筑波大学

博士（医学）学位論文

Conditional bias-adjusted estimator in clinical trials with interim analysis

（中間解析を伴う臨床試験における条件付き
バイアス調整推定量に関する研究）

２０１８

筑波大学大学院博士課程人間総合科学研究科

志　村　将　司

Contents

List of Tables

List of Figures

Preface

Randomized controlled trials (RCTs) are clinical trials wherein at least two interventions are evaluated and are the gold standard to establish evidence in clinical research. A new experimental treatment must demonstrate efficacy and safety comparable to that of a reference treatment for a cohort of individuals with the targeted disease in RCTs. In confirmatory trials including phase III studies, the sample sizes and number of events are calculated to achieve a desired power (generally 80% or 90%) before the start of RCTs given a planned effect size. The sample size and number of events must be indicated in the study protocol with the rationale.

It is important to determine the effect size to estimate the sample size because the actual power of the trials depends on the planned effect size. If the actual effect size is smaller than the planned effect size, the actual power may markedly decrease. The planned effect size is generally estimated on the basis of all available prior information. The result of previous clinical trials for the same, similar, and rival drugs are considered useful prior information.

Hay et al.[17] reported a success probability of 60.1% for phase III trials upon investigating the percentage of phase III trials wherein the experimental treatment was approved thereafter. This indicates that the planned effect size used to estimate the sample size may actually be smaller than the true effect size owing to overestimation

– exaggeration– of the efficacious results of the experimental treatment in the previous clinical trials. Thus, reducing the overestimation of the effect size is of clinical importance.

Additionally, in clinical practice, the effect size for candidate drugs is important to select treatment particularly when some available treatments are established for similar diseases. Therefore, it is also crucial to determine the true effect size after confirmatory clinical trials.

Group sequential designs are widely used in clinical trials to determine whether a trial should be terminated early. In such trials, maximum likelihood estimates are often used to describe the difference in efficacy between the experimental and reference treatments; however, these are well known for displaying conditional and unconditional biases.

The bias caused by the interim analysis is categorized into two groups, namely overestimation and underestimation of the efficacy. All of clinical trials with interim analysis result in the four scenarios: Scenario 1) the study stops at the interim analysis for efficacy–overestimation of efficacy, Scenario 2) the study does not stop at the interim analysis for efficacy–underestimation of efficacy, Scenario 3) the study stops at the interim analysis for futility–underestimation of efficacy, and Scenario 4) the study does not stop at the interim analysis for futility–overestimation of efficacy.

The bias caused by the interim analysis that overestimates or underestimates the treatment effect is problematic in the design of clinical research.

For Scenarios 1 and 4 (overestimation), the bias may cause future clinical trials to fail. As described above, a success probability of the confirmatory clinical trial was lower due to overestimation of the treatment effect for the experimental treatment. The bias of overestimation caused by the interim analysis may seriously affect the success probability as well as the “publication bias”. In addition, if the bias is significant, it may also affect the proper choice of drugs in clinical practice.

Underestimation (Scenarios 2 and 3) of the efficacy might increase the required sample size for clinical trials. Therefore, reducing the bias is of clinical importance.

Several statisticians have noted the existence of a bias, i.e., overestimation of the treatment effect when the trial is terminated early at the interim analysis.[47,53] Additionally, in medical journals, numerous researchers have discussed the interpretation of the treatment effect observed in the early terminated trial.[2–4,28,51] Regulators have noted that considering the potential of the overestimation is important when designing and analyzing clinical trials with an interim analysis.[48] Nevertheless, effort to avoid overestimation is limited in practical situations because few have a clear understanding of why overestimation occurs.[55]

Established bias-adjusted estimators include the conditional mean-adjusted estimator (CMAE), conditional median unbiased estimator (CMUE), conditional uniformly minimum variance unbiased estimator (CUMVUE), and weighted estimator (WE). However, their performances have been inadequately investigated. In addition, they may result in absolute non-negligible bias upon early termination of the trial. In this study, we clarify three focal issues of reducing conditional bias as follows.

- **Issue 1 : Comparison of existing conditional bias-adjusted estimators**

  Established bias-adjusted estimators include the conditional mean-adjusted estimator (CMAE), conditional median unbiased estimator, conditional uniformly minimum variance unbiased estimator (CUMVUE), and weighted estimator. However, their performances have been inadequately investigated. In this study, we review the characteristics of these bias-adjusted estimators and compare their conditional bias, overall bias, and conditional mean-squared errors in clinical trials with survival endpoints through simulation studies. The coverage probabilities of the confidence intervals for the four estimators are also evaluated. The first objective is to compare the performance of the existing conditional bias-adjusted estimators in terms of the conditional bias and coverage probability.

- **Issue 2 : Extension of conditional estimation using prior information**

  Shimura et al.[39] compared the performance of existing biasadjusted estimators in settings in which the trial does or does not stop for efficacy at the interim analysis. The use of the CMAE is recommended when in the former case, although the remaining bias may be nonnegligible. We propose a new estimator for adjusting the conditional bias of the treatment effect by extending the idea of the CMAE. This estimator is calculated by weighting the maximum likelihood estimate obtained at the interim analysis and the effect size prespecified when calculating the sample size. We evaluate the performance of the proposed estimator through analytical and simulation studies in various settings in which a trial is stopped for efficacy or futility at the interim analysis. The second objective is to propose a new bias-adjusted estimator to improve the performance of the extending existing estimator.

- **Issue 3 : Application of the bias-adjusted estimators to actual clinical trials**

  In published clinical trials, the bias-adjusted estimators have been rarely reported. The final objective is to quantitatively evaluate the difference in the results via the MLE and bias-adjusted estimators in actual oncological clinical trials.

This dissertation comprises five chapters. Chapter 1 provides a background regarding GSD and the conditional bias. The primary topic in Chapter 2 is to compare the existing conditional bias-adjusted estimators. This chapter is based on Shimura et al.[39] In Chapter 3, the primary focus is to propose a new conditional bias-adjusted estimator by extending the existing estimator. This chapter is entirely based on Shimura et al.[40] Chapter 4 highlights the application result of the conditional bias-adjusted estimators to 19 oncological clinical trials. This chapter is based on Shimura et al.[41] and revealed how clinical trials have suppress the conditional bias without using the conditional bias-adjusted estimators. Finally, Chapter 5 discusses the issues related to this study and presents the conclusions of this study.

Chapter 1

Introduction

1.1 Two-stage group sequential design

Unlike a fixed sample design, wherein the sample size in the clinical trial is fixed, with no interim analyses being performed, group sequential design (GSD) can help terminate potential early trials before completion. GSD controls the type I error rate of the entire trial by suppressing the significance level of each stage. The reasons for conducting interim analyses in clinical trials can be categorized in three classes: ethical, economic, and administrative.[20] Depending on the target disease, the clinical trial duration may be several years. It is ethical to promptly approve (discontinue) an effective (ineffective or harmful) treatment. In addition, GSD allows a sponsor or investigator, who conducts clinical trials, to reduce resources and costs via early termination of the trial. For the administrative advantage, validity of the study procedures can be confirmed after starting the trial. The Food and Drug Administration has issued guidance on adaptive designs including GSD and describes the design as being well-understood.[48] In reality, GSD is widely used to determine whether a trial should be terminated early. Of these, two-stage designs are one of the simplest because of their ease of conductance and interpretation.[8, 38, 42, 45]

1.2 Conditional bias via the two-stage group sequential design

GSD has many advantages; however, it has a disadvantage in terms of estimating the effect size. At the end of the trial, the effect size of the experimental treatment is estimated relative to that of the reference. A maximum likelihood estimate (MLE) is generally obtained to quantify the effect size; however, the MLE is biased in GSDs.[21, 33, 53] As described in Section 1.2, GSD can control the type I error rate; however, ths does not address the appropriate estimation of the effect size.

To explain how the MLE causes the bias, we initially consider hypothetical trials using GSD, e.g., trials with the primary endpoint of overall survival time including a total of 300 enrolled patients and 150 required deaths. The trial is terminated early for efficacy if the hazard ratio becomes lower than a termination criterion (e.g., 0.56). This termination criterion is called the stopping boundary, which is expressed by several scales such as the p-value and the upper limit of the confidence interval; however, these are essentially equivalent. The allocation ratio between the treatment and control groups was 1:1. The interim analysis was performed when 75 deaths were observed. To determine the termination boundary, a Lan–DeMets alpha-spending function with an O'Brien–Fleming type was used. The true hazard ratio was set to 1.0. We designed hypothetical clinical trials using computer simulations with 1,000 replicates and calculated the hazard ratios at the 75 and 150 deaths.

Figure 1.1 shows the sequence of the hazard ratios at each analysis. The true hazard ratio is 1.0 (dashed reference line). That is, the hazard ratios are expected to be approximately 1.0. The solid reference line describes the termination boundary for efficacy at the interim analysis. The red lines indicate the sequence of the hazard ratios for the trials that were terminated at the interim analysis for efficacy. Most of those hazard ratios for the terminated trials approached 1.0 at the final analysis. Thus, we may overestimate the effect size if we interpret the results about the effect size, only on the basis of the interim analysis.

