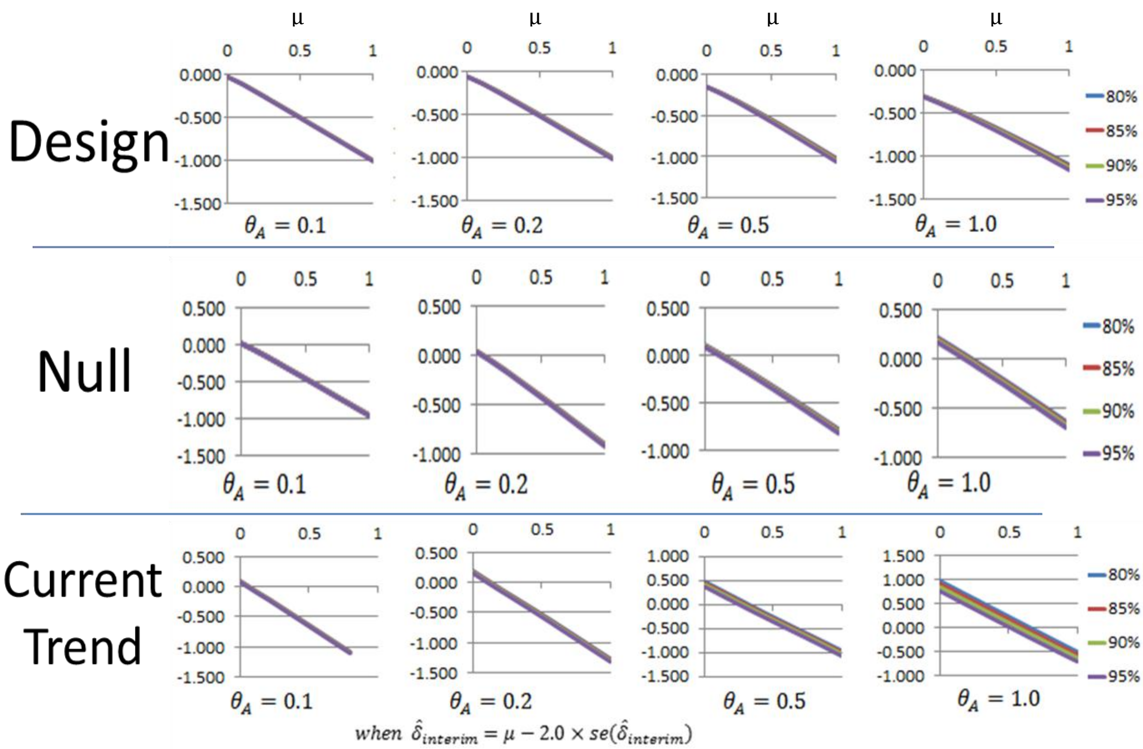

Figure 6-2: Errors in estimating the treatment effect in tRCTs under the three assumptions with two planned interim analysis, at the 2nd interim analysis

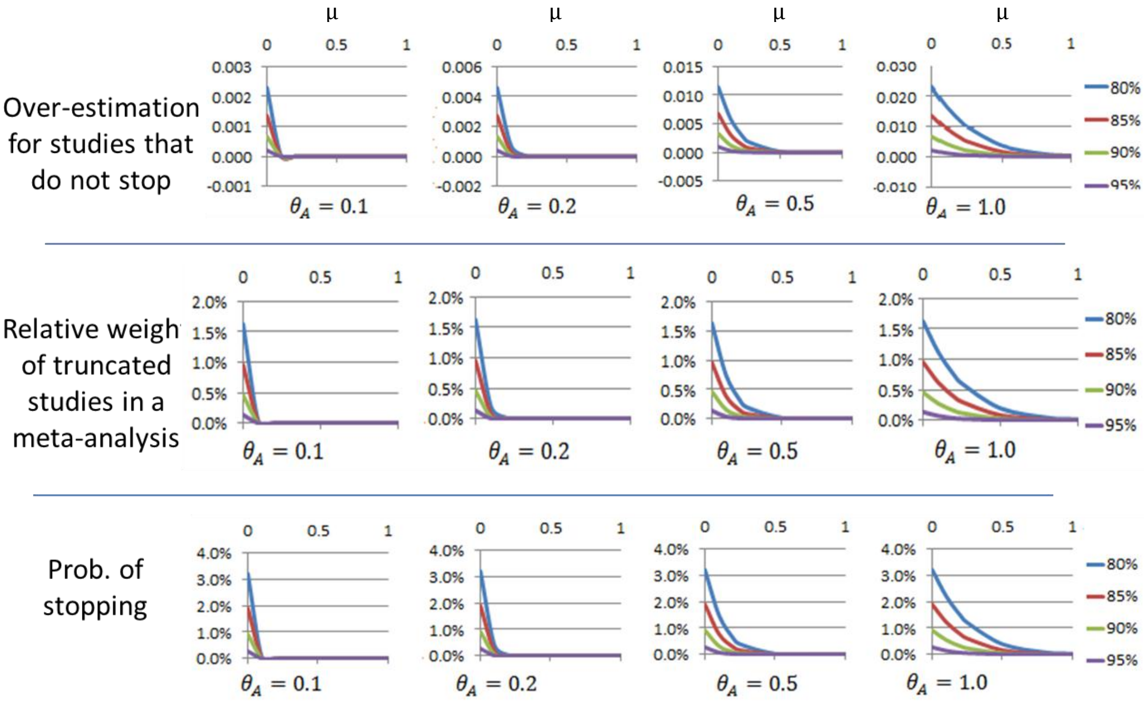

Figure 7-1: Over-estimation for non-tRCTs, relative weight of tRCTs and probability of stopping under the original design assumption with two planned interim analysis, at the 1st interim analysis

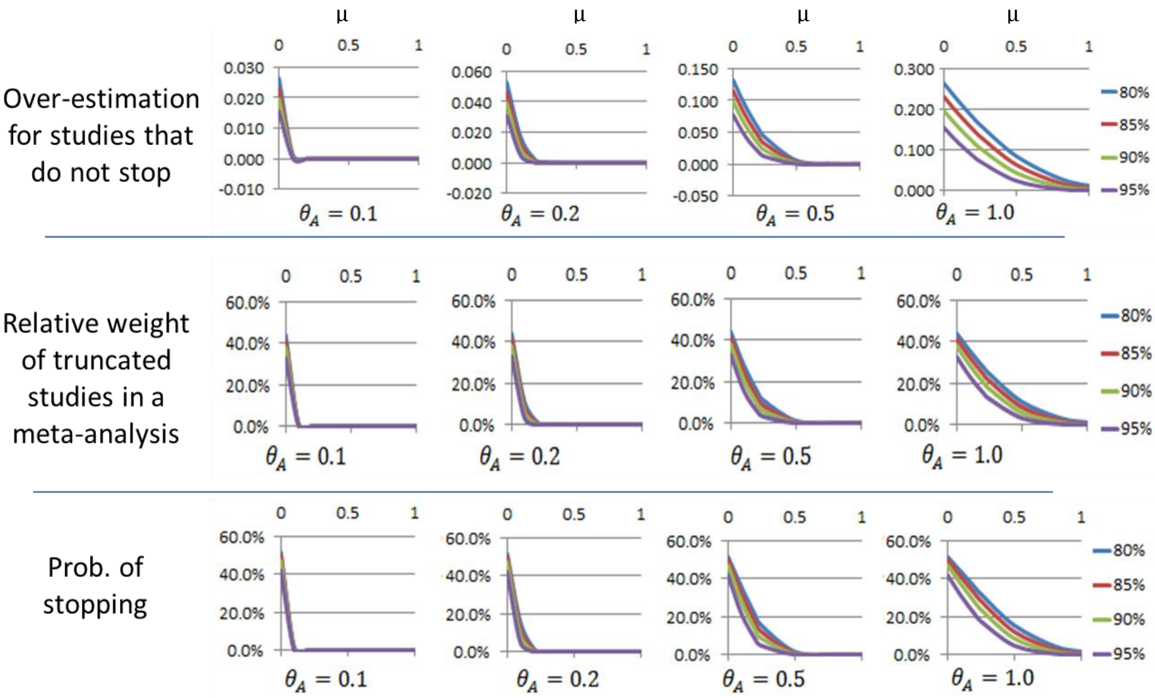

Figure 7-2: Over-estimation for non-tRCTs, relative weight of tRCTs and probability of stopping under the original design assumption with two planned interim analysis, at the 2nd interim analysis

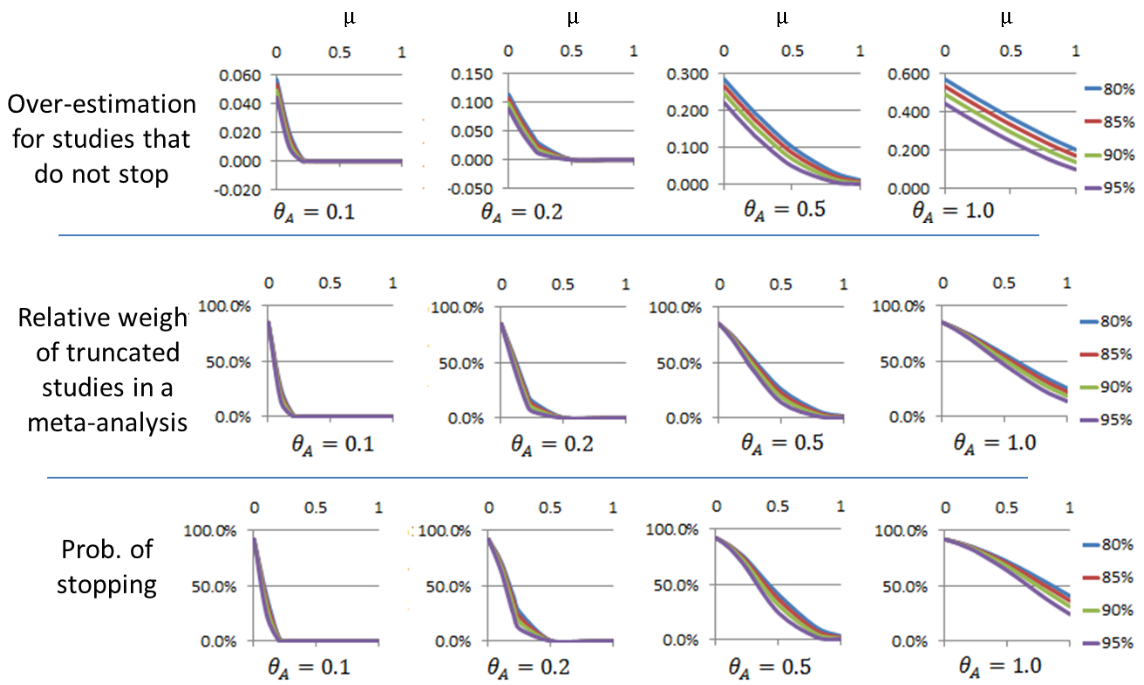

Figure 8-1: Over-estimation for non-tRCTs, relative weight of tRCTs and probability of stopping under the null assumption with two planned interim analysis, at the 1st interim analysis

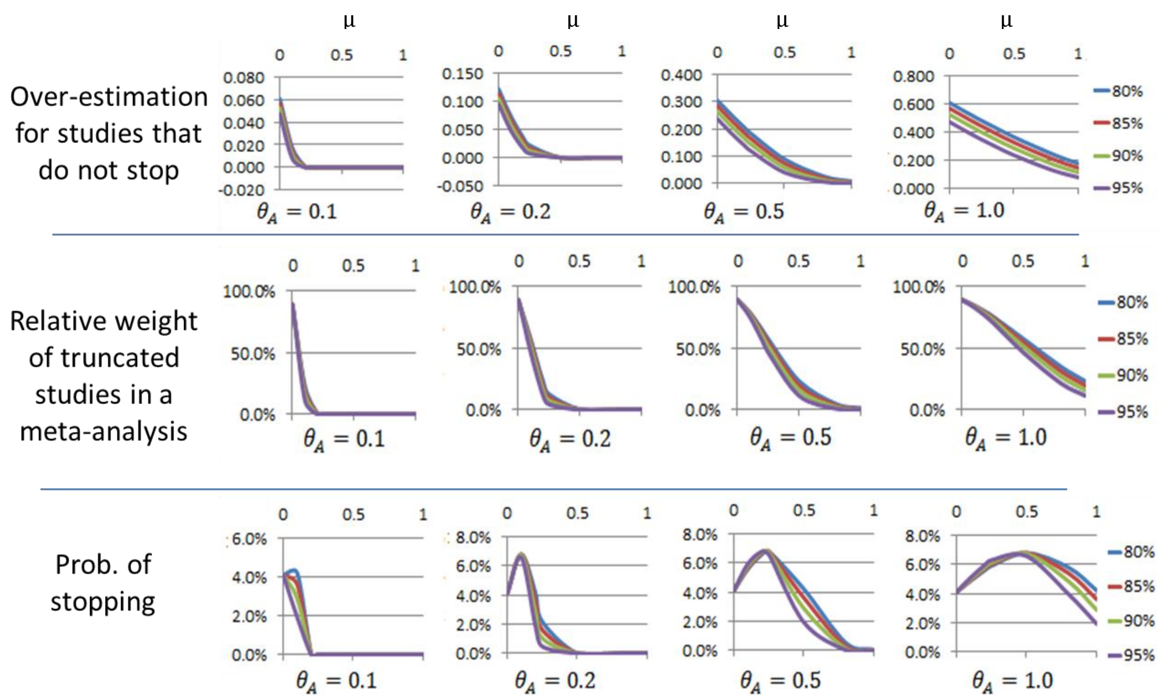

Figure 8-2: Over-estimation for non-tRCTs, relative weight of tRCTs and probability of stopping under the null assumption with two planned interim analysis, at the 2nd interim analysis