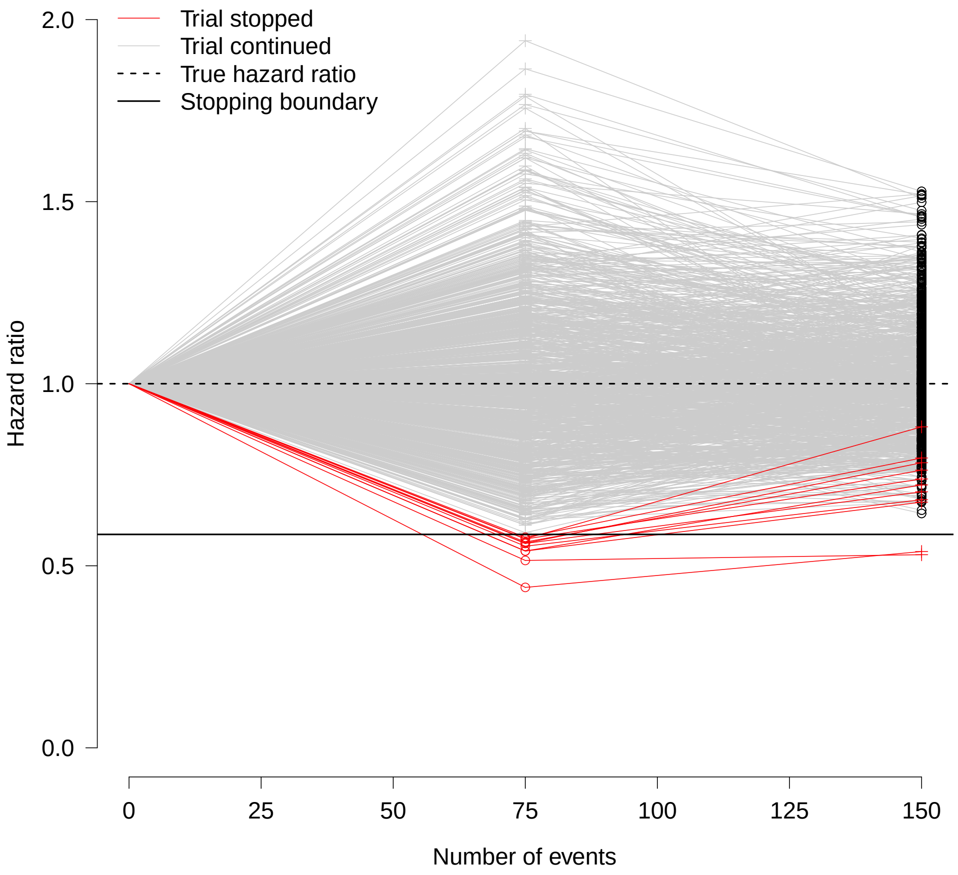


Figure 1.1: Sequence of the hazard ratios at the interim and final analyses. Circles (plus signs) indicate observed (unobserved) hazard ratios

Figure 1.2 shows the frequency distribution of the hazard ratios at the interim analysis in the figure 1.1. The solid, dashed, and dotted reference lines show the termination boundary of 0.56, true hazard ratio of 1.0, and the average hazard ratio of 0.52 for the trial terminated early for efficacy, respectively. The red area corresponds to the hazard ratios of the trial terminated at the interim analysis for efficacy. If the hazard ratio > 0.56, the result of the interim analysis would not be observed until the final analysis. The difference between this average and the true hazard ratios, $0.52 - 1.0 = -0.48$, was a problematic overestimation of the GSD, implying that a hazard ratio of 0.48 was overestimated.

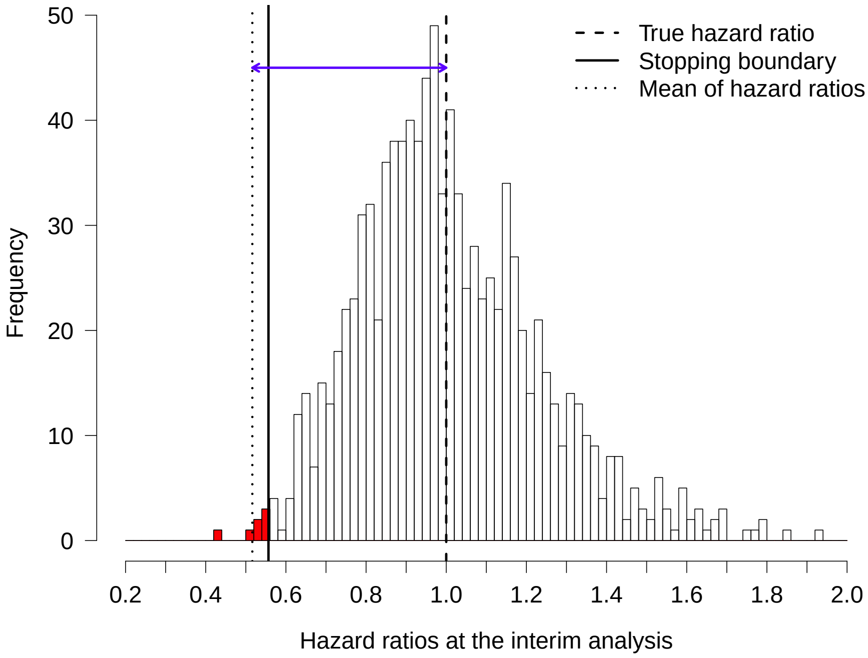


Figure 1.2: Observed and unobserved hazard ratios at the interim analysis

This dissertation focuses on the comparison of the treatment effect both at the interim analysis when the trial stopped and at the final analysis when the trial did not stop. The follow-up analysis after the trial stopped for efficacy is not considered in this research. This is because the comparison of the treatment effect after the interim analysis with positive result is difficult to interpret due to the confounding by treatment switching and the potential bias on unblinded assessments. Actually, the follow-up analysis has not frequently been reported.

1.3 Statistical definition

1.3.1 Two-stage group sequential design

The principles of GSD were originally introduced by Pocock[34] and O'Brien and Fleming.[29] The sequence of the standardized test statistics approximates the canonical joint distribution in the survival data as a normal response.[20] Hence, survival responses can be considered a normal response. In the two-stage design setting, let m ($m = 1, 2$) be the stage index. Let us suppose that θ is a true log hazard ratio, which is the true difference in the treatment efficacy of an experimental drug compared with that of a reference drug. The experimental drug is more efficacious than the reference drug if the hazard ratio is less than 1 and the log hazard ratio is less than 0. A significant advantage for the experimental drug is considered if the null hypothesis H_0: $\theta = 0$ is rejected in any stage. Further, suppose that $\hat{\theta}_{MLE,m}$ is an MLE of θ in the mth stage. Let σ_m^2 be the variance of $\hat{\theta}_{MLE,m}$, the reciprocal of the Fisher information I_m^{-1}. In our setting, I_2, the Fisher information in the second stage is equal to the maximum information. In a GSD, in the sequence $Z_m = \hat{\theta}_{MLE,m}/\sqrt{\sigma_m^2}$, ($m = 1, 2$), the standardized test statistics have a canonical joint distribution and follow the multivariate normal distribution. Thus, the multivariate normal distribution of the MLE can be obtained via simple transformation of the standardized test statistic Z_m. In particular, in the two-stage design, $(\hat{\theta}_{MLE,1}, \hat{\theta}_{MLE,2})$ follows a bivariate normal distribution:[20]

$$\begin{pmatrix} \hat{\theta}_{MLE,1} \\ \hat{\theta}_{MLE,2} \end{pmatrix} \sim MVN \left\{ \begin{pmatrix} \theta \\ \theta \end{pmatrix}, \begin{pmatrix} {\sigma_1}^2 & {\sigma_2}^2 \\ {\sigma_2}^2 & {\sigma_2}^2 \end{pmatrix} \right\}.$$

Notably, this distribution does not account for the decision to stop early made in stage 1. Here, $R_m^{(MLE)} = (-\infty, a_m) \cup (b_m, \infty)$ is the rejection region in stage m in the MLE scale and a_m and b_m are the termination boundaries in stage m for efficacy and futility, respectively. If $b_m = \infty$, only efficacy termination is considered. $R_m^{(MLE)}$ is determined by the alpha-spending function introduced by Lan and DeMets.[24] For a GSD with one interim analysis, $R_2^{(MLE)}$ is used when the trial continued to the second stage. The study is terminated at the first point at which $\hat{\theta}_{MLE,m}$ is in the rejection region.