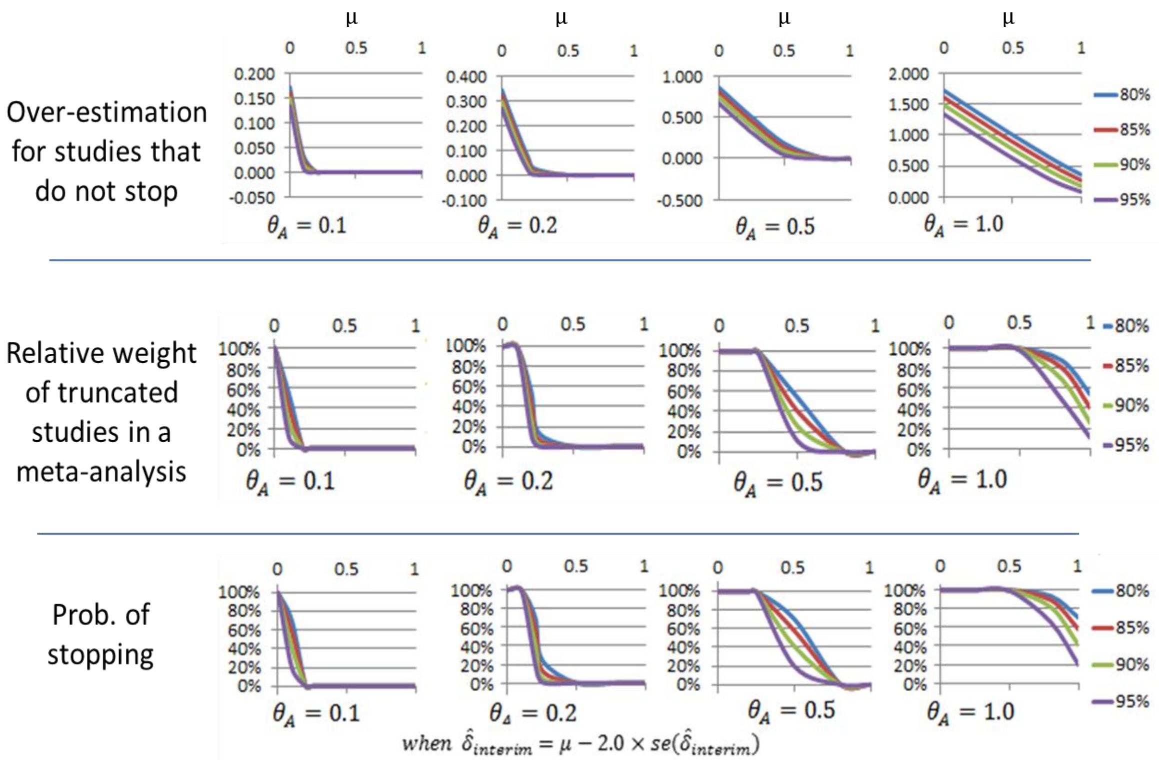

Figure 9-1: Over-estimation for non-tRCTs, relative weight of tRCTs and probability of stopping under the current trend assumption with two planned interim analysis, at the 1st interim analysis

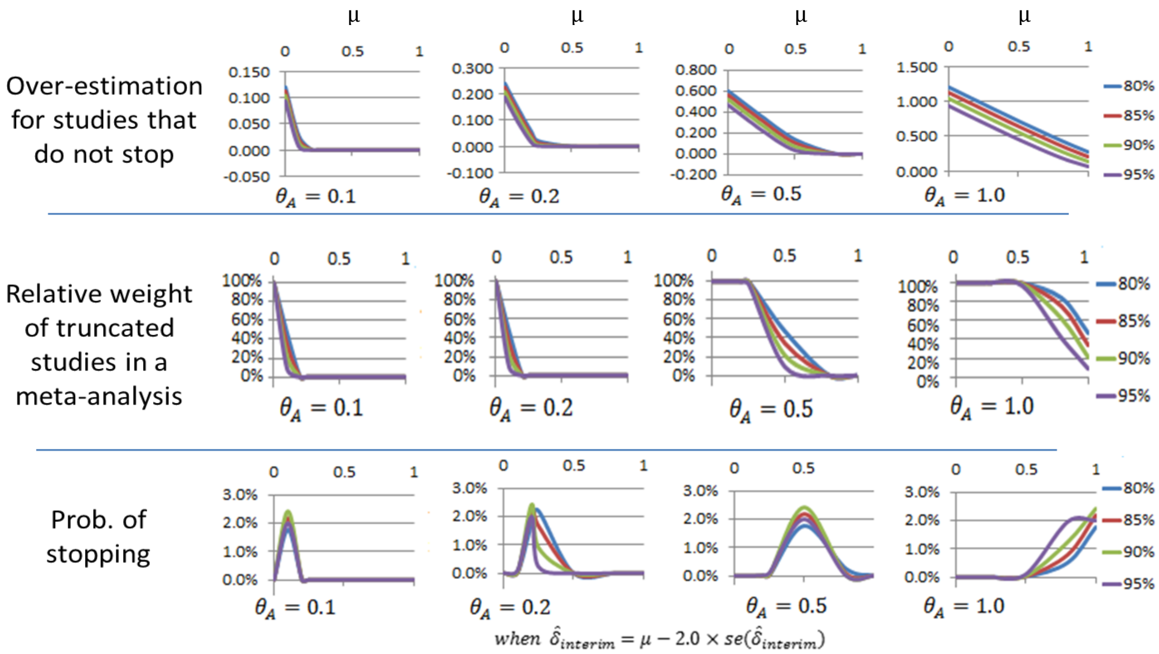

Figure 9-2: Over-estimation for non-tRCTs, relative weight of tRCTs and probability of stopping under the current trend assumption with two planned interim analysis, at the 2nd interim analysis

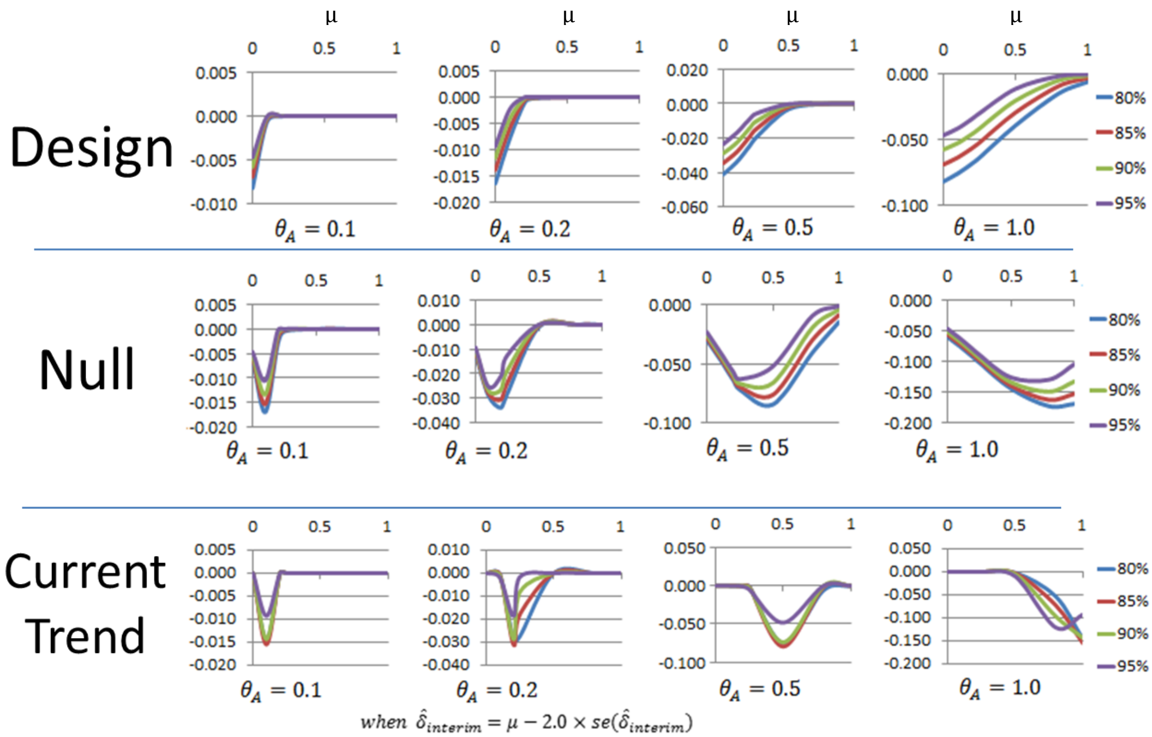

Figure 10: Overall bias for a single study under the three assumptions with two planned interim analyses

From Figure 8-1, 8-2, 9-1 and 9-2, we can see that for the null and current trend assumptions, generally it is true that the over-estimation for studies that do not stop and relative weight of truncated studies in a meta-analysis decline with the true treatment effect μ at the 1st interim analysis or the 2nd interim analysis.

However probability of stopping declines with μ at the 1st interim analysis rather than the 2nd interim analysis under the null or current trend analysis.

Figure 10 shows that the maximum overall bias for a single study with two planned interim analyses varies with the assumptions about the future data. The original design assumption has smaller maximum overall bias which is less than 0.1, the null and current trend assumption have bigger maximum overall bias which is somewhat less than 0.2.

Chapter 5

Illustrative Examples

5.1 An Example of One Interim Analysis, Two-Sided Testing

According to Carducci et al[14], a phase 3 randomized, multinational, double-blind, placebo-controlled trial was conducted to evaluate the efficacy and safety of atrasentan, a selective endothelin－receptor antagonist in patients with metastatic hormone-refractory prostate cancer (HRPC). The primary endpoint was time to disease progression in the intent-to-treat population, which was defined as the first occurrence of any one of the events designated. The protocol estimated 650 events would be needed to achieve 90% power at the 2-sided 0.05 level of significance to detect a treatment difference with a magnitude 0.77 in the hazards ratio (HR). The authors didn’t mention if there was any interim analysis and how many interim analyses there would be, but based on the fact that the study was decided by an independent data monitoring committee to be stopped at 343 confirmed events plus additional events not yet adjudicated, a reasonable guess is that this trial was planned with a single interim analysis at the halfway point, after half of a total planned 650 events had occurred. When the study stopped with 343 confirmed events, the observed hazard ratio was 0.89.

For analytic simplicity, we express our results in standardized normal units, as we did in methods section. We will interpret an effect size as small, medium, large and so on. Cohen[15] advised to classify effect size as small (d=0.2), medium (d=0.5) and large (d=0.8). How to interpret an effect size depends on its substantive context and its operational definition[16]. It is a bit arbitrary, but in this paper we regard values below 0.10 as "trivial" effects, within .10 to .29 range as "small" effects, values from .30 to .49 as "medium" effects, values from .50 to .79 as "large" effects and values exceeding .80 as "very large" effects.

Using the formulae in section 3.3, if the projected original hazard ratio of 0.77 was correct, with one planned interim analysis a study that is stopped at half of the target number of events would give an expected under-estimation of 0.1396. 0.2094 or 0.3414 (Table 10) corresponding to the null, current trend and original design assumptions respectively. It shows that the calculation under the null assumption gives the smallest bias and smallest conditional power. According to our effect size grading, truncated studies with this protocol will under-estimate the true effect size by a small amount under the null and current trend assumptions, and a medium amount under the original design assumption, e.g. 0.1396, 0.2094 and 0.3414 respectively. Compared with the test statistic of Z_t=1.0791 at the interim analysis, the truncated study under-estimates the effect size by 12.93%, 19.41% and 31.64%. If the study had not stopped, the expected over-estimation at the final analysis would be 0.0443, 0.0145 and 0.0007 respectively, which means a non-truncated study would over-estimate the true effect by 4.11%, 1.34% and 0.07%. These are all small as we anticipated. The probability of actually stopping early are 24.11%, 6.46% and 0.21%, which means most of them will not stop early and should carry on to the end even under the null assumption. That leaves very small deviations for non-stopped studies. The values of overall under-estimation with this design are 0.0168, 0.0068 and 0.0004 and the relative weights of such truncated studies in a meta-analysis are 13.71%, 3.34% and 0.10%.

Hence from this example, we can see if the trial had stopped, we should be concerned that we would under-estimate the true effect size by a possibly important amount.

5.2 Another Example of One Interim Analysis, One-Sided Testing

Champlin et al.[17] reported a randomized controlled study of conditioning regimens to determine whether the addition of antithymocyte globulin (ATG) to cyclophosphamide (Cy) would improve survival at 1 year from 65% to 85%, which is the main endpoint, and lower the rate of graft rejection from the anticipated 15% to less than 5%. The study planned to accrue 224 patients, who have severe aplasitc anemia and need bone marrow transplantation, based on 90% power at the 0.05% level to detect a 20% difference in overall survival and 80% power to detect a 15% decrease in the rate of graft failure. An interim analysis was conducted on 130 eligible patients, with 60 in Cy group and 70 in Cy+AGT group, with a 1 year survival rate of 83% in Cy versus 89% in Cy+ATG group. The study was closed after the interim analysis.

Using the formulae in section 3.1, if the projected effect size of 65% versus 85% was true, a study truncated at a single planned interim analysis at designated portion of sample size would give under-estimation of 0.1410, 0.1508 and 0.1739 (Table 11) assuming the null, current trend and original design effect respectively, which are all small according to our grading. If the study had not stopped, it would be expected to over-estimate the true effect by 0.0020, 0.0014 and 0.0005 assuming the null, current trend and original design effect respectively. The overall over-estimation is 0.0019, 0.0013 and 0.0005, which are all trivial effects, and the probability of stopping early is 2.74%, 1.79% and 0.59% assuming the null, current trend and original design effect respectively, which means most of them will not stop even under the null assumption in this case. Corresponding to the small possibilities of stopping early, the truncated studies provide a small proportion in a meta-analysis, numerically 1.39%, 0.90% and 0.30% respectively.

Chapter 6

Summary and Limitations

The investigation on the bias of trials with stopping rules for futility should be treated with the same importance as that with stopping rules for efficacy. One of the most commonly used methods for futility stopping is based on conditional power, which is defined as the probability of rejecting the null hypothesis at the final analysis given the currently observed data and a hypothesized true treatment effect for the future data[5]. The assumption we make about the future data affects how the bias varies.

With one planned interim analysis, truncated studies that stop early at the interim analysis are expected to under-estimate the true effect by a substantial amount, which always increases with the true effect size, no matter what assumption we make. Serious under-estimation can occur if the true effect size or the minimally clinically meaningful treatment effect θ_A is large. But this is unlikely to occur because the probability of stopping decreases with the true effect size or θ_A, i.e. less chance the study has to stop, the larger the under-estimation is. The non-truncated studies that do not stop early and thus continue to the final analysis tend to over-estimate the true effect under all three assumptions. The over-estimation always increases with θ_A and declines with the true effect, the power to detect θ_A or the effect observed from the current data (if assuming the effect size of the future data is the same trend as in the observed data). It becomes substantially large when the true effect size is small and θ_A is large (θ_A>0.2).

The overall bias in the estimated treatment effect from a single study with one planned interim analysis is relatively small. Thus, to quantify the true effect, a reasonable evaluation of the true effect size can be made by taking a weighted average of truncated and non-truncated studies, with the respective weights being the probabilities of stopping or not.

If two interim analyses are planned, the pattern at the 1st interim analysis is generally the same as the one with only one planned interim analysis. However, the degree of under-estimation or over-estimation at the 1st interim analysis is considerable if two interim analyses are planned under the null and current trend assumptions. At the 2nd interim analysis, the estimation error in truncated studies can change from positive value to negative value, i.e. from over-estimation to under-estimation, while the non-truncated studies still tend to over-estimate the true effect. The overall under-estimation is again small but relatively larger than the one with one planned interim analysis. So in other words, there is a penalty to pay if an additional interim analysis is planned. For the design assumption, a trial has higher chance to stop at the 2nd interim analysis than at the 1st interim analysis, while it is the opposite for the null assumption case.

Since the degree of bias associated with a trial with stopping rules for futility varies with how we design the analysis protocol, in order to make appropriate inferences on the trial results, we need to have detailed information about the protocol, including whether there are any planned interim analyses, and if so, how many, whether there is any stopping rule and if so, details of the rule. The more we know about the protocol, the more accurately we can assess the possibility of bias, over- or under-estimation of the treatment effect. One should conduct a trial according to its pre-planned protocol. However, protocols are frequently not obeyed in practice. Briel et al.[17] indicated that, by investigating a cohort of RCT protocols approved between 2000 and 2003 by six research ethics committees in Switzerland, Germany and Canada, few (15%) of prematurely discontinued RCTs were stopped “on the basis of a matching interim analysis and stopping rule”, most of RCTs stopped early “independent of pre-planned interim analyses and stopping rules”, and “ discrepancies in the reporting of interim analyses, stopping rules and data safety and monitoring boards(DSMBs) between protocols and publications were common.” Based on these facts, we suggest that when trials are stopped early, interim analysis plans, stopping guides and DSMBs should increase transparency and credibility by fully reporting trials both with respect to the protocol (SPIRIT initiative, where SPIRIT stands for the Standard Protocol Items Recommendations for Interventional Trials) and in publications of results (CONSORT – the Consolidated Standards of Reporting Trials)[17].

There are two limitations of this paper, similar to those stated in the paper of Walter et al.[5] One is that the derivations of the formulae only cover trials with one or two planned interim analyses, but the algebraic approach can be extendedly apply to trials with more planned interim analyses. Based on the results of comparisons of one planned interim analysis with two planned interim analyses, it is expected that the overall bias of a single study would increase with the number of planned interim analyses.

The other limitation is that our results are based on assuming the outcome variable is normally distributed. However, if sample sizes are big enough, studies with binomial or survival endpoints can also be treated as approximately normally distributed, as we showed in our examples. Hence, accordingly we expect that our analytic results will also apply to other types of summary endpoints as long as their distributions are approximately normal.

Appendix

Tables

Table 1a: Expected under-estimation of treatment effect in studies that stop at a single planned interim analysis, under the design assumption (one-sided testing)

True Mean (when θ_A=0.1)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0648	-0.1504	-0.2449	-0.2933	-0.5396	-0.8379	-1.0373
0.85	-0.0656	-0.1527	-0.2478	-0.2965	-0.5432	-0.8418	-1.0413
0.9	-0.0666	-0.1552	-0.2511	-0.3000	-0.5472	-0.8460	-1.0455
0.95	-0.0681	-0.1586	-0.2553	-0.3044	-0.5522	-0.8512	-1.0509

True Mean (when θ_A=0.2)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.1296	-0.2115	-0.3009	-0.3472	-0.5867	-0.8812	-1.0792
0.85	-0.1312	-0.2147	-0.3053	-0.3521	-0.5929	-0.8882	-1.0864
0.9	-0.1332	-0.2186	-0.3105	-0.3577	-0.5999	-0.8959	-1.0944
0.95	-0.1362	-0.2239	-0.3172	-0.3650	-0.6088	-0.9056	-1.1044

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.1620	-0.2429	-0.3306	-0.3761	-0.6123	-0.9045	-1.1015
0.85	-0.1640	-0.2466	-0.3357	-0.3817	-0.6196	-0.9128	-1.1102
0.9	-0.1666	-0.2511	-0.3415	-0.3881	-0.6278	-0.9221	-1.1199
0.95	-0.1703	-0.2572	-0.3493	-0.3965	-0.6382	-0.9337	-1.1319

True Mean (when θ_A=0.5)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.3240	-0.4027	-0.4858	-0.5287	-0.7522	-1.0331	-1.2246
0.85	-0.3280	-0.4085	-0.4932	-0.5368	-0.7633	-1.0466	-1.2391
0.9	-0.3331	-0.4156	-0.5021	-0.5465	-0.7762	-1.0618	-1.2555
0.95	-0.3406	-0.4257	-0.5144	-0.5597	-0.7931	-1.0815	-1.2765

True Mean (when θ_A=0.8)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.5185	-0.5962	-0.6770	-0.7184	-0.9332	-1.2036	-1.3889
0.85	-0.5248	-0.6044	-0.6870	-0.7292	-0.9476	-1.2213	-1.4084
0.9	-0.5330	-0.6147	-0.6991	-0.7422	-0.9645	-1.2419	-1.4308
0.95	-0.5450	-0.6293	-0.7162	-0.7604	-0.9874	-1.2689	-1.4600

True Mean (when θ_A=1.0)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.6481	-0.7255	-0.8055	-0.8463	-1.0573	-1.3225	-1.5044
0.85	-0.6560	-0.7353	-0.8170	-0.8587	-1.0736	-1.3426	-1.5267
0.9	-0.6662	-0.7476	-0.8313	-0.8739	-1.0930	-1.3661	-1.5523
0.95	-0.6812	-0.7653	-0.8514	-0.8952	-1.1195	-1.3973	-1.5862

Table 1b-1: Expected under-estimation of treatment effect
in studies that stop at a single planned interim analysis (one-sided testing),
under the current trend assumption when $\hat{\delta}_{\text{interim}} = \mu - 2.0 \times se(\hat{\delta}_{\text{interim}})$

True Mean (when θ_A=0.1)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0001	-0.0641	-0.2440	-0.3413	-0.8370	-1.4357	-1.8353
0.85	-0.0001	-0.0700	-0.2533	-0.3510	-0.8474	-1.4463	#DIV/0!
0.9	-0.0001	-0.0770	-0.2635	-0.3617	-0.8587	-1.4578	#DIV/0!
0.95	-0.0001	-0.0865	-0.2764	-0.3750	-0.8727	-1.4720	#DIV/0!

True Mean (when θ_A=0.2)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0002	-0.0189	-0.1282	-0.2098	-0.6825	-1.2758	-1.6739
0.85	-0.0002	-0.0216	-0.1400	-0.2246	-0.7019	-1.2963	-1.6947
0.9	-0.0002	-0.0252	-0.1541	-0.2416	-0.7233	-1.3187	-1.7174
0.95	-0.0001	-0.0309	-0.1730	-0.2636	-0.7500	-1.3464	-1.7454

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0002	-0.0123	-0.0912	-0.1602	-0.6100	-1.1980	-1.5948
0.85	-0.0002	-0.0140	-0.1018	-0.1750	-0.6331	-1.2231	-1.6204
0.9	-0.0002	-0.0163	-0.1148	-0.1926	-0.6588	-1.2506	-1.6484
0.95	-0.0002	-0.0200	-0.1332	-0.2162	-0.6910	-1.2848	-1.6830

True Mean (when θ_A=0.5)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0004	-0.0045	-0.0246	-0.0473	-0.3204	-0.8403	-1.2200
0.85	-0.0004	-0.0048	-0.0279	-0.0539	-0.3501	-0.8831	-1.2663
0.9	-0.0004	-0.0053	-0.0325	-0.0629	-0.3852	-0.9311	-1.3176
0.95	-0.0003	-0.0060	-0.0400	-0.0772	-0.4324	-0.9922	-1.3820

True Mean (when θA=0.8)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0007	-0.0033	-0.0116	-0.0200	-0.1461	-0.5126	-0.8392
0.85	-0.0007	-0.0033	-0.0127	-0.0224	-0.1659	-0.5601	-0.8985
0.9	-0.0006	-0.0035	-0.0142	-0.0256	-0.1915	-0.6163	-0.9663
0.95	-0.0006	-0.0037	-0.0168	-0.0311	-0.2301	-0.6919	-1.0546

True Mean (when θA=1.0)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0009	-0.0031	-0.0090	-0.0145	-0.0945	-0.3650	-0.6408
0.85	-0.0008	-0.0031	-0.0096	-0.0159	-0.1079	-0.4072	-0.7002
0.9	-0.0008	-0.0032	-0.0105	-0.0178	-0.1259	-0.4593	-0.7703
0.95	-0.0007	-0.0033	-0.0120	-0.0210	-0.1544	-0.5330	-0.8649

Table 1b-2: Expected under-estimation of treatment effect in studies that stop at a single planned interim analysis (one-sided testing), under the current trend assumption when θA=0.25

$$when\ \ \hat{\delta}_{interim} = \mu - 2.0 \times se(\hat{\delta}_{interim})$$

True Mean (when θA=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0002	-0.0123	-0.0912	-0.1602	-0.6100	-1.1980	-1.5948
0.85	-0.0002	-0.0140	-0.1018	-0.1750	-0.6331	-1.2231	-1.6204
0.9	-0.0002	-0.0163	-0.1148	-0.1926	-0.6588	-1.2506	-1.6484
0.95	-0.0002	-0.0200	-0.1332	-0.2162	-0.6910	-1.2848	-1.6830

$$when\ \ \hat{\delta}_{interim} = \mu - 1.5 \times se(\hat{\delta}_{interim})$$

True Mean (when θA=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0010	-0.0275	-0.1331	-0.2101	-0.6709	-1.2605	-1.6575
0.85	-0.0010	-0.0300	-0.1434	-0.2235	-0.6905	-1.2816	-1.6791
0.9	-0.0009	-0.0332	-0.1558	-0.2391	-0.7121	-1.3048	-1.7027
0.95	-0.0008	-0.0382	-0.1727	-0.2598	-0.7392	-1.3336	-1.7319

$$when\ \hat{\delta}_{interim} = \mu - 1.0 \times se(\hat{\delta}_{interim})$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0037	-0.0517	-0.1804	-0.2633	-0.7322	-1.3230	-1.7202
0.85	-0.0035	-0.0545	-0.1895	-0.2746	-0.7480	-1.3401	-1.7378
0.9	-0.0032	-0.0580	-0.2002	-0.2876	-0.7655	-1.3590	-1.7570
0.95	-0.0029	-0.0633	-0.2145	-0.3046	-0.7875	-1.3824	-1.7808

$$when\ \hat{\delta}_{interim} = \mu - 0.5 \times se(\hat{\delta}_{interim})$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0105	-0.0848	-0.2318	-0.3189	-0.7937	-1.3855	-1.7830
0.85	-0.0099	-0.0871	-0.2389	-0.3275	-0.8058	-1.3987	-1.7965
0.9	-0.0091	-0.0901	-0.2471	-0.3373	-0.8191	-1.4132	-1.8113
0.95	-0.0082	-0.0945	-0.2581	-0.3502	-0.8358	-1.4312	-1.8297

$$when\ \hat{\delta}_{interim} = \mu$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0243	-0.1255	-0.2861	-0.3761	-0.8555	-1.4482	-1.8458
0.85	-0.0227	-0.1266	-0.2905	-0.3817	-0.8637	-1.4573	-1.8553
0.9	-0.0210	-0.1280	-0.2958	-0.3881	-0.8728	-1.4674	-1.8657
0.95	-0.0189	-0.1303	-0.3028	-0.3965	-0.8843	-1.4800	-1.8786

$$when\ \hat{\delta}_{interim} = \mu + 0.5 \times se(\hat{\delta}_{interim})$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0469	-0.1721	-0.3425	-0.4346	-0.9174	-1.5108	-1.9086
0.85	-0.0439	-0.1711	-0.3439	-0.4368	-0.9217	-1.5159	-1.9141
0.9	-0.0406	-0.1702	-0.3457	-0.4396	-0.9265	-1.5216	-1.9200
0.95	-0.0365	-0.1695	-0.3484	-0.4433	-0.9327	-1.5288	-1.9275

$$when\ \ \hat{\delta}_{interim} = \mu + 1.0 \times se(\hat{\delta}_{interim})$$

True Mean (when θA=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0786	-0.2229	-0.4003	-0.4940	-0.9794	-1.5735	-1.9715
0.85	-0.0735	-0.2193	-0.3984	-0.4927	-0.9798	-1.5746	-1.9728
0.9	-0.0680	-0.2155	-0.3966	-0.4916	-0.9804	-1.5759	-1.9744
0.95	-0.0611	-0.2112	-0.3947	-0.4905	-0.9813	-1.5777	-1.9764

$$when\ \ \hat{\delta}_{interim} = \mu + 1.5 \times se(\hat{\delta}_{interim})$$

True Mean (when θA=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.1182	-0.2768	-0.4591	-0.5540	-1.0416	-1.6362	-2.0343
0.85	-0.1105	-0.2702	-0.4538	-0.5492	-1.0381	-1.6333	-2.0316
0.9	-0.1021	-0.2631	-0.4482	-0.5440	-1.0343	-1.6302	-2.0288
0.95	-0.0918	-0.2546	-0.4415	-0.5380	-1.0299	-1.6265	-2.0254

$$when\ \ \hat{\delta}_{interim} = \mu + 2.0 \times se(\hat{\delta}_{interim})$$

True Mean (when θA=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.1638	-0.3328	-0.5188	-0.6146	-1.1039	-1.6990	-2.0972
0.85	-0.1532	-0.3229	-0.5099	-0.6060	-1.0964	-1.6920	-2.0904
0.9	-0.1416	-0.3122	-0.5003	-0.5968	-1.0883	-1.6845	-2.0832
0.95	-0.1273	-0.2993	-0.4887	-0.5857	-1.0785	-1.6754	-2.0743