1.3.2 Conditional bias

The overall bias is defined as the weighted average of the conditional expectations of the difference between an estimator and the true parameter:

$$E\left(\hat{\theta} - \theta\right) = \sum_{m=1}^{2} P(M = m) E\left(\hat{\theta} - \theta \middle| M = m\right),$$

where $\hat{\theta}$ is an estimator, M is a random variable expressing the termination stage, and $P(M = m)$ is the termination probability at each interim analysis.[27] The conditional expectation of $\hat{\theta}_{MLE,1}$ given stopping stage $m = 1$ in a trial with one interim analysis for efficacy is

$$\begin{aligned}
E\left(\hat{\theta}_{MLE,1}\middle| M=1\right) &= \int_{-\infty}^{\infty} xf(x|M=1)dx \\
&= \frac{1}{\Phi\left(\frac{a_1-\theta}{\sigma_1}\right)} \int_{-\infty}^{a_1} xf(x)dx \\
&= \theta - \sigma_1 \frac{\phi\left(\frac{a_1-\theta}{\sigma_1}\right)}{\Phi\left(\frac{a_1-\theta}{\sigma_1}\right)}, \qquad (1.1)
\end{aligned}$$

where $f(x|M=1)$ is the conditional probability density function of $\hat{\theta}_{MLE,1}$ given stopping stage $m=1$, $f(x)$ is the probability density function of $\hat{\theta}_{MLE,1}$, $\phi(\bullet)$ is a standard normal probability density function, and $\Phi(\bullet)$ is a standard normal cumulative distribution function. The conditional bias in stage 1 is defined as

$$B(\theta|\sigma_1, a_1, M=1) = E\left(\hat{\theta}_{MLE,1} - \theta\middle| M=1\right) = -\sigma_1 \frac{\phi\left(\frac{a_1-\theta}{\sigma_1}\right)}{\Phi\left(\frac{a_1-\theta}{\sigma_1}\right)}. \qquad (1.2)$$

Equation (1.2) estimates the interim analysis with the efficacy boundary. Even in the presence of two termination boundaries for efficacy and futility, the bias is conditioned by the efficacy or futility termination. Thus, the formula can easily be modified for futility. Similarly, the conditional bias in stage 2 is defined as

$$B(\theta|\sigma_1, a_1, \sigma_2, M=2) = E\left(\hat{\theta}_{MLE,2} - \theta\middle| M=2\right) = \left(\frac{\sigma_2^2}{\sigma_1}\right) \frac{\phi\left(\frac{a_1-\theta}{\sigma_1}\right)}{1-\Phi\left(\frac{a_1-\theta}{\sigma_1}\right)}. \quad (1.3)$$

The overall bias may approach zero despite the magnitudes of the conditional biases in each stage being large. This discrepancy implies that a reduction in the conditional bias can be more important than a reduction in the overall bias. An unbiased estimator with a substantial conditional bias may be undesirable, especially if early termination occurs.

Chapter 2

Comparison of existing conditional bias-adjusted estimators

2.1 Introduction

In clinical trials, GSDs are widely used to determine whether a trial should be terminated early. After the completion of group sequential trials, the difference in the treatment effect between the experimental and reference drugs is estimated to measure the former's effectiveness for addressing a certain disease in comparison with the latter. The maximum likelihood estimate (MLE) is then generally used to quantify the difference in the treatment effect between the two groups; however, it is well known that the MLE is biased in GSDs.[21,33,53] To address this issue, some researchers have proposed bias-adjusted estimators. For example, Whitehead[53,54] developed the median unbiased estimator (MUE) and mean-adjusted estimator (MAE) to reduce bias by evaluating the distribution of the MLE. Emerson and Fleming[12] proposed the uniformly minimum variance unbiased estimator (UMVUE), based on the Rao–Blackwell theorem, and showed that the biases of these three estimators are smaller than those of the MLE. Liu and Hall[26] clarified that there is a UMVUE in the class of estimators does not depend on future stopping information. These estimators are referred to as "unconditional" bias-adjusted estimators and can be adapted regardless of whether the trial terminates at the interim analysis or continues until the final analysis. However, Troendle and Yu[47] pointed out that unconditional bias-adjusted estimators lead to substantial "conditional" bias, which is defined as the bias conditioned by the stopping stage, because they can be unbiased by the trade-off relationship owing to the bias of the treatment effect between termination and continuation. For example, in the case of planning an interim analysis, the average unconditional estimates with and without early termination may be approximately zero.

A reduction in conditional bias is as important as a reduction in overall bias because, in practice, researchers can only obtain an estimate that is conditional on the stopping stages. Four conditional bias-adjusted estimators have been proposed. Troendle and Yu[47] proposed a conditional MAE (CMAE) and Pepe et al.[32] developed the conditional UMVUE (CUMVUE) in binomial responses by evaluating the conditional distribution of the MLE. Koopmeiners et al.[23] extended the CUMVUE from binomial responses to normal responses and proposed a conditional MUE (CMUE). Zhong and Prentice[56] developed a weighted estimator (WE) as a linear combination of the conditional estimations . Further, some researchers have discussed the performance of confidence interval methods. Ohman Strickland and Casalla[30] showed that a Wald confidence interval of the MLE fails to provide the nominal coverage probability and proposed conditional confidence intervals for normal responses. Pepe et al.[32] also developed a non-parametric bootstrap confidence interval of the CUMVUE for binomial responses.

Although the MUE has been applied to an actual trial,[11, 18, 50] the above-mentioned estimators have not been widely used in practice for two main reasons. First, the performance of existing bias-adjusted estimators has not been evaluated from a practical point of view. Second, few statistical software programs can easily calculate these conditional estimators. With regard to these issues, Zhang et al.[55] discussed the reason why bias-adjusted estimators are not popular.

Koopmeiners et al.[23] compared the performances of the original MLE, stage 2 estimator (an MLE calculated by using only second-stage data when the study continued onto a second stage), CMUE, CMAE, and CUMVUE through a simulation study in which a trial with a sample size of 20−60 does not stop for futility. They found the CUMVUE to be an appropriate method in terms of conditional bias and standard error. However, the performance of bias-adjusted estimators including WEs has been inadequately studied. Although Koopmeiners et al.[23] used the conditional confidence interval of the estimators developed by Ohman Strickland and Casalla,[30] the confidence interval did not achieve the nominal coverage probability. Therefore, we focus on the non-parametric bootstrap confidence interval, which can easily be applied to all bias-adjusted estimators, without distribution assumptions.

In the present study, we compare the performances of the CMUE, CMAE, CUMVUE, and WE in terms of conditional bias, overall bias, and the conditional mean-squared error in group sequential trials with survival data that include an interim analysis. The conditional coverage probability of the confidence interval for the estimators is also evaluated. To this end, we conduct simulation studies under four settings: 1) stopping early for efficacy, 2) not stopping early for efficacy, 3) stopping early for futility, and 4) not stopping early for futility. We focus on the survival response because GSDs are frequently applied to clinical trials in oncology with time-to-event variables such as overall survival time and progression-free survival time.

2.2 Scenarios

In our setting, we consider only one interim analysis and cases where the trial could stop for futility or efficacy but not both at the same time.

There are two types of early stops and four scenarios. Firstly, the trial allows the possibility of early stopping only for efficacy with $a_1 = -\infty$. Scenarios 1 and 2 represent those cases when the study stops ($m = 1$) and does not stop ($m = 2$) early for efficacy, respectively. Secondly, the trial allows the possibility of early stopping only for futility with $b_1 = \infty$. Scenarios 3 and 4 represent those cases when the study stops ($m = 1$) and does not stop ($m = 2$) early for futility, respectively.

2.3 Existing conditional bias-adjusted estimators

2.3.1 With early termination

In this section, we describe only the conditional estimators for efficacy (Scenario 1) because those for futility are calculated by adopting the same approach. The stage 2 estimator and CUMVUE do not exist in Scenarios 1 and 3.

CMAE

The CMAE is derived from the mean-adjusted estimator proposed by Whitehead.[53] Troendle and Yu[47] proposed a simple conditional bias-adjusted estimator by reducing the bias from the MLE under the condition that the bias function was obtained. Note that in the two-stage GSD, the conditional bias functions for the first- and second-order methods are identical when the trial is stopped early. Ideally, this bias would be calculated at the true θ. Again, the CMAE is calculated by reducing the bias from $\hat{\theta}_{MLE,1}$ by using the following formula if the trial is terminated early for efficacy:

$$\hat{\theta}_{CMAE,1} = \hat{\theta}_{MLE,1} - B(\hat{\theta}_{MLE,1}|\sigma_1, a_1, M = 1). \tag{2.1}$$

CMUE

Zhong and Prentice[56] and Koopmeiners et al.[23] extended the MUE to a conditional estimator. The CMUE with study termination for efficacy, denoted by $\hat{\theta}_{CMUE,1}$, is the value of θ obtained by solving

$$0.5 = \frac{\Phi\left(\frac{\theta - \hat{\theta}_{MLE,1}}{\sigma_1}\right)}{\Phi\left(\frac{a_1 - \theta}{\sigma_1}\right)}. \tag{2.2}$$

WE

Zhong and Prentice[56] proposed the WE, defined by the linear combination of a conditional bias-adjusted estimator with a combined estimator, using two-stage data. The idea behind this combination approach comes from the overestimation of the MLE and underestimation of the bias-adjusted estimator. Two types of WEs are derived from two bias-adjusted estimators (i.e., CMAE and CMUE) that can be calculated in each stage. The first weighted estimator (termed WE1) given the stopping at stage 1, is a combination of the MLE and CMAE defined as

$$\hat{\theta}_{WE1,1} = (1 - \hat{K}_{WE1,1})\, \hat{\theta}_{MLE,1} + \hat{K}_{WE1,1}\, \hat{\theta}_{CMAE,1}, \tag{2.3}$$

where $\hat{K}_{WE1,1} = \frac{I_2^{-1}}{I_2^{-1}+(\hat{\theta}_{MLE,1}-\hat{\theta}_{CMAE,1})^2}$ is the weight of each estimator and I_2 is the maximum information. The second weighted estimator (termed WE2) is a combination of the MLE and CMUE defined as

$$\hat{\theta}_{WE2,1} = (1 - \hat{K}_{WE2,1})\, \hat{\theta}_{MLE,1} + \hat{K}_{WE2,1}\, \hat{\theta}_{CMUE,1}, \tag{2.4}$$

where $\hat{K}_{WE2,1} = \frac{I_2^{-1}}{I_2^{-1}+(\hat{\theta}_{MLE,1}-\hat{\theta}_{CMUE,1})^2}$ is the weight of each estimator.

2.3.2 Without early termination

We consider Scenarios 2 and 4 in this section; note that all estimators exist in these scenarios.

CMAE

The CMAE, denoted by $\hat{\theta}_{CMAE,2}$, is the value of θ obtained by solving the following equation, using the Newton–Raphson iteration:

$$\theta - (-1)^{D_{ef}} \left(\frac{\sigma_2^2}{\sigma_1}\right) \times \frac{\phi\left(\frac{-a_1 D_{ef} - b_1(1-D_{ef}) + \theta}{\sigma_1}\right)}{\Phi\left(\frac{-a_1 D_{ef} - b_1(1-D_{ef}) + \theta}{\sigma_1}\right) D_{ef} + \Phi\left(\frac{-\theta + a_1 D_{ef} + b_1(1-D_{ef})}{\sigma_1}\right)(1 - D_{ef})} = \hat{\theta}_{MLE,2}, \tag{2.5}$$

where D_{ef} is an indicator function, which becomes 1 (0) when the trial allows an early stop for efficacy (futility).

CMUE

When the trial continues after the interim analysis, the CMUE, denoted by $\hat{\theta}_{CMUE,2}$, is the value of θ obtained by solving

$$\begin{aligned} 0.5 &= \int_{-\hat{\theta}_{MLE,2}}^{\infty} f(y|m=2)dy \\ &= \int_{-\hat{\theta}_{MLE,2}}^{\infty} \frac{\Phi\left(\frac{-a_1 D_{ef} - b_1(1-D_{ef}) + y}{\sqrt{\sigma_1^2 - \sigma_2^2}}\right) D_{ef} + \Phi\left(\frac{-y + a_1 D_{ef} + b_1(1-D_{ef})}{\sqrt{\sigma_1^2 - \sigma_2^2}}\right)(1 - D_{ef})}{\Phi\left(\frac{-a_1 D_{ef} - b_1(1-D_{ef}) + \theta}{\sigma_1}\right) D_{ef} + \Phi\left(\frac{-\theta + a_1 D_{ef} + b_1(1-D_{ef})}{\sigma_1}\right)(1 - D_{ef})} \times \\ &\qquad \frac{1}{\sigma_2}\phi\left(\frac{-y+\theta}{\sigma_2}\right) dy. \end{aligned} \tag{2.6}$$

CUMVUE

Pepe et al.[32] and Koopmeiners et al.[23] proposed the CUMVUE for binomial and normal responses, respectively. We note that the CUMVUE can only be defined if the study continued onto the second stage:

$$
\begin{aligned}
&E\left[\hat{\theta}_{S2E}\middle|\hat{\theta}_{MLE,2}, m=2\right]\\
&=\hat{\theta}_{MLE,2}-(-1)^{(1-D_{ef})}\frac{\sigma_2^2}{\sqrt{\sigma_1^2-\sigma_2^2}}\times\\
&\qquad\frac{\phi\left(\frac{-a_1D_{ef}-b_1(1-D_{ef})+\hat{\theta}_{MLE,2}}{\sqrt{\sigma_1^2-\sigma_2^2}}\right)}{\Phi\left(\frac{-a_1D_{ef}-b_1(1-D_{ef})+\hat{\theta}_{MLE,2}}{\sqrt{\sigma_1^2-\sigma_2^2}}\right)D_{ef}+\Phi\left(\frac{-\hat{\theta}_{MLE,2}+a_1D_{ef}+b_1(1-D_{ef})}{\sqrt{\sigma_1^2-\sigma_2^2}}\right)(1-D_{ef})}.
\end{aligned}
\tag{2.7}
$$

WE

The definition of WE is independent of the trials' early termination. WE1, given the stopping at stage 2, is a combination of the MLE and CMAE. This is defined as follows:

$$\hat{\theta}_{WE1,2}=(1-\hat{K}_{WE1,2})\,\hat{\theta}_{COM}+\hat{K}_{WE1,2}\,\hat{\theta}_{CMAE,2}, \tag{2.8}$$

where $\hat{\theta}_{COM}$ is a combination estimator, as named by Zhong and Prentice,[56] defined as the combination of the MLE in each stage. $\hat{\theta}_{COM}$ is calculated as follows:

$$\hat{\theta}_{COM}=\frac{I_1}{I_1+I_2}\hat{\theta}_{MLE,1}+\frac{I_2}{I_1+I_2}\hat{\theta}_{MLE,2}, \tag{2.9}$$

where $\hat{K}_{WE1,2}=\frac{I_2^{-1}}{I_2^{-1}+(\hat{\theta}_{COM}-\hat{\theta}_{CMAE,2})^2}$ is the weight of each estimator, I_1 is the Fisher information at stage 1, and I_2 is the Fisher information at stage 2.

WE2 is a combination of the MLE and CMUE. This is defined as follows:

$$\hat{\theta}_{WE2,2} = (1 - \hat{K}_{WE2,2})\,\hat{\theta}_{COM} + \hat{K}_{WE2,2}\,\hat{\theta}_{CMUE,2}, \tag{2.10}$$

where $\hat{K}_{WE2,2} = \frac{I_2^{-1}}{I_2^{-1} + (\hat{\theta}_{COM} - \hat{\theta}_{CMUE,2})^2}$ is the weight of each estimator.

2.4 Bootstrap conditional confidence interval

To estimate the confidence interval of the bias-adjusted estimators as described in Section 2.3, we use the non-parametric bootstrap method originally proposed by Pepe et al.[32] The bootstrap confidence interval can be calculated for any of the proposed conditional bias-adjusted estimators. In our simulation, we compare the performance of the non-parametric bootstrap confidence interval for each estimator.

We obtain the same number of observations drawn with replacement from the original dataset. All the resampled datasets are used for the following procedure and the stopping boundary at the interim analysis is not considered. Next, the MLE, stage 2 estimator, and each conditional bias-adjusted estimator are calculated from the resampled dataset. We repeat the above procedures B times (e.g., $B > 500$) and obtain x_i $(i = 1, 2, ..., B)$, where x_i is the ith independent bootstrap estimate of each estimator. The empirical distribution of x_i is a bootstrap distribution. The cumulative distribution of x is denoted by $F(x)$. The 2.5th and 97.5th percentiles of the bootstrap distribution, $F^{-1}(0.025)$ and $F^{-1}(0.975)$, are used as the lower and upper 95% confidence limits, $\tilde{\theta}_{LCL}$ and $\tilde{\theta}_{UCL}$, respectively.

2.5 Simulation study

2.5.1 Data generation and scenarios

In our simulation study, we assumed a randomized, parallel two-group comparison study with one interim analysis for efficacy or futility. As noted in the Introduction, we set four scenarios: 1) stopping early for efficacy, 2) not stopping early for efficacy, 3) stopping early for futility, and 4) not stopping early for futility. Overall type I and type II errors were set to 0.05 and 0.20, respectively. The Lan–DeMets alpha- or beta-spending function with the Pocock- or O'BrienFleming-type boundary was used. We assumed the information time for the interim analysis to be 35%, 50%, and 70%. The experimental drug is more efficacious than the reference drug if the hazard ratio is less than 1 and the log hazard ratio is less than 0. The planned hazard ratio, which is used to calculate the number of events and sample size, was set to 0.7, corresponding to a log hazard ratio of -0.357. The accrual and follow-up times were set to 3 and 5 years, respectively. We considered the Weibull distribution, which has a hazard function defined as

$$\frac{\lambda t^{\lambda-1}}{\exp(\theta G)},$$

where λ is the shape parameter of the Weibull distribution, t is the time from enrolment, and G is an indicator function, which becomes 1 for the experimental drug and 0 for the reference drug. We set λ to 0.5, 1, and 2. The hazard ratios for each shape parameter, $\exp(-\theta)$, were 1.0, 0.9, 0.7, and 0.5. The number of bootstrap samples for constructing the bootstrap confidence interval was 500. To calculate overall bias, we ran 10,000 simulations. To calculate conditional bias, we set the number of times that the estimation could be obtained to 5,000 for each condition. We needed more than 6,000,000 replications to obtain 5,000 cases when the stopping probability of each condition was small (e.g., when the hazard ratio was 0.5 and the information time was 35% with the O'Brien–Fleming-type boundary in Scenario 3).

We compared the performance of the MLE, stage 2 estimate (S2E), CMAE, CMUE, CUMVUE, WE1, and WE2 defined in Section 2.3. Koopmeiners et al.[23] calculated the bias-adjusted estimators by using an iterative procedure to account for the dependence between θ and σ_2^2. However, during the iteration, the variance in the MLE was occasionally too small, preventing us from obtaining the bias-adjusted estimator (i.e., the CMUE). Therefore, we chose an approach without an iterative procedure in our study.