Table 1c: Expected under-estimation of treatment effect in studies that stop at a single planned interim analysis (one-sided testing), under the null assumption

True Mean (when θ_A=0.1)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0097	-0.0648	-0.1504	-0.1971	-0.4406	-0.7383	-0.9376
0.85	-0.0091	-0.0656	-0.1527	-0.1998	-0.4441	-0.7421	-0.9415
0.9	-0.0084	-0.0666	-0.1552	-0.2027	-0.4479	-0.7463	-0.9457
0.95	-0.0076	-0.0681	-0.1586	-0.2066	-0.4528	-0.7514	-0.9510

True Mean (when θ_A=0.2)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0194	-0.0624	-0.1296	-0.1693	-0.3943	-0.6844	-0.8812
0.85	-0.0182	-0.0620	-0.1312	-0.1717	-0.3995	-0.6909	-0.8882
0.9	-0.0168	-0.0618	-0.1332	-0.1748	-0.4055	-0.6982	-0.8959
0.95	-0.0151	-0.0618	-0.1362	-0.1790	-0.4132	-0.7074	-0.9056

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0243	-0.0644	-0.1255	-0.1620	-0.3761	-0.6606	-0.8555
0.85	-0.0227	-0.0636	-0.1266	-0.1640	-0.3817	-0.6681	-0.8637
0.9	-0.0210	-0.0629	-0.1280	-0.1666	-0.3881	-0.6765	-0.8728
0.95	-0.0189	-0.0622	-0.1303	-0.1703	-0.3965	-0.6872	-0.8843

True Mean (when θ_A=0.5)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0486	-0.0830	-0.1289	-0.2511	-0.4027	-0.5722	-0.7522
0.85	-0.0455	-0.0802	-0.1273	-0.2531	-0.4085	-0.5811	-0.7633
0.9	-0.0420	-0.0773	-0.1258	-0.2560	-0.4156	-0.5915	-0.7762
0.95	-0.0378	-0.0738	-0.1244	-0.2606	-0.4257	-0.6056	-0.7931

True Mean (when θ_A=0.8)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0778	-0.1100	-0.1494	-0.1718	-0.3091	-0.5185	-0.6770
0.85	-0.0727	-0.1051	-0.1453	-0.1683	-0.3096	-0.5248	-0.6870
0.9	-0.0672	-0.1000	-0.1411	-0.1648	-0.3110	-0.5330	-0.6991
0.95	-0.0605	-0.0937	-0.1362	-0.1609	-0.3140	-0.5450	-0.7162

True Mean (when θ_A=1.0)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0972	-0.1287	-0.1659	-0.1867	-0.3118	-0.5021	-0.6481
0.85	-0.0909	-0.1226	-0.1604	-0.1817	-0.3102	-0.5062	-0.6560
0.9	-0.0840	-0.1159	-0.1546	-0.1764	-0.3092	-0.5119	-0.6662
0.95	-0.0756	-0.1078	-0.1476	-0.1703	-0.3092	-0.5211	-0.6812

Table 2a: Expected over-estimation of treatment effect in studies that continues to the final analysis with a single planned interim analysis (one-sided testing), under the design assumption

True Mean (when θ_A=0.1)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	0.0105	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000
0.85	0.0084	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
0.9	0.0063	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.95	0.0040	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

True Mean (when θ_A=0.2)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	0.0210	0.0051	0.0006	0.0001	0.0000	0.0000	0.0000
0.85	0.0168	0.0033	0.0002	0.0000	0.0000	0.0000	0.0000
0.9	0.0125	0.0018	0.0001	0.0000	0.0000	0.0000	0.0000
0.95	0.0079	0.0006	0.0000	0.0000	0.0000	0.0000	0.0000

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	0.0262	0.0090	0.0019	0.0007	0.0000	0.0000	0.0000
0.85	0.0210	0.0061	0.0010	0.0003	0.0000	0.0000	0.0000
0.9	0.0156	0.0035	0.0004	0.0001	0.0000	0.0000	0.0000
0.95	0.0099	0.0015	0.0001	0.0000	0.0000	0.0000	0.0000

True Mean (when θ_A=0.5)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	0.0525	0.0323	0.0180	0.0128	0.0014	0.0000	0.0000
0.85	0.0419	0.0240	0.0121	0.0082	0.0006	0.0000	0.0000
0.9	0.0313	0.0161	0.0071	0.0044	0.0002	0.0000	0.0000
0.95	0.0198	0.0085	0.0029	0.0016	0.0000	0.0000	0.0000

True Mean (when θ_A=0.8)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	0.0840	0.0628	0.0452	0.0377	0.0129	0.0023	0.0005
0.85	0.0671	0.0480	0.0328	0.0266	0.0075	0.0010	0.0002
0.9	0.0500	0.0337	0.0214	0.0166	0.0036	0.0003	0.0000
0.95	0.0316	0.0191	0.0107	0.0077	0.0011	0.0000	0.0000

True Mean (when θ_A=1.0)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	0.1050	0.0834	0.0647	0.0565	0.0257	0.0076	0.0028
0.85	0.0838	0.0644	0.0480	0.0410	0.0163	0.0039	0.0012
0.9	0.0625	0.0458	0.0323	0.0267	0.0088	0.0016	0.0004
0.95	0.0395	0.0266	0.0170	0.0133	0.0032	0.0003	0.0001

Table 2b-1: Expected over-estimation of treatment effect in studies that continues to the final analysis with a single planned interim analysis (one-sided testing), under the current trend assumption and $\mathbf{when}\ \hat{\boldsymbol{\delta}}_{\mathbf{interim}} = \boldsymbol{\mu} - \mathbf{2.0} \times \mathbf{se}(\hat{\boldsymbol{\delta}}_{\mathbf{interim}})$

True Mean (when θ_A=0.1)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	0.0897	0.0107	0.0000	0.0000	0.0000	0.0000	0.0000
0.85	0.0839	0.0073	0.0000	0.0000	0.0000	0.0000	0.0000
0.9	0.0775	0.0042	0.0000	0.0000	0.0000	0.0000	0.0000
0.95	0.0697	0.0016	0.0000	0.0000	0.0000	0.0000	0.0000

True Mean (when θ_A=0.2)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	0.1794	0.0897	0.0214	0.0053	0.0000	0.0000	0.0000
0.85	0.1677	0.0785	0.0145	0.0026	0.0000	0.0000	0.0000
0.9	0.1550	0.0665	0.0083	0.0009	0.0000	0.0000	0.0000
0.95	0.1394	0.0519	0.0032	0.0001	0.0000	0.0000	0.0000

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	0.2242	0.1334	0.0553	0.0268	0.0000	0.0000	0.0000
0.85	0.2097	0.1191	0.0434	0.0181	0.0000	0.0000	0.0000
0.9	0.1938	0.1037	0.0313	0.0104	0.0000	0.0000	0.0000
0.95	0.1743	0.0848	0.0183	0.0040	0.0000	0.0000	0.0000

True Mean (when θ_A=0.5)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	0.4485	0.3559	0.2669	0.2243	0.0536	0.0005	0.0000
0.85	0.4193	0.3269	0.2383	0.1963	0.0363	0.0001	0.0000
0.9	0.3876	0.2953	0.2074	0.1661	0.0209	0.0000	0.0000
0.95	0.3485	0.2564	0.1696	0.1296	0.0079	0.0000	0.0000

True Mean (when θ_A=0.8)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	0.7175	0.6245	0.5332	0.4884	0.2782	0.0858	0.0212
0.85	0.6709	0.5779	0.4868	0.4423	0.2352	0.0580	0.0103
0.9	0.6202	0.5272	0.4365	0.3922	0.1896	0.0334	0.0036
0.95	0.5577	0.4648	0.3745	0.3307	0.1357	0.0127	0.0005

True Mean (when θ_A=1.0)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	0.8969	0.8037	0.7118	0.6665	0.4487	0.2211	0.1072
0.85	0.8386	0.7454	0.6537	0.6085	0.3926	0.1735	0.0725
0.9	0.7752	0.6821	0.5906	0.5456	0.3323	0.1252	0.0417
0.95	0.6971	0.6040	0.5129	0.4682	0.2593	0.0731	0.0158

Table 2b-2: Expected over-estimation of treatment effect in studies that continues to the final analysis with a single planned interim analysis (one-sided testing), under the current trend assumption and when $\theta_A = 0.25$

$$when\ \ \hat{\delta}_{interim} = \mu - 2.0 \times se(\hat{\delta}_{interim})$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	0.2242	0.1334	0.0553	0.0268	0.0000	0.0000	0.0000
0.85	0.2097	0.1191	0.0434	0.0181	0.0000	0.0000	0.0000
0.9	0.1938	0.1037	0.0313	0.0104	0.0000	0.0000	0.0000
0.95	0.1743	0.0848	0.0183	0.0040	0.0000	0.0000	0.0000

$$when\ \ \hat{\delta}_{interim} = \mu - 1.5 \times se(\hat{\delta}_{interim})$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	0.1949	0.1067	0.0361	0.0144	0.0000	0.0000	0.0000
0.85	0.1822	0.0945	0.0271	0.0089	0.0000	0.0000	0.0000
0.9	0.1684	0.0813	0.0183	0.0045	0.0000	0.0000	0.0000
0.95	0.1515	0.0652	0.0095	0.0014	0.0000	0.0000	0.0000

$$when\ \ \hat{\delta}_{interim} = \mu - 1.0 \times se(\hat{\delta}_{interim})$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	0.1661	0.0817	0.0211	0.0065	0.0000	0.0000	0.0000
0.85	0.1553	0.0714	0.0149	0.0036	0.0000	0.0000	0.0000
0.9	0.1435	0.0605	0.0092	0.0016	0.0000	0.0000	0.0000
0.95	0.1291	0.0473	0.0041	0.0004	0.0000	0.0000	0.0000

$$when\ \ \hat{\delta}_{interim} = \mu - 0.5 \times se(\hat{\delta}_{interim})$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	0.1381	0.0589	0.0106	0.0024	0.0000	0.0000	0.0000
0.85	0.1291	0.0507	0.0069	0.0012	0.0000	0.0000	0.0000
0.9	0.1193	0.0420	0.0038	0.0004	0.0000	0.0000	0.0000
0.95	0.1073	0.0317	0.0014	0.0001	0.0000	0.0000	0.0000

$$when\ \ \hat{\delta}_{interim} = \mu$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	0.1112	0.0391	0.0044	0.0007	0.0000	0.0000	0.0000
0.85	0.1039	0.0330	0.0026	0.0003	0.0000	0.0000	0.0000
0.9	0.0961	0.0266	0.0013	0.0001	0.0000	0.0000	0.0000
0.95	0.0864	0.0191	0.0004	0.0000	0.0000	0.0000	0.0000

$$when\ \ \hat{\delta}_{interim} = \mu + 0.5 \times se(\hat{\delta}_{interim})$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	0.0858	0.0233	0.0015	0.0002	0.0000	0.0000	0.0000
0.85	0.0802	0.0191	0.0008	0.0001	0.0000	0.0000	0.0000
0.9	0.0741	0.0148	0.0003	0.0000	0.0000	0.0000	0.0000
0.95	0.0667	0.0100	0.0001	0.0000	0.0000	0.0000	0.0000

$$when\ \ \hat{\delta}_{interim} = \mu + 1.0 \times se(\hat{\delta}_{interim})$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	0.0625	0.0121	0.0004	0.0000	0.0000	0.0000	0.0000
0.85	0.0585	0.0095	0.0002	0.0000	0.0000	0.0000	0.0000
0.9	0.0540	0.0070	0.0001	0.0000	0.0000	0.0000	0.0000
0.95	0.0486	0.0044	0.0000	0.0000	0.0000	0.0000	0.0000

$$when\ \ \hat{\delta}_{interim} = \mu + 1.5 \times se(\hat{\delta}_{interim})$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	0.0422	0.0052	0.0001	0.0000	0.0000	0.0000	0.0000
0.85	0.0395	0.0039	0.0000	0.0000	0.0000	0.0000	0.0000
0.9	0.0365	0.0028	0.0000	0.0000	0.0000	0.0000	0.0000
0.95	0.0328	0.0016	0.0000	0.0000	0.0000	0.0000	0.0000

$$when\ \ \hat{\delta}_{interim} = \mu + 2.0 \times se(\hat{\delta}_{interim})$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	0.0257	0.0018	0.0000	0.0000	0.0000	0.0000	0.0000
0.85	0.0240	0.0013	0.0000	0.0000	0.0000	0.0000	0.0000
0.9	0.0222	0.0009	0.0000	0.0000	0.0000	0.0000	0.0000
0.95	0.0200	0.0004	0.0000	0.0000	0.0000	0.0000	0.0000

Table 2c: Expected over-estimation of treatment effect in studies that continues to the final analysis with a single planned interim analysis (one-sided testing), under the null assumption

True Mean (when θ_A=0.1)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	0.0445	0.0105	0.0003	0.0000	0.0000	0.0000	0.0000
0.85	0.0416	0.0084	0.0001	0.0000	0.0000	0.0000	0.0000
0.9	0.0384	0.0063	0.0000	0.0000	0.0000	0.0000	0.0000
0.95	0.0346	0.0040	0.0000	0.0000	0.0000	0.0000	0.0000