We evaluated overall bias, conditional bias, the mean-squared error, the conditional mean-squared error, the coverage probability of the nominal 95% confidence intervals, and the width of the confidence intervals for the log hazard ratio for the bias-adjusted estimators. Overall bias was calculated as the average of the difference between the estimator and true log hazard ratio regardless of the stopping stage:

$$\frac{1}{s_1 + s_2} \sum_{k=1}^{s_1+s_2} \left(\exp(\hat{\theta}_k) - \exp(\theta) \right),$$

where s_1 and s_2 are the number of simulations in which the trial stopped and did not stop, respectively. Note that $s_1 + s_2$ is 10,000 for the calculation of overall bias. Conditional bias was calculated as the average of the difference between the estimator and true hazard ratio by stopping stage. Conditional bias at stage i is defined as

$$\frac{1}{s_i}\sum_{k=1}^{s_1+s_2}\left(\exp(\hat{\theta}_k)-\exp(\theta)\right)D(m=i),$$

where $D(\bullet)$ is an indicator function, which becomes 1 when $m=i$ and 0 when $m \neq i$. Note that if $i = 1$, s_1 is 5,000, whereas $s_1 + s_2$ is more than or equal to 5,000. On the contrary, if $i = 2$, s_2 is 5,000, whereas $s_1 + s_2$ is more than or equal to 5,000. The conditional mean-squared error at stage i is defined as

$$\frac{1}{s_i}\sum_{k=1}^{s_1+s_2}\left(\exp(\hat{\theta}_k)-\exp(\theta)\right)^2 D(m=i).$$

The conditional coverage probability is defined as

$$\frac{1}{s_i}\sum_{k=1}^{s_1+s_2} D\left(\tilde{\theta}_{LCL,k} < \theta < \tilde{\theta}_{UCL,k}, m=i\right),$$

where $\tilde{\theta}_{LCL}$ and $\tilde{\theta}_{UCL}$ are the lower and upper confidence limits calculated by the bootstrap method of the kth simulation. Finally, the width of the confidence interval is defined as

$$\frac{1}{s_i}\sum_{k=1}^{s_1+s_2}\left(\exp(\tilde{\theta}_{UCL,k})-\exp(\tilde{\theta}_{LCL,k})\right)D(m=i).$$

All simulation studies were performed with R version 3.1.1.[36]

2.5.2 Results

The results for overall bias, conditional bias, and the conditional coverage probability for Scenarios 1 and 2 when $\lambda = 1$ are shown in Tables 2.1 to 2.5. For Scenarios 3 and 4, the results for futility are discussed next to the results for Scenarios 1 and 2. Each of these tables consists of six rows. Rows 1, 2, and 3 show the cases of the O'Brien–Fleming-type boundary at 35%, 50%, and 75% information time, respectively. Rows 4, 5, and 6 show the cases of Pocock-type boundary at 35%, 50%, and 75% information time, respectively. The columns contain the true hazard ratio, true log hazard ratio, and simulation results of the compared estimators.

Table 2.1 shows the conditional bias when the trial terminated early for efficacy (Scenario 1). The CUMVUE and S2E are not included in Table 2.1 because these are not defined in the case of early termination. The conditional bias for all of the estimators based on the O'Brien–Fleming-type boundary was larger than that of the Pocock-type boundary because the former had a higher stopping boundary at the interim analysis than that of the latter. Therefore, the second term of Equation (1) and the conditional bias became larger when the Pocock-type boundary was used. However, the differences between the two types of boundaries decreased if the experimental drug was more effective and the information time approached 1. The conditional biases for the CMAE, CMUE, WE1, and WE2 were smaller than those for the MLE regardless of the setting. The rank order among the estimators was slightly dependent on the true hazard ratio, information time, and spending function. The conditional bias of the CMAE was smaller than or equal to that of WE1. In addition, the conditional bias of the CMAE was smaller than or equal to that of WE2 except when the O'Brien–Fleming-type boundary was used with 35% information time. The CMUE had a relatively small conditional

Table 2.1: Conditional bias at stage 1 for efficacy ($\lambda = 1$, Scenario 1)

Scenario	HR	logHR	MLE	CMAE	CMUE	WE1	WE2
OF Type	1.0	0.000	−0.569	−0.502	−0.371	−0.551	−0.429
35% IT	0.9	−0.105	−0.473	−0.408	−0.291	−0.456	−0.347
	0.7	−0.357	−0.284	−0.226	−0.137	−0.270	−0.187
	0.5	−0.693	−0.110	−0.067	−0.013	−0.101	−0.052
OF Type	1.0	0.000	−0.444	−0.371	−0.294	−0.424	−0.351
50% IT	0.9	−0.105	−0.350	−0.281	−0.217	−0.331	−0.271
	0.7	−0.357	−0.170	−0.112	−0.070	−0.155	−0.117
	0.5	−0.693	−0.034	−0.003	0.007	−0.028	−0.018
OF Type	1.0	0.000	−0.338	−0.267	−0.229	−0.318	−0.283
70% IT	0.9	−0.105	−0.246	−0.180	−0.152	−0.228	−0.202
	0.7	−0.357	−0.084	−0.037	−0.028	−0.073	−0.064
	0.5	−0.693	0.002	0.015	0.014	0.004	0.003
P Type	1.0	0.000	−0.422	−0.338	−0.297	−0.399	−0.361
35% IT	0.9	−0.105	−0.329	−0.250	−0.213	−0.308	−0.274
	0.7	−0.357	−0.157	−0.093	−0.075	−0.142	−0.125
	0.5	−0.693	−0.029	0.005	0.007	−0.023	−0.020
P Type	1.0	0.000	−0.350	−0.272	−0.246	−0.329	−0.304
50% IT	0.9	−0.105	−0.260	−0.189	−0.170	−0.242	−0.224
	0.7	−0.357	−0.098	−0.045	−0.040	−0.086	−0.081
	0.5	−0.693	−0.004	0.014	0.013	−0.001	−0.001
P Type	1.0	0.000	−0.295	−0.225	−0.208	−0.276	−0.260
70% IT	0.9	−0.105	−0.205	−0.141	−0.131	−0.188	−0.179
	0.7	−0.357	−0.058	−0.018	−0.021	−0.050	−0.052
	0.5	−0.693	0.006	0.013	0.012	0.007	0.007

OF Type: O'Brien–Fleming-type boundary, P Type: Pocock-type boundary, IT: Information time, HR: Hazard ratio.

bias among the conditional bias-adjusted estimators.

Table 2.2 shows the coverage probability when the trial terminated early for efficacy (Scenario 1). The coverage probability of the MLE, CMUE, and WE2 was relatively low when the true hazard ratio was 1.0 or 0.9. The CMAE outperformed the other methods in this scenario, although it tended to exceed the nominal confidence level. The coverage probability of WE1 and WE2 was generally lower than that of the CMAE and CMUE, especially when the true hazard ratio was 1.0. This is because WE1 and WE2 consist of the linear combination of the MLE and the CMAE or CMUE, and the coverage probability of the MLE lies relatively far from the nominal confidence level. Although the conditional bias of the CMAE was larger than that of the CMUE, the mean-squared error of the CMAE was smaller. In addition, the conditional coverage probability of the CMAE was relatively better than that of the CMUE. Therefore, the CMAE would be preferable in terms of the trade-off relationship between the conditional bias, mean-squared error, and coverage probability.

Table 2.3 indicates the conditional bias when the trial did not terminate early for efficacy (Scenario 2). The MLE, stage 2 estimator, and conditional bias-adjusted estimators in the second stage were calculated regardless of whether the null hypothesis was rejected. The stage 2 estimator and five conditional bias-adjusted estimators had much smaller conditional biases than those of the MLE. The CUMVUE and S2E were unbiased. The conditional bias of the CUMVUE was smaller than that of the S2E when the information time was 70%. This is because the CUMVUE had more information than the S2E and the asymptotics worked well in the CUMVUE. The conditional bias of WE1 and WE2 was somewhat greater than that of the CMAE, CMUE, and CUMVUE. The conditional mean-squared error of the MLE was smaller than that of any other estimators at stage 2 (see Table 2.4). The conditional mean-squared error of the S2E was larger than those of the other estimators. The absolute conditional bias when the trial did not terminate in Table 2.3 was smaller than that obtained when the trial terminated early as presented in Table 2.1 because the estimators were calculated including the unbiased stage 2 data. This was particularly the case when the MLE was used. WE1 and WE2 could not correct the negative conditional bias sufficiently as shown in Table 2.1. This is because WE1 and WE2 were derived from the MLE, which had a very large

Table 2.2: Conditional coverage probability at stage 1 for efficacy ($\lambda = 1$, Scenario 1)

Scenario	HR	logHR	MLE	CMAE	CMUE	WE1	WE2
OF Type	1.0	0.000	0.000	0.886	0.000	0.749	0.000
35% IT	0.9	−0.105	0.000	0.936	0.067	0.866	0.000
	0.7	−0.357	0.402	0.976	0.939	0.940	0.813
	0.5	−0.693	0.923	0.983	0.978	0.964	0.955
OF Type	1.0	0.000	0.000	0.949	0.000	0.878	0.000
50% IT	0.9	−0.105	0.027	0.965	0.385	0.924	0.003
	0.7	−0.357	0.865	0.985	0.976	0.961	0.932
	0.5	−0.693	0.961	0.979	0.975	0.966	0.957
OF Type	1.0	0.000	0.016	0.972	0.000	0.928	0.000
70% IT	0.9	−0.105	0.637	0.984	0.790	0.959	0.048
	0.7	−0.357	0.949	0.986	0.979	0.968	0.949
	0.5	−0.693	0.963	0.962	0.969	0.963	0.966
P Type	1.0	0.000	0.060	0.966	0.000	0.928	0.000
35% IT	0.9	−0.105	0.561	0.975	0.546	0.948	0.000
	0.7	−0.357	0.912	0.983	0.967	0.961	0.908
	0.5	−0.693	0.961	0.978	0.972	0.963	0.952
P Type	1.0	0.000	0.178	0.974	0.000	0.936	0.000
50% IT	0.9	−0.105	0.704	0.981	0.753	0.960	0.010
	0.7	−0.357	0.944	0.983	0.972	0.965	0.933
	0.5	−0.693	0.968	0.973	0.970	0.968	0.966
P Type	1.0	0.000	0.315	0.976	0.000	0.944	0.000
70% IT	0.9	−0.105	0.811	0.984	0.872	0.960	0.096
	0.7	−0.357	0.953	0.983	0.973	0.963	0.941
	0.5	−0.693	0.957	0.957	0.972	0.957	0.972

OF Type: O'Brien–Fleming-type boundary, P Type: Pocock-type boundary, IT: Information time, HR: Hazard ratio.

Table 2.3: Conditional bias at stage 2 for efficacy ($\lambda = 1$, Scenario 2)

Scenario	HR	logHR	MLE	S2E	CMAE	CMUE	CUMVUE	WE1	WE2
OF Type	1.0	0.000	0.012	0.016	0.011	0.012	0.012	0.014	0.014
35% IT	0.9	−0.105	0.012	0.017	0.010	0.010	0.011	0.012	0.012
	0.7	−0.357	0.013	0.012	0.004	0.005	0.007	0.012	0.012
	0.5	−0.693	0.030	0.006	0.000	0.001	0.003	0.022	0.023
OF Type	1.0	0.000	0.012	0.023	0.009	0.010	0.011	0.012	0.012
50% IT	0.9	−0.105	0.012	0.019	0.005	0.006	0.009	0.011	0.011
	0.7	−0.357	0.030	0.014	0.003	0.005	0.009	0.021	0.022
	0.5	−0.693	0.070	0.009	0.002	0.004	0.007	0.026	0.028
OF Type	1.0	0.000	0.015	0.032	0.002	0.005	0.011	0.012	0.013
70% IT	0.9	−0.105	0.023	0.029	−0.004	0.001	0.011	0.012	0.014
	0.7	−0.357	0.076	0.017	−0.007	−0.001	0.012	0.021	0.026
	0.5	−0.693	0.167	0.013	0.003	0.006	0.012	0.021	0.024
P Type	1.0	0.000	0.015	0.018	0.009	0.010	0.011	0.015	0.015
35% IT	0.9	−0.105	0.020	0.015	0.008	0.009	0.011	0.019	0.020
	0.7	−0.357	0.037	0.010	0.004	0.005	0.008	0.031	0.031
	0.5	−0.693	0.070	0.006	0.004	0.004	0.006	0.030	0.031
P Type	1.0	0.000	0.014	0.016	0.003	0.005	0.008	0.012	0.013
50% IT	0.9	−0.105	0.026	0.020	0.005	0.008	0.013	0.021	0.022
	0.7	−0.357	0.060	0.014	0.001	0.004	0.009	0.031	0.032
	0.5	−0.693	0.119	0.009	0.006	0.007	0.009	0.027	0.028
P Type	1.0	0.000	0.021	0.038	0.001	0.005	0.013	0.013	0.014
70% IT	0.9	−0.105	0.032	0.019	−0.007	−0.002	0.010	0.014	0.017
	0.7	−0.357	0.095	0.021	−0.008	−0.002	0.009	0.019	0.024
	0.5	−0.693	0.194	0.012	0.007	0.010	0.015	0.023	0.026

OF Type: O'Brien–Fleming-type boundary, P Type: Pocock-type boundary, IT: Information time, HR: Hazard ratio.

negative conditional bias in Scenario 1. On the contrary, as shown in Table 2.3, WE1 and WE2 had a positive conditional bias because the MLE had a very large positive conditional bias in Scenario 2.

Table 2.5 shows the coverage probability of the estimators when the trial did not terminate early for efficacy (Scenario 2). The coverage probability of the MLE decreased

Table 2.4: Conditional mean-squared error at stage 2 for efficacy ($\lambda = 1$, Scenario 2)

Scenario	HR	logHR	MLE	S2E	CMAE	CMUE	CUMVUE	WE1	WE2
OF Type	1.0	0.000	0.022	0.034	0.022	0.022	0.022	0.024	0.024
35% IT	0.9	−0.105	0.018	0.029	0.018	0.018	0.018	0.019	0.019
	0.7	−0.357	0.010	0.017	0.012	0.012	0.011	0.011	0.011
	0.5	−0.693	0.006	0.009	0.007	0.007	0.007	0.007	0.007
OF Type	1.0	0.000	0.021	0.044	0.022	0.022	0.022	0.023	0.023
50% IT	0.9	−0.105	0.017	0.035	0.018	0.018	0.018	0.018	0.018
	0.7	−0.357	0.010	0.021	0.013	0.013	0.013	0.012	0.012
	0.5	−0.693	0.010	0.011	0.008	0.008	0.009	0.010	0.010
OF Type	1.0	0.000	0.020	0.076	0.024	0.023	0.022	0.023	0.022
70% IT	0.9	−0.105	0.015	0.061	0.021	0.020	0.019	0.019	0.018
	0.7	−0.357	0.013	0.037	0.017	0.018	0.018	0.018	0.018
	0.5	−0.693	0.031	0.020	0.013	0.013	0.014	0.015	0.016
P Type	1.0	0.000	0.019	0.030	0.020	0.020	0.020	0.020	0.020
35% IT	0.9	−0.105	0.016	0.026	0.018	0.018	0.017	0.017	0.017
	0.7	−0.357	0.010	0.015	0.011	0.011	0.011	0.011	0.011
	0.5	−0.693	0.009	0.008	0.007	0.007	0.007	0.009	0.009
P Type	1.0	0.000	0.018	0.041	0.021	0.020	0.020	0.019	0.019
50% IT	0.9	−0.105	0.014	0.032	0.018	0.018	0.017	0.016	0.016
	0.7	−0.357	0.011	0.019	0.013	0.013	0.013	0.013	0.013
	0.5	−0.693	0.018	0.010	0.008	0.008	0.009	0.010	0.011
P Type	1.0	0.000	0.018	0.074	0.023	0.023	0.021	0.021	0.021
70% IT	0.9	−0.105	0.014	0.056	0.021	0.021	0.020	0.019	0.019
	0.7	−0.357	0.015	0.035	0.017	0.018	0.019	0.018	0.019
	0.5	−0.693	0.041	0.018	0.012	0.013	0.014	0.015	0.015

OF Type: O'Brien–Fleming-type boundary, P Type: Pocock-type boundary, IT: Information time, HR: Hazard ratio.

when the true hazard ratio decreased, except in the case when the O'Brien–Fleming-type boundary was used and the information time was 35%. The conditional bias at stage 2 was small when the high O'Brien–Fleming-type boundary was used with 35% information time. The reason for the small conditional bias was that the large MLE at the interim analysis would stop the trial early and obtaining a large MLE at stage 2 was thus unlikely. Then, the small conditional bias raised the conditional coverage probability even if the true hazard ratio decreased. The S2E had a lower conditional coverage probability than that of the nominal level because of the lower amount of information used to calculate the S2E. The coverage probability did not differ between the CMAE, CMUE, and CUMVUE. However, the conditional coverage probability of these three estimators exceeded the nominal level of 95%, especially when the true hazard ratio was small. The conditional coverage probability of WE1 and WE2 did not reach the nominal level when the true hazard ratio was near 1.0. As shown in Table 2.6, the confidence interval of the S2E was wider than those of the other estimators when the true hazard ratio was near 1.0. The confidence interval of the CUMVUE was slightly wider than that of the CMAE and CMUE when the true hazard ratio was 0.5. However, the confidence interval of the CUMVUE was narrower than that of the CMAE and CMUE at a true hazard ratio ≥ 0.9. This is because the asymptotics of the bootstrap distribution worked well when the difference in the treatment effect was small.

Table 2.7 presents the evaluation results for overall bias for efficacy (Scenarios 1 and 2). The CUMVUE and S2E are not included in this table because they are only defined in the case of no early termination. The overall bias for the compared estimators depended on the true hazard ratio. WE2 was better at a true hazard ratio was 0.5 and information time of 70%.