True Mean (when θ_A=0.2)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	0.0889	0.0504	0.0210	0.0113	0.0000	0.0000	0.0000
0.85	0.0832	0.0448	0.0168	0.0082	0.0000	0.0000	0.0000
0.9	0.0769	0.0389	0.0125	0.0053	0.0000	0.0000	0.0000
0.95	0.0691	0.0317	0.0079	0.0027	0.0000	0.0000	0.0000

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	0.1112	0.0719	0.0391	0.0262	0.0007	0.0000	0.0000
0.85	0.1039	0.0648	0.0330	0.0210	0.0003	0.0000	0.0000
0.9	0.0961	0.0572	0.0266	0.0156	0.0001	0.0000	0.0000
0.95	0.0864	0.0479	0.0191	0.0099	0.0000	0.0000	0.0000

True Mean (when θ_A=0.5)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	0.2223	0.1818	0.1438	0.0782	0.0323	0.0089	0.0014
0.85	0.2079	0.1674	0.1297	0.0660	0.0240	0.0053	0.0006
0.9	0.1922	0.1518	0.1145	0.0531	0.0161	0.0026	0.0002
0.95	0.1728	0.1326	0.0959	0.0382	0.0085	0.0008	0.0000

True Mean (when θ_A=0.8)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	0.3557	0.3148	0.2752	0.2561	0.1678	0.0840	0.0452
0.85	0.3326	0.2917	0.2523	0.2333	0.1465	0.0671	0.0328
0.9	0.3075	0.2666	0.2274	0.2086	0.1238	0.0500	0.0214
0.95	0.2765	0.2357	0.1969	0.1783	0.0966	0.0316	0.0107

True Mean (when θ_A=1.0)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	0.4447	0.4036	0.3636	0.3440	0.2518	0.1565	0.1050
0.85	0.4158	0.3747	0.3348	0.3154	0.2242	0.1319	0.0838
0.9	0.3843	0.3433	0.3036	0.2843	0.1945	0.1062	0.0625
0.95	0.3456	0.3047	0.2652	0.2461	0.1584	0.0764	0.0395

Table 3a: Overall bias for a single study with a stopping rule and one planned interim analysis (one-sided testing), under the design assumption

True Mean (when θ_A=0.1)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0079	-0.0003	0.0000	0.0000	0.0000	0.0000	0.0000
0.85	-0.0067	-0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
0.9	-0.0053	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.95	-0.0035	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

True Mean (when θ_A=0.2)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0159	-0.0049	-0.0006	-0.0001	0.0000	0.0000	0.0000
0.85	-0.0134	-0.0032	-0.0002	0.0000	0.0000	0.0000	0.0000
0.9	-0.0105	-0.0017	-0.0001	0.0000	0.0000	0.0000	0.0000
0.95	-0.0071	-0.0006	0.0000	0.0000	0.0000	0.0000	0.0000

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0198	-0.0084	-0.0019	-0.0007	0.0000	0.0000	0.0000
0.85	-0.0167	-0.0058	-0.0010	-0.0003	0.0000	0.0000	0.0000
0.9	-0.0132	-0.0034	-0.0004	-0.0001	0.0000	0.0000	0.0000
0.95	-0.0089	-0.0015	-0.0001	0.0000	0.0000	0.0000	0.0000

True Mean (when θ_A=0.5)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0396	-0.0279	-0.0167	-0.0122	-0.0014	0.0000	0.0000
0.85	-0.0334	-0.0215	-0.0115	-0.0079	-0.0006	0.0000	0.0000
0.9	-0.0263	-0.0150	-0.0069	-0.0043	-0.0002	0.0000	0.0000
0.95	-0.0177	-0.0082	-0.0029	-0.0016	0.0000	0.0000	0.0000

True Mean (when θ_A=0.8)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0634	-0.0518	-0.0398	-0.0341	-0.0125	-0.0023	-0.0005
0.85	-0.0534	-0.0414	-0.0299	-0.0248	-0.0074	-0.0010	-0.0002
0.9	-0.0421	-0.0303	-0.0201	-0.0159	-0.0036	-0.0003	0.0000
0.95	-0.0283	-0.0180	-0.0104	-0.0076	-0.0011	0.0000	0.0000

True Mean (when θ_A=1.0)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0793	-0.0678	-0.0557	-0.0498	-0.0245	-0.0076	-0.0028
0.85	-0.0668	-0.0548	-0.0430	-0.0374	-0.0158	-0.0039	-0.0012
0.9	-0.0527	-0.0408	-0.0299	-0.0252	-0.0087	-0.0016	-0.0004
0.95	-0.0354	-0.0249	-0.0164	-0.0130	-0.0032	-0.0003	-0.0001

Table 3b-1: Overall bias for a single study with a stopping rule and one planned interim analysis (one-sided testing), under the current trend assumption and when $\hat{\delta}_{\mathbf{interim}} = \mu - 2.0 \times se(\hat{\delta}_{\mathbf{interim}})$

True Mean (when θ_A=0.1)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	0.0000	-0.0080	0.0000	0.0000	0.0000	0.0000	0.0000
0.85	0.0000	-0.0060	0.0000	0.0000	0.0000	0.0000	0.0000
0.9	0.0000	-0.0038	0.0000	0.0000	0.0000	0.0000	0.0000
0.95	0.0000	-0.0015	0.0000	0.0000	0.0000	0.0000	0.0000

True Mean (when θ_A=0.2)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0001	-0.0086	-0.0161	-0.0051	0.0000	0.0000	0.0000
0.85	-0.0001	-0.0095	-0.0120	-0.0025	0.0000	0.0000	0.0000
0.9	-0.0001	-0.0106	-0.0075	-0.0009	0.0000	0.0000	0.0000
0.95	-0.0001	-0.0119	-0.0031	-0.0001	0.0000	0.0000	0.0000

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0001	-0.0059	-0.0250	-0.0201	0.0000	0.0000	0.0000
0.85	-0.0001	-0.0066	-0.0234	-0.0150	0.0000	0.0000	0.0000
0.9	-0.0001	-0.0075	-0.0203	-0.0094	0.0000	0.0000	0.0000
0.95	-0.0001	-0.0090	-0.0143	-0.0038	0.0000	0.0000	0.0000

True Mean (when θ_A=0.5)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0002	-0.0022	-0.0118	-0.0214	-0.0402	-0.0005	0.0000
0.85	-0.0002	-0.0024	-0.0132	-0.0237	-0.0300	-0.0001	0.0000
0.9	-0.0002	-0.0026	-0.0151	-0.0265	-0.0188	0.0000	0.0000
0.95	-0.0002	-0.0030	-0.0179	-0.0297	-0.0076	0.0000	0.0000

True Mean (when θ_A=0.8)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0004	-0.0016	-0.0057	-0.0098	-0.0579	-0.0643	-0.0202
0.85	-0.0003	-0.0017	-0.0063	-0.0109	-0.0613	-0.0481	-0.0100
0.9	-0.0003	-0.0017	-0.0070	-0.0124	-0.0636	-0.0301	-0.0035
0.95	-0.0003	-0.0019	-0.0082	-0.0148	-0.0623	-0.0122	-0.0005

True Mean (when θ_A=1.0)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0004	-0.0015	-0.0045	-0.0072	-0.0428	-0.1000	-0.0803
0.85	-0.0004	-0.0015	-0.0048	-0.0078	-0.0474	-0.0937	-0.0601
0.9	-0.0004	-0.0016	-0.0052	-0.0087	-0.0529	-0.0810	-0.0377
0.95	-0.0003	-0.0016	-0.0059	-0.0103	-0.0595	-0.0573	-0.0153

Table 3b-2: Overall bias for a single study with a stopping rule and one planned interim analysis (one-sided testing), under the current trend assumption and when θ_A=0.25

$when\ \hat{\delta}_{interim} = \mu - 2.0 \times se(\hat{\delta}_{interim})$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0001	-0.0059	-0.0250	-0.0201	0.0000	0.0000	0.0000
0.85	-0.0001	-0.0066	-0.0234	-0.0150	0.0000	0.0000	0.0000
0.9	-0.0001	-0.0075	-0.0203	-0.0094	0.0000	0.0000	0.0000
0.95	-0.0001	-0.0090	-0.0143	-0.0038	0.0000	0.0000	0.0000

$when\ \hat{\delta}_{interim} = \mu - 1.5 \times se(\hat{\delta}_{interim})$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0005	-0.0122	-0.0234	-0.0127	0.0000	0.0000	0.0000
0.85	-0.0005	-0.0129	-0.0197	-0.0083	0.0000	0.0000	0.0000
0.9	-0.0004	-0.0138	-0.0148	-0.0043	0.0000	0.0000	0.0000
0.95	-0.0004	-0.0148	-0.0085	-0.0014	0.0000	0.0000	0.0000

$$when \ \hat{\delta}_{interim} = \mu - 1.0 \times se(\hat{\delta}_{interim})$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0018	-0.0196	-0.0171	-0.0062	0.0000	0.0000	0.0000
0.85	-0.0017	-0.0197	-0.0128	-0.0035	0.0000	0.0000	0.0000
0.9	-0.0016	-0.0196	-0.0084	-0.0016	0.0000	0.0000	0.0000
0.95	-0.0014	-0.0190	-0.0040	-0.0004	0.0000	0.0000	0.0000

$$when \ \hat{\delta}_{interim} = \mu - 0.5 \times se(\hat{\delta}_{interim})$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0051	-0.0246	-0.0097	-0.0024	0.0000	0.0000	0.0000
0.85	-0.0047	-0.0234	-0.0065	-0.0012	0.0000	0.0000	0.0000
0.9	-0.0044	-0.0217	-0.0037	-0.0004	0.0000	0.0000	0.0000
0.95	-0.0039	-0.0190	-0.0014	-0.0001	0.0000	0.0000	0.0000

$$when \ \hat{\delta}_{interim} = \mu$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0110	-0.0241	-0.0043	-0.0007	0.0000	0.0000	0.0000
0.85	-0.0102	-0.0217	-0.0026	-0.0003	0.0000	0.0000	0.0000
0.9	-0.0095	-0.0188	-0.0013	-0.0001	0.0000	0.0000	0.0000
0.95	-0.0085	-0.0148	-0.0004	0.0000	0.0000	0.0000	0.0000

$$when \ \hat{\delta}_{interim} = \mu + 0.5 \times se(\hat{\delta}_{interim})$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0184	-0.0184	-0.0015	-0.0002	0.0000	0.0000	0.0000
0.85	-0.0172	-0.0156	-0.0008	-0.0001	0.0000	0.0000	0.0000
0.9	-0.0159	-0.0126	-0.0003	0.0000	0.0000	0.0000	0.0000
0.95	-0.0143	-0.0090	-0.0001	0.0000	0.0000	0.0000	0.0000

$$when\ \ \hat{\delta}_{interim} = \mu + 1.0 \times se(\hat{\delta}_{interim})$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0241	-0.0109	-0.0004	0.0000	0.0000	0.0000	0.0000
0.85	-0.0226	-0.0088	-0.0002	0.0000	0.0000	0.0000	0.0000
0.9	-0.0209	-0.0066	-0.0001	0.0000	0.0000	0.0000	0.0000
0.95	-0.0188	-0.0042	0.0000	0.0000	0.0000	0.0000	0.0000

$$when\ \ \hat{\delta}_{interim} = \mu + 1.5 \times se(\hat{\delta}_{interim})$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0246	-0.0050	-0.0001	0.0000	0.0000	0.0000	0.0000
0.85	-0.0230	-0.0038	0.0000	0.0000	0.0000	0.0000	0.0000
0.9	-0.0213	-0.0027	0.0000	0.0000	0.0000	0.0000	0.0000
0.95	-0.0191	-0.0016	0.0000	0.0000	0.0000	0.0000	0.0000

$$when\ \ \hat{\delta}_{interim} = \mu + 2.0 \times se(\hat{\delta}_{interim})$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0196	-0.0018	0.0000	0.0000	0.0000	0.0000	0.0000
0.85	-0.0183	-0.0013	0.0000	0.0000	0.0000	0.0000	0.0000
0.9	-0.0169	-0.0009	0.0000	0.0000	0.0000	0.0000	0.0000
0.95	-0.0152	-0.0004	0.0000	0.0000	0.0000	0.0000	0.0000

Table 3c: Overall bias for a single study with a stopping rule and one planned interim analysis (one-sided testing), under the null assumption

True Mean (when θ_A=0.1)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0044	-0.0079	-0.0003	0.0000	0.0000	0.0000	0.0000
0.85	-0.0041	-0.0067	-0.0001	0.0000	0.0000	0.0000	0.0000
0.9	-0.0038	-0.0053	0.0000	0.0000	0.0000	0.0000	0.0000
0.95	-0.0034	-0.0035	0.0000	0.0000	0.0000	0.0000	0.0000

True Mean (when θ_A=0.2)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0088	-0.0193	-0.0159	-0.0100	0.0000	0.0000	0.0000
0.85	-0.0082	-0.0183	-0.0134	-0.0075	0.0000	0.0000	0.0000
0.9	-0.0076	-0.0172	-0.0105	-0.0050	0.0000	0.0000	0.0000
0.95	-0.0068	-0.0156	-0.0071	-0.0026	0.0000	0.0000	0.0000

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0088	-0.0178	-0.0193	-0.0159	-0.0006	0.0000	0.0000
0.85	-0.0082	-0.0171	-0.0173	-0.0134	-0.0002	0.0000	0.0000
0.9	-0.0076	-0.0162	-0.0150	-0.0105	-0.0001	0.0000	0.0000
0.95	-0.0068	-0.0151	-0.0118	-0.0071	0.0000	0.0000	0.0000

True Mean (when θ_A=0.5)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0219	-0.0338	-0.0445	-0.0482	-0.0279	-0.0086	-0.0014
0.85	-0.0205	-0.0324	-0.0427	-0.0434	-0.0215	-0.0052	-0.0006
0.9	-0.0189	-0.0308	-0.0406	-0.0375	-0.0150	-0.0026	-0.0002
0.95	-0.0170	-0.0289	-0.0377	-0.0295	-0.0082	-0.0008	0.0000