The simulation results in the case of the interim analysis for futility (Scenarios 3 and

Table 2.5: Conditional coverage probability at stage 2 for efficacy ($\lambda = 1$, Scenario 2)

Scenario	HR	logHR	MLE	S2E	CMAE	CMUE	CUMVUE	WE1	WE2
OF Type	1.0	0.000	0.940	0.946	0.940	0.940	0.940	0.890	0.890
35% IT	0.9	−0.105	0.944	0.936	0.945	0.945	0.944	0.893	0.892
	0.7	−0.357	0.950	0.943	0.952	0.951	0.951	0.906	0.905
	0.5	−0.693	0.956	0.944	0.969	0.970	0.971	0.959	0.958
OF Type	1.0	0.000	0.941	0.941	0.942	0.942	0.941	0.889	0.889
50% IT	0.9	−0.105	0.950	0.943	0.952	0.952	0.951	0.902	0.901
	0.7	−0.357	0.956	0.945	0.964	0.965	0.964	0.942	0.940
	0.5	−0.693	0.903	0.943	0.980	0.979	0.978	0.975	0.975
OF Type	1.0	0.000	0.956	0.945	0.956	0.957	0.956	0.901	0.899
70% IT	0.9	−0.105	0.967	0.946	0.969	0.968	0.967	0.933	0.933
	0.7	−0.357	0.946	0.944	0.989	0.988	0.986	0.985	0.983
	0.5	−0.693	0.531	0.947	0.998	0.998	0.998	0.998	0.998
P Type	1.0	0.000	0.953	0.944	0.954	0.954	0.954	0.905	0.905
35% IT	0.9	−0.105	0.950	0.942	0.954	0.954	0.952	0.907	0.907
	0.7	−0.357	0.961	0.950	0.971	0.971	0.969	0.958	0.957
	0.5	−0.693	0.896	0.943	0.976	0.977	0.977	0.972	0.972
P Type	1.0	0.000	0.957	0.935	0.959	0.959	0.958	0.910	0.907
50% IT	0.9	−0.105	0.958	0.946	0.963	0.962	0.962	0.927	0.925
	0.7	−0.357	0.946	0.950	0.979	0.978	0.977	0.971	0.970
	0.5	−0.693	0.741	0.946	0.989	0.989	0.989	0.989	0.988
P Type	1.0	0.000	0.959	0.941	0.961	0.960	0.959	0.917	0.916
70% IT	0.9	−0.105	0.966	0.946	0.972	0.970	0.968	0.945	0.941
	0.7	−0.357	0.918	0.946	0.990	0.991	0.989	0.988	0.987
	0.5	−0.693	0.279	0.942	0.998	0.998	0.997	0.997	0.997

OF Type: O'Brien–Fleming-type boundary, P Type: Pocock-type boundary, IT: Information time, HR: Hazard ratio.

Table 2.6: Width of the confidence interval at stage 2 for efficacy ($\lambda = 1$, Scenario 2)

Scenario	HR	logHR	MLE	S2E	CMAE	CMUE	CUMVUE	WE1	WE2
OF Type	1.0	0.000	0.576	0.728	0.582	0.581	0.580	0.489	0.488
35% IT	0.9	−0.105	0.519	0.656	0.530	0.529	0.527	0.445	0.444
	0.7	−0.357	0.407	0.509	0.435	0.435	0.433	0.368	0.367
	0.5	−0.693	0.307	0.367	0.355	0.356	0.357	0.321	0.320
OF Type	1.0	0.000	0.575	0.820	0.597	0.595	0.591	0.496	0.494
50% IT	0.9	−0.105	0.518	0.736	0.555	0.553	0.549	0.464	0.462
	0.7	−0.357	0.415	0.571	0.488	0.488	0.488	0.425	0.425
	0.5	−0.693	0.329	0.411	0.425	0.427	0.431	0.404	0.404
OF Type	1.0	0.000	0.572	1.132	0.665	0.660	0.650	0.569	0.563
70% IT	0.9	−0.105	0.520	1.018	0.659	0.656	0.648	0.578	0.574
	0.7	−0.357	0.438	0.780	0.660	0.662	0.665	0.606	0.607
	0.5	−0.693	0.378	0.563	0.629	0.635	0.647	0.609	0.611
P Type	1.0	0.000	0.548	0.689	0.574	0.573	0.571	0.482	0.481
35% IT	0.9	−0.105	0.496	0.619	0.533	0.532	0.531	0.451	0.450
	0.7	−0.357	0.398	0.481	0.457	0.457	0.458	0.406	0.406
	0.5	−0.693	0.313	0.348	0.374	0.375	0.378	0.376	0.376
P Type	1.0	0.000	0.544	0.786	0.597	0.595	0.591	0.503	0.501
50% IT	0.9	−0.105	0.496	0.712	0.571	0.569	0.567	0.488	0.486
	0.7	−0.357	0.408	0.553	0.524	0.525	0.528	0.476	0.476
	0.5	−0.693	0.336	0.400	0.457	0.460	0.465	0.462	0.463
P Type	1.0	0.000	0.548	1.078	0.665	0.661	0.651	0.576	0.571
70% IT	0.9	−0.105	0.500	0.952	0.668	0.666	0.661	0.594	0.592
	0.7	−0.357	0.427	0.743	0.672	0.675	0.680	0.624	0.625
	0.5	−0.693	0.375	0.534	0.637	0.644	0.657	0.627	0.629

OF Type: O'Brien–Fleming-type boundary, P Type: Pocock-type boundary, IT: Information time, HR: Hazard ratio.

Table 2.7: Overall bias at for efficacy

Scenario	HR	logHR	PC	MLE	CMAE	CMUE	WE1	WE2
OF Type 35% IT	1.0	0.000	0.999	0.011	0.010	0.010	0.012	0.012
	0.9	−0.105	0.996	0.009	0.008	0.009	0.009	0.010
	0.7	−0.357	0.953	0.000	−0.005	0.001	0.000	0.006
	0.5	−0.693	0.625	−0.022	−0.024	−0.005	−0.023	−0.004
OF Type 50% IT	1.0	0.000	0.993	0.007	0.004	0.006	0.007	0.009
	0.9	−0.105	0.976	0.004	−0.003	0.002	0.003	0.007
	0.7	−0.357	0.789	−0.010	−0.023	−0.009	−0.016	0.000
	0.5	−0.693	0.210	−0.008	0.001	0.011	−0.015	−0.004
OF Type 70% IT	1.0	0.000	0.980	0.009	−0.002	0.004	0.006	0.010
	0.9	−0.105	0.930	0.006	−0.015	−0.004	−0.003	0.007
	0.7	−0.357	0.488	−0.007	−0.023	−0.015	−0.029	−0.017
	0.5	−0.693	0.024	0.005	0.014	0.014	0.004	0.004
P Type 35% IT	1.0	0.000	0.975	0.006	0.002	0.007	0.007	0.012
	0.9	−0.105	0.938	−0.003	−0.009	0.000	−0.003	0.007
	0.7	−0.357	0.678	−0.026	−0.027	−0.016	−0.025	−0.010
	0.5	−0.693	0.166	−0.014	0.003	0.005	−0.016	−0.012
P Type 50% IT	1.0	0.000	0.967	0.008	0.000	0.006	0.007	0.013
	0.9	−0.105	0.907	−0.004	−0.017	−0.006	−0.007	0.004
	0.7	−0.357	0.504	−0.019	−0.022	−0.015	−0.028	−0.017
	0.5	−0.693	0.041	0.002	0.015	0.013	0.001	0.000
P Type 70% IT	1.0	0.000	0.962	0.006	−0.011	−0.003	−0.001	0.007
	0.9	−0.105	0.863	−0.001	−0.026	−0.013	−0.014	0.000
	0.7	−0.357	0.329	−0.005	−0.010	−0.009	−0.023	−0.019
	0.5	−0.693	0.006	0.005	0.011	0.009	0.005	0.004

OF Type: O'Brien–Fleming-type boundary, P Type: Pocock-type boundary, IT: Information time, HR: Hazard ratio.

4) when $\lambda = 1$ are provided below. According to Table 2.8, the MLE was positively biased, which is contrary to that shown in Table 2.1. This is because the expectation of a truncated normal distribution in Scenario 3 has negative bias, as described in Equation (ES 6). The CMAE reduced the conditional bias of the MLE in all of the settings used. On the contrary, the CMUE tended to markedly overcorrect the conditional bias of the MLE. The maximum difference of the conditional bias between the CMUE and CMAE was approximately 0.65 when the information time was 35% and a Pocock-type boundary was used. The conditional coverage probability of the CMAE exceeded the nominal confidence level of 95% except for a true hazard ratio of 0.5 (Table 2.9). On the contrary, the conditional coverage probability of the CMUE did not reach the nominal confidence level. As shown in Table 2.10, the MLE was negatively biased, which is contrary to that shown in Table 2.3. The CUMVUE and S2E were unbiased.

The results for λ values of 0.5 and 2 are shown below. The results for all the estimators except for the S2E were similar to the case that λ = 1. Tables 2.11 to 2.14 show that the S2E was not unbiased if λ and the true hazard ratio $\neq 1$. In the cases that $\lambda \neq 1$ and the true hazard ratio = 1.0, the hazard functions for each patient were equal between the two groups. However, if λ and the true hazard ratio $\neq 1.0$, the hazard for each group changed over time and the deduction of the survival period to calculate the S2E could not thus reflect the true hazard. These results indicate that the S2E was valid only for the exponential distribution, confirming the advantage of using conditional bias-adjusted estimators.