True Mean (when θ_A=0.8)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0351	-0.0468	-0.0588	-0.0643	-0.0804	-0.0634	-0.0398
0.85	-0.0328	-0.0445	-0.0564	-0.0619	-0.0753	-0.0534	-0.0299
0.9	-0.0303	-0.0421	-0.0539	-0.0591	-0.0689	-0.0421	-0.0201
0.95	-0.0273	-0.0391	-0.0506	-0.0554	-0.0598	-0.0283	-0.0104

True Mean (when θ_A=1.0)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0438	-0.0555	-0.0675	-0.0734	-0.0963	-0.0964	-0.0793
0.85	-0.0410	-0.0527	-0.0647	-0.0705	-0.0917	-0.0867	-0.0668
0.9	-0.0379	-0.0496	-0.0616	-0.0673	-0.0861	-0.0751	-0.0527
0.95	-0.0341	-0.0458	-0.0577	-0.0633	-0.0782	-0.0591	-0.0354

Table 4a: Probability of a study stopping at a single planned interim analysis (one-sided testing), under the design assumption

True Mean (when θ_A=0.1)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	24.47%	0.38%	0.00%	0.00%	0.00%	0.00%	0.00%
0.85	20.35%	0.16%	0.00%	0.00%	0.00%	0.00%	0.00%
0.9	15.81%	0.05%	0.00%	0.00%	0.00%	0.00%	0.00%
0.95	10.40%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%

True Mean (when θ_A=0.2)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	24.47%	4.63%	0.38%	0.08%	0.00%	0.00%	0.00%
0.85	20.35%	2.95%	0.16%	0.03%	0.00%	0.00%	0.00%
0.9	15.81%	1.58%	0.05%	0.01%	0.00%	0.00%	0.00%
0.95	10.40%	0.56%	0.01%	0.00%	0.00%	0.00%	0.00%

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	24.47%	6.89%	1.14%	0.38%	0.00%	0.00%	0.00%
0.85	20.35%	4.68%	0.58%	0.16%	0.00%	0.00%	0.00%
0.9	15.81%	2.75%	0.23%	0.05%	0.00%	0.00%	0.00%
0.95	10.40%	1.13%	0.05%	0.01%	0.00%	0.00%	0.00%

True Mean (when θ_A=0.5)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	24.47%	13.84%	6.89%	4.63%	0.38%	0.01%	0.00%
0.85	20.35%	10.51%	4.68%	2.95%	0.16%	0.00%	0.00%
0.9	15.81%	7.20%	2.75%	1.58%	0.05%	0.00%	0.00%
0.95	10.40%	3.84%	1.13%	0.56%	0.01%	0.00%	0.00%

True Mean (when θ_A=0.8)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	24.47%	17.39%	11.77%	9.50%	2.68%	0.38%	0.08%
0.85	20.35%	13.70%	8.71%	6.80%	1.56%	0.16%	0.03%
0.9	15.81%	9.87%	5.76%	4.28%	0.74%	0.05%	0.01%
0.95	10.40%	5.73%	2.89%	1.99%	0.22%	0.01%	0.00%

True Mean (when θ_A=1.0)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	24.47%	18.69%	13.84%	11.77%	4.63%	1.14%	0.38%
0.85	20.35%	14.90%	10.51%	8.71%	2.95%	0.58%	0.16%
0.9	15.81%	10.91%	7.20%	5.76%	1.58%	0.23%	0.05%
0.95	10.40%	6.50%	3.84%	2.89%	0.56%	0.05%	0.01%

Table 4b-1: Probability of a study stopping at a single planned interim analysis (one-sided testing), under the current trend assumption, when $\widehat{\boldsymbol{\delta}}_{\mathbf{interim}} = \boldsymbol{\mu} - \mathbf{2.0} \times \mathbf{se}(\widehat{\boldsymbol{\delta}}_{\mathbf{interim}})$

True Mean (when θ_A=0.1)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	99.95%	25.07%	0.00%	0.00%	0.00%	0.00%	0.00%
0.85	99.95%	17.16%	0.00%	0.00%	0.00%	0.00%	0.00%
0.9	99.95%	9.78%	0.00%	0.00%	0.00%	0.00%	0.00%
0.95	99.95%	3.53%	0.00%	0.00%	0.00%	0.00%	0.00%

True Mean (when θ_A=0.2)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	99.95%	90.47%	25.07%	4.82%	0.00%	0.00%	0.00%
0.85	99.95%	87.92%	17.16%	2.24%	0.00%	0.00%	0.00%
0.9	99.95%	84.08%	9.78%	0.73%	0.00%	0.00%	0.00%
0.95	99.95%	77.06%	3.53%	0.10%	0.00%	0.00%	0.00%

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	99.95%	95.59%	54.78%	25.07%	0.00%	0.00%	0.00%
0.85	99.95%	94.46%	46.01%	17.16%	0.00%	0.00%	0.00%
0.9	99.95%	92.73%	35.29%	9.78%	0.00%	0.00%	0.00%
0.95	99.95%	89.45%	21.52%	3.53%	0.00%	0.00%	0.00%

True Mean (when θ_A=0.5)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	99.95%	99.37%	95.59%	90.47%	25.07%	0.11%	0.00%
0.85	99.95%	99.27%	94.46%	87.92%	17.16%	0.02%	0.00%
0.9	99.95%	99.12%	92.73%	84.08%	9.78%	0.00%	0.00%
0.95	99.95%	98.84%	89.45%	77.06%	3.53%	0.00%	0.00%

True Mean (when θ_A=0.8)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	99.95%	99.74%	98.93%	97.99%	79.20%	25.07%	4.82%
0.85	99.95%	99.71%	98.71%	97.53%	73.93%	17.16%	2.24%
0.9	99.95%	99.67%	98.40%	96.84%	66.44%	9.78%	0.73%
0.95	99.95%	99.60%	97.81%	95.51%	54.12%	3.53%	0.10%

True Mean (when θ_A=1.0)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	99.95%	99.81%	99.37%	98.93%	90.47%	54.78%	25.07%
0.85	99.95%	99.79%	99.27%	98.71%	87.92%	46.01%	17.16%
0.9	99.95%	99.77%	99.12%	98.40%	84.08%	35.29%	9.78%
0.95	99.95%	99.73%	98.84%	97.81%	77.06%	21.52%	3.53%

Table 4b-2: Probability of a study stopping at a single planned interim analysis (one-sided testing), under the current trend assumption, when θ_A=0.25

$$when\ \ \hat{\delta}_{interim} = \mu - 2.0 \times se(\hat{\delta}_{interim})$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	99.95%	95.59%	54.78%	25.07%	0.00%	0.00%	0.00%
0.85	99.95%	94.46%	46.01%	17.16%	0.00%	0.00%	0.00%
0.9	99.95%	92.73%	35.29%	9.78%	0.00%	0.00%	0.00%
0.95	99.95%	89.45%	21.52%	3.53%	0.00%	0.00%	0.00%

$$when\ \ \hat{\delta}_{interim} = \mu - 1.5 \times se(\hat{\delta}_{interim})$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	99.74%	88.59%	35.20%	12.05%	0.00%	0.00%	0.00%
0.85	99.74%	86.32%	27.42%	7.38%	0.00%	0.00%	0.00%
0.9	99.74%	83.05%	19.01%	3.64%	0.00%	0.00%	0.00%
0.95	99.74%	77.35%	9.88%	1.05%	0.00%	0.00%	0.00%

$$when\ \hat{\delta}_{interim} = \mu - 1.0 \times se(\hat{\delta}_{interim})$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	98.90%	75.96%	18.95%	4.72%	0.00%	0.00%	0.00%
0.85	98.90%	72.40%	13.56%	2.57%	0.00%	0.00%	0.00%
0.9	98.90%	67.58%	8.42%	1.09%	0.00%	0.00%	0.00%
0.95	98.90%	59.89%	3.68%	0.25%	0.00%	0.00%	0.00%

$$when\ \hat{\delta}_{interim} = \mu - 0.5 \times se(\hat{\delta}_{interim})$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	96.33%	58.12%	8.38%	1.49%	0.00%	0.00%	0.00%
0.85	96.33%	53.77%	5.48%	0.72%	0.00%	0.00%	0.00%
0.9	96.33%	48.25%	3.02%	0.26%	0.00%	0.00%	0.00%
0.95	96.33%	40.15%	1.11%	0.05%	0.00%	0.00%	0.00%

$$when\ \hat{\delta}_{interim} = \mu$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	90.14%	38.40%	3.01%	0.38%	0.00%	0.00%	0.00%
0.85	90.14%	34.26%	1.79%	0.16%	0.00%	0.00%	0.00%
0.9	90.14%	29.32%	0.87%	0.05%	0.00%	0.00%	0.00%
0.95	90.14%	22.68%	0.26%	0.01%	0.00%	0.00%	0.00%

$$when\ \hat{\delta}_{interim} = \mu + 0.5 \times se(\hat{\delta}_{interim})$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	78.52%	21.33%	0.87%	0.08%	0.00%	0.00%	0.00%
0.85	78.52%	18.27%	0.47%	0.03%	0.00%	0.00%	0.00%
0.9	78.52%	14.83%	0.20%	0.01%	0.00%	0.00%	0.00%
0.95	78.52%	10.57%	0.05%	0.00%	0.00%	0.00%	0.00%

$$when\ \hat{\delta}_{interim} = \mu + 1.0 \times se(\hat{\delta}_{interim})$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	61.40%	9.76%	0.20%	0.01%	0.00%	0.00%	0.00%
0.85	61.40%	8.00%	0.10%	0.00%	0.00%	0.00%	0.00%
0.9	61.40%	6.13%	0.04%	0.00%	0.00%	0.00%	0.00%
0.95	61.40%	4.01%	0.01%	0.00%	0.00%	0.00%	0.00%

$$when\ \hat{\delta}_{interim} = \mu + 1.5 \times se(\hat{\delta}_{interim})$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	41.67%	3.63%	0.04%	0.00%	0.00%	0.00%	0.00%
0.85	41.67%	2.84%	0.02%	0.00%	0.00%	0.00%	0.00%
0.9	41.67%	2.05%	0.01%	0.00%	0.00%	0.00%	0.00%
0.95	41.67%	1.22%	0.00%	0.00%	0.00%	0.00%	0.00%

$$when\ \hat{\delta}_{interim} = \mu + 2.0 \times se(\hat{\delta}_{interim})$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	23.88%	1.09%	0.01%	0.00%	0.00%	0.00%	0.00%
0.85	23.88%	0.81%	0.00%	0.00%	0.00%	0.00%	0.00%
0.9	23.88%	0.55%	0.00%	0.00%	0.00%	0.00%	0.00%
0.95	23.88%	0.30%	0.00%	0.00%	0.00%	0.00%	0.00%

Table 4c: Probability of a study stopping at a single planned interim analysis (one-sided testing), under the null assumption

True Mean (when θ_A=0.1)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	90.14%	24.47%	0.38%	0.01%	0.00%	0.00%	0.00%
0.85	90.14%	20.35%	0.16%	0.00%	0.00%	0.00%	0.00%
0.9	90.14%	15.81%	0.05%	0.00%	0.00%	0.00%	0.00%
0.95	90.14%	10.40%	0.01%	0.00%	0.00%	0.00%	0.00%

True Mean (when θ_A=0.2)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	90.14%	61.76%	24.47%	11.77%	0.01%	0.00%	0.00%
0.85	90.14%	59.11%	20.35%	8.71%	0.00%	0.00%	0.00%
0.9	90.14%	55.71%	15.81%	5.76%	0.00%	0.00%	0.00%
0.95	90.14%	50.61%	10.40%	2.89%	0.00%	0.00%	0.00%

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	90.14%	69.05%	38.40%	24.47%	0.38%	0.00%	0.00%
0.85	90.14%	67.08%	34.26%	20.35%	0.16%	0.00%	0.00%
0.9	90.14%	64.54%	29.32%	15.81%	0.05%	0.00%	0.00%
0.95	90.14%	60.65%	22.68%	10.40%	0.01%	0.00%	0.00%

True Mean (when θ_A=0.5)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	90.14%	81.42%	69.05%	38.40%	13.84%	3.01%	0.38%
0.85	90.14%	80.68%	67.08%	34.26%	10.51%	1.79%	0.16%
0.9	90.14%	79.71%	64.54%	29.32%	7.20%	0.87%	0.05%
0.95	90.14%	78.23%	60.65%	22.68%	3.84%	0.26%	0.01%

True Mean (when θ_A=0.8)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	90.14%	85.13%	78.65%	74.88%	52.06%	24.47%	11.77%
0.85	90.14%	84.73%	77.64%	73.49%	48.62%	20.35%	8.71%
0.9	90.14%	84.21%	76.32%	71.68%	44.32%	15.81%	5.76%
0.95	90.14%	83.43%	74.30%	68.91%	38.08%	10.40%	2.89%

True Mean (when θ_A=1.0)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	90.14%	86.25%	81.42%	78.65%	61.76%	38.40%	24.47%
0.85	90.14%	85.95%	80.68%	77.64%	59.11%	34.26%	20.35%
0.9	90.14%	85.55%	79.71%	76.32%	55.71%	29.32%	15.81%
0.95	90.14%	84.96%	78.23%	74.30%	50.61%	22.68%	10.40%

Table 5a: Relative weight of stopped studies in a meta-analysis, with one planned interim analysis and under the design assumption

True Mean (when θ_A=0.1)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	13.94%	0.19%	0.00%	0.00%	0.00%	0.00%	0.00%
0.85	11.33%	0.08%	0.00%	0.00%	0.00%	0.00%	0.00%
0.9	8.58%	0.02%	0.00%	0.00%	0.00%	0.00%	0.00%
0.95	5.48%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

True Mean (when θ_A=0.2)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	13.94%	2.37%	0.19%	0.04%	0.00%	0.00%	0.00%
0.85	11.33%	1.50%	0.08%	0.01%	0.00%	0.00%	0.00%
0.9	8.58%	0.80%	0.02%	0.00%	0.00%	0.00%	0.00%
0.95	5.48%	0.28%	0.00%	0.00%	0.00%	0.00%	0.00%

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	13.94%	3.57%	0.57%	0.19%	0.00%	0.00%	0.00%
0.85	11.33%	2.40%	0.29%	0.08%	0.00%	0.00%	0.00%
0.9	8.58%	1.39%	0.11%	0.02%	0.00%	0.00%	0.00%
0.95	5.48%	0.57%	0.02%	0.00%	0.00%	0.00%	0.00%

True Mean (when θ_A=0.5)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	13.94%	7.44%	3.57%	2.37%	0.19%	0.00%	0.00%
0.85	11.33%	5.55%	2.40%	1.50%	0.08%	0.00%	0.00%
0.9	8.58%	3.74%	1.39%	0.80%	0.02%	0.00%	0.00%
0.95	5.48%	1.96%	0.57%	0.28%	0.00%	0.00%	0.00%

True Mean (when θ_A=0.8)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	13.94%	9.52%	6.25%	4.99%	1.36%	0.19%	0.04%
0.85	11.33%	7.35%	4.55%	3.52%	0.79%	0.08%	0.01%
0.9	8.58%	5.19%	2.96%	2.19%	0.37%	0.02%	0.00%
0.95	5.48%	2.95%	1.47%	1.01%	0.11%	0.00%	0.00%

True Mean (when θ_A=1.0)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	13.94%	10.31%	7.44%	6.25%	2.37%	0.57%	0.19%
0.85	11.33%	8.05%	5.55%	4.55%	1.50%	0.29%	0.08%
0.9	8.58%	5.77%	3.74%	2.96%	0.80%	0.11%	0.02%
0.95	5.48%	3.36%	1.96%	1.47%	0.28%	0.02%	0.00%

Table 5b-1: Relative weight of stopped studies in a meta-analysis, with one planned interim analysis (one-sided testing), under the current trend assumption and when $\hat{\delta}_{\mathbf{interim}} = \mu - 2.0 \times \mathrm{se}(\hat{\delta}_{\mathbf{interim}})$

True Mean (when θ_A=0.1)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	99.90%	14.33%	0.00%	0.00%	0.00%	0.00%	0.00%
0.85	99.90%	9.39%	0.00%	0.00%	0.00%	0.00%	0.00%
0.9	99.90%	5.14%	0.00%	0.00%	0.00%	0.00%	0.00%
0.95	99.90%	1.80%	0.00%	0.00%	0.00%	0.00%	0.00%

True Mean (when θ_A=0.2)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	99.90%	82.60%	14.33%	2.47%	0.00%	0.00%	0.00%
0.85	99.90%	78.44%	9.39%	1.13%	0.00%	0.00%	0.00%
0.9	99.90%	72.53%	5.14%	0.37%	0.00%	0.00%	0.00%
0.95	99.90%	62.68%	1.80%	0.05%	0.00%	0.00%	0.00%

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	99.90%	91.55%	37.72%	14.33%	0.00%	0.00%	0.00%
0.85	99.90%	89.50%	29.87%	9.39%	0.00%	0.00%	0.00%
0.9	99.90%	86.45%	21.42%	5.14%	0.00%	0.00%	0.00%
0.95	99.90%	80.91%	12.05%	1.80%	0.00%	0.00%	0.00%

True Mean (when θ_A=0.5)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	99.90%	98.76%	91.55%	82.60%	14.33%	0.06%	0.00%
0.85	99.90%	98.55%	89.50%	78.44%	9.39%	0.01%	0.00%
0.9	99.90%	98.25%	86.45%	72.53%	5.14%	0.00%	0.00%
0.95	99.90%	97.71%	80.91%	62.68%	1.80%	0.00%	0.00%

True Mean (when θA=0.8)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	99.90%	99.48%	97.87%	96.06%	65.57%	14.33%	2.47%
0.85	99.90%	99.42%	97.46%	95.18%	58.65%	9.39%	1.13%
0.9	99.90%	99.34%	96.84%	93.87%	49.75%	5.14%	0.37%
0.95	99.90%	99.20%	95.71%	91.41%	37.10%	1.80%	0.05%

True Mean (when θA=1.0)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	99.90%	99.62%	98.76%	97.87%	82.60%	37.72%	14.33%
0.85	99.90%	99.59%	98.55%	97.46%	78.44%	29.87%	9.39%
0.9	99.90%	99.54%	98.25%	96.84%	72.53%	21.42%	5.14%
0.95	99.90%	99.46%	97.71%	95.71%	62.68%	12.05%	1.80%

Table 5b-2: Relative weight of stopped studies in a meta-analysis, with one planned interim analysis (one-sided testing), under the current trend assumption and when θA=0.25

$$when\ \ \hat{\delta}_{interim} = \mu - 2.0 \times se(\hat{\delta}_{interim})$$

True Mean (when θA=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	99.90%	91.55%	37.72%	14.33%	0.00%	0.00%	0.00%
0.85	99.90%	89.50%	29.87%	9.39%	0.00%	0.00%	0.00%
0.9	99.90%	86.45%	21.42%	5.14%	0.00%	0.00%	0.00%
0.95	99.90%	80.91%	12.05%	1.80%	0.00%	0.00%	0.00%

$$when\ \ \hat{\delta}_{interim} = \mu - 1.5 \times se(\hat{\delta}_{interim})$$

True Mean (when θA=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	99.47%	79.51%	21.36%	6.41%	0.00%	0.00%	0.00%
0.85	99.47%	75.93%	15.89%	3.83%	0.00%	0.00%	0.00%
0.9	99.47%	71.01%	10.50%	1.85%	0.00%	0.00%	0.00%
0.95	99.47%	63.07%	5.19%	0.53%	0.00%	0.00%	0.00%

$$when\ \hat{\delta}_{interim} = \mu - 1.0 \times se(\hat{\delta}_{interim})$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	97.82%	61.23%	10.46%	2.42%	0.00%	0.00%	0.00%
0.85	97.82%	56.74%	7.27%	1.30%	0.00%	0.00%	0.00%
0.9	97.82%	51.04%	4.39%	0.55%	0.00%	0.00%	0.00%
0.95	97.82%	42.75%	1.88%	0.12%	0.00%	0.00%	0.00%

$$when\ \hat{\delta}_{interim} = \mu - 0.5 \times se(\hat{\delta}_{interim})$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	92.91%	40.96%	4.37%	0.75%	0.00%	0.00%	0.00%
0.85	92.91%	36.77%	2.82%	0.36%	0.00%	0.00%	0.00%
0.9	92.91%	31.79%	1.53%	0.13%	0.00%	0.00%	0.00%
0.95	92.91%	25.12%	0.56%	0.02%	0.00%	0.00%	0.00%

$$when\ \hat{\delta}_{interim} = \mu$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	82.05%	23.76%	1.53%	0.19%	0.00%	0.00%	0.00%
0.85	82.05%	20.67%	0.90%	0.08%	0.00%	0.00%	0.00%
0.9	82.05%	17.18%	0.44%	0.02%	0.00%	0.00%	0.00%
0.95	82.05%	12.79%	0.13%	0.00%	0.00%	0.00%	0.00%

$$when\ \hat{\delta}_{interim} = \mu + 0.5 \times se(\hat{\delta}_{interim})$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	64.63%	11.94%	0.43%	0.04%	0.00%	0.00%	0.00%
0.85	64.63%	10.05%	0.23%	0.01%	0.00%	0.00%	0.00%
0.9	64.63%	8.01%	0.10%	0.00%	0.00%	0.00%	0.00%
0.95	64.63%	5.58%	0.03%	0.00%	0.00%	0.00%	0.00%

$$when\ \hat{\delta}_{interim} = \mu + 1.0 \times se(\hat{\delta}_{interim})$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	44.30%	5.13%	0.10%	0.01%	0.00%	0.00%	0.00%
0.85	44.30%	4.17%	0.05%	0.00%	0.00%	0.00%	0.00%
0.9	44.30%	3.16%	0.02%	0.00%	0.00%	0.00%	0.00%
0.95	44.30%	2.05%	0.00%	0.00%	0.00%	0.00%	0.00%

$$when\ \hat{\delta}_{interim} = \mu + 1.5 \times se(\hat{\delta}_{interim})$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	26.32%	1.85%	0.02%	0.00%	0.00%	0.00%	0.00%
0.85	26.32%	1.44%	0.01%	0.00%	0.00%	0.00%	0.00%
0.9	26.32%	1.03%	0.00%	0.00%	0.00%	0.00%	0.00%
0.95	26.32%	0.62%	0.00%	0.00%	0.00%	0.00%	0.00%

$$when\ \hat{\delta}_{interim} = \mu + 2.0 \times se(\hat{\delta}_{interim})$$

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	13.56%	0.55%	0.00%	0.00%	0.00%	0.00%	0.00%
0.85	13.56%	0.41%	0.00%	0.00%	0.00%	0.00%	0.00%
0.9	13.56%	0.27%	0.00%	0.00%	0.00%	0.00%	0.00%
0.95	13.56%	0.15%	0.00%	0.00%	0.00%	0.00%	0.00%

Table 5c: Relative weight of stopped studies in a meta-analysis, with one planned interim analysis (one-sided testing) and under the null assumption

True Mean (when θ_A=0.1)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	82.05%	13.94%	0.19%	0.01%	0.00%	0.00%	0.00%
0.85	82.05%	11.33%	0.08%	0.00%	0.00%	0.00%	0.00%
0.9	82.05%	8.58%	0.02%	0.00%	0.00%	0.00%	0.00%
0.95	82.05%	5.48%	0.00%	0.00%	0.00%	0.00%	0.00%

True Mean (when θ_A=0.2)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	82.05%	44.68%	13.94%	6.25%	0.01%	0.00%	0.00%
0.85	82.05%	41.95%	11.33%	4.55%	0.00%	0.00%	0.00%
0.9	82.05%	38.61%	8.58%	2.96%	0.00%	0.00%	0.00%
0.95	82.05%	33.88%	5.48%	1.47%	0.00%	0.00%	0.00%

True Mean (when θ_A=0.25)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	82.05%	52.73%	23.76%	13.94%	0.19%	0.00%	0.00%
0.85	82.05%	50.47%	20.67%	11.33%	0.08%	0.00%	0.00%
0.9	82.05%	47.64%	17.18%	8.58%	0.02%	0.00%	0.00%
0.95	82.05%	43.52%	12.79%	5.48%	0.00%	0.00%	0.00%

True Mean (when θ_A=0.5)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	82.05%	68.66%	52.73%	23.76%	7.44%	1.53%	0.19%
0.85	82.05%	67.61%	50.47%	20.67%	5.55%	0.90%	0.08%
0.9	82.05%	66.27%	47.64%	17.18%	3.74%	0.44%	0.02%
0.95	82.05%	64.24%	43.52%	12.79%	1.96%	0.13%	0.00%

True Mean (when θ_A=0.8)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	82.05%	74.11%	64.82%	59.84%	35.19%	13.94%	6.25%
0.85	82.05%	73.50%	63.45%	58.09%	32.12%	11.33%	4.55%
0.9	82.05%	72.73%	61.71%	55.86%	28.47%	8.58%	2.96%
0.95	82.05%	71.56%	59.10%	52.56%	23.52%	5.48%	1.47%

True Mean (when θ_A=1.0)

Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	82.05%	75.82%	68.66%	64.82%	44.68%	23.76%	13.94%
0.85	82.05%	75.35%	67.61%	63.45%	41.95%	20.67%	11.33%
0.9	82.05%	74.76%	66.27%	61.71%	38.61%	17.18%	8.58%
0.95	82.05%	73.86%	64.24%	59.10%	33.88%	12.79%	5.48%

Table 6: Overall bias for a single study with a stopping rule, whenθA=0.1

a. with one planned interim analysis, one-sided testing and under the design assumption

	True Mean						
Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0079	-0.0003	0.0000	0.0000	0.0000	0.0000	0.0000
0.85	-0.0067	-0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
0.9	-0.0053	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.95	-0.0035	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

b. With two planned interim analysis, one-sided testing and under the design assumption

	True Mean						
Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	-0.0082	-0.0006	0.0000	0.0000	0.0000	0.0000	0.0000
0.85	-0.0069	-0.0003	0.0000	0.0000	0.0000	0.0000	0.0000
0.9	-0.0057	-0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
0.95	-0.0047	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 7: Probability of stopping, when θ_A=0.1

a. Probability of a study stopping at a single planned interim analysis, under the design assumption, one-sided testing

Power	True Mean 0	0.1	0.2	0.25	0.5	0.8	1
0.8	24.47%	0.38%	0.00%	0.00%	0.00%	0.00%	0.00%
0.85	20.35%	0.16%	0.00%	0.00%	0.00%	0.00%	0.00%
0.9	15.81%	0.05%	0.00%	0.00%	0.00%	0.00%	0.00%
0.95	10.40%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%

b. Probability of a study stopping at the 1st interim analysis with two planned interim analysis, under the design assumption, one-sided testing

Power	True Mean 0	0.1	0.2	0.25	0.5	0.8	1
0.8	3.20%	0.03%	0.00%	0.00%	0.00%	0.00%	0.00%
0.85	1.89%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%
0.9	0.91%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
0.95	0.27%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

c. Probability of a study stopping at the 2nd interim analysis with two planned interim analysis, under the design assumption, one-sided testing

Power	True Mean 0	0.1	0.2	0.25	0.5	0.8	1
0.8	51.69%	1.58%	0.00%	0.00%	0.00%	0.00%	0.00%
0.85	50.04%	0.85%	0.00%	0.00%	0.00%	0.00%	0.00%
0.9	47.20%	0.36%	0.00%	0.00%	0.00%	0.00%	0.00%
0.95	42.12%	0.09%	0.00%	0.00%	0.00%	0.00%	0.00%

Table 8: Probability of stopping, when θ_A=0.1

a. Probability of a study stopping at a single planned interim analysis (one-sided), under the null assumption, one-sided testing

Power	True Mean 0	0.1	0.2	0.25	0.5	0.8	1
0.8	90.14%	24.47%	0.38%	0.01%	0.00%	0.00%	0.00%
0.85	90.14%	20.35%	0.16%	0.00%	0.00%	0.00%	0.00%
0.9	90.14%	15.81%	0.05%	0.00%	0.00%	0.00%	0.00%
0.95	90.14%	10.40%	0.01%	0.00%	0.00%	0.00%	0.00%

b. Probability of a study stopping at the 1st interim analysis with two planned interim analysis, under the null assumption, one-sided testing

Power	True Mean 0	0.1	0.2	0.25	0.5	0.8	1
0.8	91.67%	40.74%	3.20%	0.39%	0.00%	0.00%	0.00%
0.85	91.67%	36.44%	1.89%	0.16%	0.00%	0.00%	0.00%
0.9	91.67%	31.27%	0.91%	0.05%	0.00%	0.00%	0.00%
0.95	91.67%	24.26%	0.27%	0.01%	0.00%	0.00%	0.00%

c. Probability of a study stopping at the 2nd interim analysis with two planned interim analysis, under the null assumption, one-sided testing

Power	True Mean 0	0.1	0.2	0.25	0.5	0.8	1
0.8	4.10%	4.20%	0.04%	0.00%	0.00%	0.00%	0.00%
0.85	4.10%	3.61%	0.01%	0.00%	0.00%	0.00%	0.00%
0.9	4.10%	2.88%	0.00%	0.00%	0.00%	0.00%	0.00%
0.95	4.10%	1.92%	0.00%	0.00%	0.00%	0.00%	0.00%

Table 9A: Relative weight of stopped studies in a meta-analysis, when θ_A=0.1

a. with one planned interim analysis, one-sided testing and under the design assumption

Power	True Mean 0	0.1	0.2	0.25	0.5	0.8	1
0.8	13.94%	0.19%	0.00%	0.00%	0.00%	0.00%	0.00%
0.85	11.33%	0.08%	0.00%	0.00%	0.00%	0.00%	0.00%
0.9	8.58%	0.02%	0.00%	0.00%	0.00%	0.00%	0.00%
0.95	5.48%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

b. that stopped at the 1st interim analysis with two planned interim analysis, one-sided testing and under the design assumption

Power	True Mean 0	0.1	0.2	0.25	0.5	0.8	1
0.8	1.63%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%
0.85	0.95%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
0.9	0.46%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
0.95	0.14%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

c. that stopped at the 2nd interim analysis with two planned interim analysis, one-sided testing and under the design assumption

Power	True Mean 0	0.1	0.2	0.25	0.5	0.8	1
0.8	44.06%	1.07%	0.00%	0.00%	0.00%	0.00%	0.00%
0.85	41.42%	0.57%	0.00%	0.00%	0.00%	0.00%	0.00%
0.9	37.98%	0.24%	0.00%	0.00%	0.00%	0.00%	0.00%
0.95	32.84%	0.06%	0.00%	0.00%	0.00%	0.00%	0.00%

Table 9B: Relative weight of stopped studies in a meta-analysis, when θ_A=0.1

a. with one planned interim analysis, one-sided testing and under the null assumption

	True Mean						
Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	82.05%	13.94%	0.19%	0.01%	0.00%	0.00%	0.00%
0.85	82.05%	11.33%	0.08%	0.00%	0.00%	0.00%	0.00%
0.9	82.05%	8.58%	0.02%	0.00%	0.00%	0.00%	0.00%
0.95	82.05%	5.48%	0.00%	0.00%	0.00%	0.00%	0.00%

b. that stopped at the 1st interim analysis with two planned interim analysis, one-sided testing and under the null assumption

	True Mean						
Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	84.62%	25.58%	1.63%	0.20%	0.00%	0.00%	0.00%
0.85	84.62%	22.28%	0.95%	0.08%	0.00%	0.00%	0.00%
0.9	84.62%	18.53%	0.46%	0.02%	0.00%	0.00%	0.00%
0.95	84.62%	13.80%	0.14%	0.00%	0.00%	0.00%	0.00%

c. that stopped at the 2nd interim analysis with two planned interim analysis, one-sided testing and under the null assumption

	True Mean						
Power	0	0.1	0.2	0.25	0.5	0.8	1
0.8	88.73%	22.93%	1.12%	0.13%	0.00%	0.00%	0.00%
0.85	88.73%	19.53%	0.65%	0.05%	0.00%	0.00%	0.00%
0.9	88.73%	15.78%	0.31%	0.02%	0.00%	0.00%	0.00%
0.95	88.73%	11.26%	0.09%	0.00%	0.00%	0.00%	0.00%

Table 9C: Relative weight of stopped studies in a meta-analysis, when θ_A=0.1

a. with one planned interim analysis, one-sided testing and under the current trend assumption

Power	True Mean 0	0.1	0.2	0.25	0.5	0.8	1
0.8	99.90%	14.33%	0.00%	0.00%	0.00%	0.00%	0.00%
0.85	99.90%	9.39%	0.00%	0.00%	0.00%	0.00%	0.00%
0.9	99.90%	5.14%	0.00%	0.00%	0.00%	0.00%	0.00%
0.95	99.90%	1.80%	0.00%	0.00%	0.00%	0.00%	0.00%

b. that stopped at the 1st interim analysis with two planned interim analysis, one-sided testing and under the current trend assumption

Power	True Mean 0	0.1	0.2	0.25	0.5	0.8	1
0.8	100.00%	54.11%	0.00%	0.00%	0.00%	0.00%	0.00%
0.85	100.00%	40.51%	0.00%	0.00%	0.00%	0.00%	0.00%
0.9	100.00%	25.67%	0.00%	0.00%	0.00%	0.00%	0.00%
0.95	100.00%	10.79%	0.00%	0.00%	0.00%	0.00%	0.00%

c. that stopped at the 2nd interim analysis with two planned interim analysis, one-sided testing and under the current trend assumption

Power	True Mean 0	0.1	0.2	0.25	0.5	0.8	1
0.8	100.00%	46.75%	0.00%	0.00%	0.00%	0.00%	0.00%
0.85	100.00%	34.00%	0.00%	0.00%	0.00%	0.00%	0.00%
0.9	100.00%	21.18%	0.00%	0.00%	0.00%	0.00%	0.00%
0.95	100.00%	9.07%	0.00%	0.00%	0.00%	0.00%	0.00%

*

Table 10: An Example of One Interim Analysis, Two-Sided Testing

1/2-sided	**α**	**γ**	**event N**	**HR_expected**	**event n**	**HR_observed**
2	0.05	0.15	650	0.77	343	0.89

I_1	I_K	**t**	z_t	B_t
85.75	162.5	0.528	1.079	0.784

Z^*_{1H}			Z^*_{1L}		
Null	Current Trend	Design	Null	Current Trend	Design
1.718	0.906	-0.449	-1.718	-0.906	-3.884

δ^*_{1H}			δ^*_{1L}		
Null	Current Trend	Design	Null	Current Trend	Design
0.185	0.098	-0.048	-0.185	-0.098	-0.419

HR^*_L			HR^*_H			**CP**		
Null	Current Trend	Design	Null	Current Trend	Design	Null	Current Trend	Design
0.831	0.907	1.050	1.204	1.103	1.521	4.35%	24.50%	71.85%

$\theta_{Futility}$			$\theta'_{Futility}$			**Pr(stop)**		
Null	Current Trend	Design	Null	Current Trend	Design	Null	Current Trend	Design
-0.703	-1.514	-2.869	-4.138	-3.327	-6.304	24.11%	6.46%	0.21%

Under-estimation in tRCTs			**Over-estimation in non-tRCTs**		
Null	Current Trend	Design	Null	Current Trend	Design
-0.1396	-0.2094	-0.3414	0.0443	0.0145	0.0007

Overall Bias			**Relative weight of tRCTs in a meta-analysis**		
Null	Current Trend	Design	Null	Current Trend	Design
-0.0168	-0.0068	-0.0004	13.71%	3.34%	0.10%

Note:

event N = target number of events event n = number of events at the interim analysis

HR_expectd = target(Expected) hazard ratio HR_observed = observed hazard ratio

Z^*_{1H} , Z^*_{1L} = upper and lower threshold for the test statistic at the interim analysis

δ^*_{1H} , δ^*_{1L} = upper and lower threshold for the treatment effect size at the interim analysis

HR^*_H, HR^*_L = upper and lower threshold for the hazard ratio at the interim analysis

Table 11: An Example of One Interim Analysis, One-Sided Testing

1/2-sided	α	γ	n_1	n_2	p_1'	p_2'	N_1	N_2	p_1	p_2
1	0.05	0.15	60	70	0.83	0.89	112	112	0.65	0.85

I_1	I_K	t	z_t	B_t
266.65	315.49	0.8452	0.980	0.901

Z_1^*			CP		
Null	Current Trend	Design	Null	Current Trend	Design
1.346	1.1661	0.747	2.93%	7.05%	31.09%

$\theta_{Futility}$			Pr(stop)		
Null	Current Trend	Design	Null	Current Trend	Design
-1.9203	-2.0998	-2.5185	2.74%	1.79%	0.59%

Under-estimation in tRCTs			Over-estimation in non-tRCTs		
Null	Current Trend	Design	Null	Current Trend	Design
-0.1410	-0.1508	-0.1739	0.0020	0.0014	0.0005

Overall Bias			Relative weight of tRCTs in a meta-analysis		
Null	Current Trend	Design	Null	Current Trend	Design
0.0019	0.0013	0.0005	1.39%	0.90%	0.30%

Note:

N_1, N_2 = target number of patients in Cy and Cy+ATG group

p_1, p_2 = target 1 year survival rate in Cy and Cy+ATG group

n_1, n_2 = number of patients in Cy and Cy+ATG group

p_1' , p_2' = 1 year survival rate in Cy and Cy+ATG group at the interim analysis

Bibliography

[1] *Interim analysis*. http://en.wikipedia.org/wiki/Interim_analysis. Oct. 16, 2014.

[2] Halperin, M., Lan, K.K.G., Ware, J.H. Johnson N.J. and DeMets, D.L. (1982) *An Aid to Data Monitoring in Long-Term Clinical Trials*. Control Clin Trials 3:311–323A

[3] Lan, K. K. G., Simon, R. and Halperin, M. (1982). *Stochastically Curtailed Tests in Long-Term Clinical Trials Theory*. Communications in Statistics, Part C: Sequential Analysis, **Vol.1: Issue 3**, pp. 207-219

[4] Lan K.K.G., Wittes J. (1988) *The B-Value: A Tool for Monitoring Data*. Biometrics **44**: 579–585

[5] Walter, S., Guyatt, G., Bassler, D., Briel, M. and Ramsay, T. (2014) *Quantifying The Bias in The Estimated Treatment Effect in Randomised Trials Having Interim Analyses and A Rule for Early Stopping*. (Unpublished manuscript).

[6] *Clinical endpoint*. http://en.wikipedia.org/wiki/Clinical_endpoint. Oct. 16, 2014.

[7] Chinn S. *A Simple Method for Converting an Odds Ratio to Effect Size for Use in Meta-Analysis*. Statistics in Medicine 2000; **19** 3127-3131.

[8] Truncated normal distribution. http://en.wikipedia.org/wiki/Truncated_normal_distribution. Oct. 16, 2014.

[9] He, P., Lai, T.L. and Liao, O.Y. (2012) *Futility Stopping in Clinical Trials*. Statistics and Its Interface; **Vol. 5** 415-423.

[10] Lachin, J.M. (2005) *A Review of Methods For Futility Stopping Based on Conditional Power*. Statistics in Medicine; **Vol 24: Iss. 18**, 2747-2764.

[11] Freidlin, B. Ch12 *Futility Analysis*. pp.176. Balakrishnan, N. ed. *Methods and Applications of Statistics in Clinical Trials*. Hoboken, New Jersey: John Wiley & Sons, 2014.

[12] Proschan, M.A., Lan, K.K.G. and Wittes, J.T. *Statistical Monitoring of Clinical Trials: A Unified Approach*. New York, New York: Springer Science+Business Medis, 2006, p.63

[13] Jennison, Christopher. and Turnbull, Bruce.W. *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton and London: Chapman and Hall/CRC, 2000, p.206,

[14] Carducci, M.A., Saad, F., Abrahamsson, P.A., Dearnaley, D.P., Schulman, C.C., North, S.A., Sleep, D.J., Isaacson, J.D., Nelson, J.B.; Atrasentan Phase III Study Group Institutions. *A Phase 3 Randomized Controlled Trial of The Efficacy and Safety of Atrasentan in Men With Metastatic Hormone-Refractory Prostate Cancer.* Cancer. 2007 Nov 1;110(9):1959-66.

[15] Cohen, Jocob. *Statistical Power Analysis for The Behavioral Sciences*. p.25. New York, New York: Academic Press, 1976.

[16] *Effect Size*. http://en.wikipedia.org/wiki/Effect_size. Oct. 20, 2014.

[17] Champlin, R.E., Perez, W.S., Passweg, J.R., Klein, J.P., Camitta, B.M., Gluckman, E., Bredeson, C.N., Eapen, M. and Horowitz, M.M. (2007) *Bone Marrow Transplantation for Severe Aplastic Anemia: A Randomized Controlled Study of Conditioning Regimens*. Blood. 109(10):4582-5.

[18] Stegert, M., Kasenda, B., Elm, E.V., You, J., Blümle, A., Tomonaga, Y., Saccilotto, R., Amstutz, A., Bengough, T., Meerpohl, J.J., Tikkinen, K.A.O., Neumann, I., Carrasco-Labra, A., Faulhaber, M., Mulla, S., Mertz, D., Akl, E.A., Bassler, D, Busse, J.W., Ferreira-González, I., Lamontagne, F., Nordmann, A., Gloy, V., Olu, K.K., Raatz, H.,Moja, L., Rosenthal, R., Ebrahim, S., Schandelmaier, S., Sun, X., Vandvik, P.O., Johnston, B.C., Walter, M.A., Burnand, B., Schwenkglenks, M., Hemkens, L., Bucher, H.C., Guyatt, G.H. and Briel, M. (2014) *Clinical trials that stop early do so mostly independent of pre-planned interim analyses and stopping rules.* (Unpublished manuscript).