In Scenario 1, the CMUE reduced the conditional bias when the stopping probability was small. However, in several of the cases tested, conditional coverage probability of the CMUE was low. The CMAE reduced the conditional bias of the MLE and had more than the 95% conditional coverage probability. In Scenario 2, the CUMVUE was

Table 2.8: Conditional bias at stage 1 for futility ($\lambda = 1$, Scenario 3)

Scenario	HR	logHR	MLE	CMAE	CMUE	WE1	WE2
OF Type	1.0	0.000	0.299	0.171	0.317	0.268	0.403
35% IT	0.9	−0.105	0.361	0.223	0.386	0.325	0.478
	0.7	−0.357	0.508	0.356	0.518	0.463	0.621
	0.5	−0.693	0.668	0.503	0.659	0.615	0.773
OF Type	1.0	0.000	0.139	0.064	−0.450	0.123	−0.414
50% IT	0.9	−0.105	0.194	0.107	−0.354	0.173	−0.318
	0.7	−0.357	0.328	0.223	−0.161	0.299	−0.126
	0.5	−0.693	0.493	0.376	0.032	0.456	0.066
OF Type	1.0	0.000	0.059	0.017	−0.218	0.052	−0.162
70% IT	0.9	−0.105	0.100	0.045	−0.143	0.089	−0.089
	0.7	−0.357	0.225	0.148	0.010	0.204	0.061
	0.5	−0.693	0.389	0.299	0.142	0.360	0.186
P Type	1.0	0.000	0.185	0.095	−0.556	0.166	−0.527
35% IT	0.9	−0.105	0.231	0.128	−0.452	0.207	−0.422
	0.7	−0.357	0.363	0.242	−0.255	0.329	−0.226
	0.5	−0.693	0.518	0.382	−0.057	0.475	−0.028
P Type	1.0	0.000	0.096	0.038	−0.291	0.085	−0.228
50% IT	0.9	−0.105	0.134	0.062	−0.209	0.118	−0.148
	0.7	−0.357	0.269	0.179	−0.028	0.245	0.030
	0.5	−0.693	0.427	0.323	0.133	0.394	0.184
P Type	1.0	0.000	0.045	0.009	−0.192	0.040	−0.142
70% IT	0.9	−0.105	0.083	0.034	−0.117	0.073	−0.067
	0.7	−0.357	0.201	0.129	0.011	0.182	0.059
	0.5	−0.693	0.364	0.280	0.141	0.338	0.183

OF Type: O'Brien–Fleming-type boundary, P Type: Pocock-type boundary, IT: Information time, HR: Hazard ratio.

Table 2.9: Conditional coverage probability at stage 1 for futility ($\lambda = 1$, Scenario 3)

Scenario	HR	logHR	MLE	CMAE	CMUE	WE1	WE2
OF Type	1.0	0.000	0.926	0.988	0.000	0.969	0.000
35% IT	0.9	−0.105	0.877	0.985	0.000	0.964	0.000
	0.7	−0.357	0.224	0.976	0.000	0.939	0.000
	0.5	−0.693	0.000	0.885	0.000	0.734	0.000
OF Type	1.0	0.000	0.957	0.988	0.019	0.972	0.035
50% IT	0.9	−0.105	0.933	0.986	0.051	0.967	0.105
	0.7	−0.357	0.638	0.981	0.951	0.956	1.000
	0.5	−0.693	0.000	0.892	0.999	0.765	0.999
OF Type	1.0	0.000	0.971	0.982	0.000	0.973	0.000
70% IT	0.9	−0.105	0.952	0.981	1.000	0.962	1.000
	0.7	−0.357	0.802	0.986	1.000	0.963	1.000
	0.5	−0.693	0.000	0.898	0.996	0.757	0.978
P Type	1.0	0.000	0.955	0.983	0.000	0.967	0.000
35% IT	0.9	−0.105	0.933	0.985	0.000	0.966	0.000
	0.7	−0.357	0.731	0.983	0.002	0.962	0.049
	0.5	−0.693	0.000	0.941	0.993	0.864	0.996
P Type	1.0	0.000	0.967	0.985	0.000	0.972	0.000
50% IT	0.9	−0.105	0.956	0.987	0.000	0.971	1.000
	0.7	−0.357	0.786	0.981	1.000	0.953	1.000
	0.5	−0.693	0.000	0.924	1.000	0.819	0.994
P Type	1.0	0.000	0.972	0.981	0.000	0.973	0.000
70% IT	0.9	−0.105	0.962	0.984	1.000	0.970	1.000
	0.7	−0.357	0.838	0.984	1.000	0.962	1.000
	0.5	−0.693	0.000	0.907	0.990	0.785	0.957

OF Type: O'Brien–Fleming-type boundary, P Type: Pocock-type boundary, IT: Information time, HR: Hazard ratio.

Table 2.10: Conditional bias at stage 2 for futility ($\lambda = 1$, Scenario 4)

Scenario	HR	logHR	MLE	S2E	CMAE	CMUE	CUMVUE	WE1	WE2
OF Type	1.0	0.000	−0.040	0.016	0.018	0.016	0.012	−0.026	−0.027
35% IT	0.9	−0.105	−0.020	0.016	0.016	0.014	0.010	−0.014	−0.015
	0.7	−0.357	0.002	0.011	0.010	0.009	0.007	0.003	0.003
	0.5	−0.693	0.004	0.007	0.004	0.004	0.004	0.004	0.004
OF Type	1.0	0.000	−0.089	0.023	0.028	0.024	0.018	−0.019	−0.022
50% IT	0.9	−0.105	−0.053	0.019	0.022	0.019	0.012	−0.017	−0.019
	0.7	−0.357	−0.003	0.017	0.016	0.014	0.009	0.002	0.001
	0.5	−0.693	0.005	0.011	0.006	0.006	0.006	0.005	0.005
OF Type	1.0	0.000	−0.159	0.033	0.044	0.037	0.023	0.006	−0.001
70% IT	0.9	−0.105	−0.098	0.033	0.043	0.036	0.021	0.007	0.000
	0.7	−0.357	−0.015	0.023	0.024	0.019	0.009	0.005	0.002
	0.5	−0.693	0.004	0.017	0.006	0.005	0.005	0.005	0.005
P Type	1.0	0.000	−0.068	0.016	0.017	0.015	0.012	−0.034	−0.035
35% IT	0.9	−0.105	−0.044	0.011	0.011	0.009	0.006	−0.029	−0.030
	0.7	−0.357	−0.004	0.012	0.012	0.010	0.008	−0.002	−0.003
	0.5	−0.693	0.004	0.008	0.005	0.005	0.005	0.004	0.004
P Type	1.0	0.000	−0.111	0.021	0.025	0.023	0.017	−0.020	−0.022
50% IT	0.9	−0.105	−0.069	0.018	0.022	0.019	0.013	−0.019	−0.021
	0.7	−0.357	−0.010	0.014	0.016	0.014	0.009	−0.001	−0.002
	0.5	−0.693	0.006	0.012	0.008	0.007	0.007	0.006	0.006
P Type	1.0	0.000	−0.171	0.034	0.042	0.035	0.022	0.005	−0.001
70% IT	0.9	−0.105	−0.108	0.033	0.043	0.036	0.021	0.007	0.001
	0.7	−0.357	−0.017	0.022	0.027	0.022	0.011	0.007	0.004
	0.5	−0.693	0.005	0.015	0.007	0.007	0.006	0.006	0.006

OF Type: O'Brien–Fleming-type boundary, P Type: Pocock-type boundary, IT: Information time, HR: Hazard ratio.

Table 2.11: Conditional bias at stage 2 for efficacy ($\lambda = 0.5$, Scenario 2)

Scenario	HR	logHR	MLE	S2E	CMAE	CMUE	CUMVUE	WE1	WE2
OF Type	1.0	0.000	0.008	0.015	0.007	0.008	0.008	0.009	0.009
35% IT	0.9	−0.105	0.014	0.047	0.013	0.013	0.014	0.013	0.013
	0.7	−0.357	0.016	0.069	0.008	0.009	0.011	0.016	0.016
	0.5	−0.693	0.032	0.063	0.004	0.005	0.006	0.025	0.025
OF Type	1.0	0.000	0.019	0.029	0.016	0.016	0.018	0.018	0.019
50% IT	0.9	−0.105	0.012	0.053	0.004	0.005	0.008	0.009	0.010
	0.7	−0.357	0.030	0.059	−0.001	0.002	0.007	0.020	0.021
	0.5	−0.693	0.082	0.056	0.008	0.009	0.012	0.032	0.034
OF Type	1.0	0.000	0.017	0.037	0.004	0.007	0.013	0.012	0.014
70% IT	0.9	−0.105	0.027	0.051	0.000	0.005	0.014	0.016	0.018
	0.7	−0.357	0.078	0.055	0.000	0.006	0.019	0.028	0.033
	0.5	−0.693	0.166	0.042	0.005	0.008	0.014	0.023	0.027
P Type	1.0	0.000	0.014	0.018	0.007	0.008	0.010	0.013	0.013
35% IT	0.9	−0.105	0.022	0.035	0.010	0.011	0.014	0.021	0.021
	0.7	−0.357	0.039	0.063	0.007	0.009	0.011	0.034	0.035
	0.5	−0.693	0.067	0.055	0.002	0.002	0.004	0.029	0.029
P Type	1.0	0.000	0.012	0.013	0.000	0.002	0.006	0.008	0.009
50% IT	0.9	−0.105	0.026	0.035	0.004	0.007	0.012	0.020	0.021
	0.7	−0.357	0.058	0.055	−0.001	0.002	0.007	0.029	0.031
	0.5	−0.693	0.118	0.047	0.006	0.007	0.009	0.027	0.028
P Type	1.0	0.000	0.025	0.029	0.005	0.009	0.016	0.018	0.020
70% IT	0.9	−0.105	0.037	0.049	−0.001	0.004	0.016	0.019	0.022
	0.7	−0.357	0.095	0.054	−0.007	−0.001	0.010	0.020	0.025
	0.5	−0.693	0.194	0.043	0.010	0.013	0.018	0.026	0.029

OF Type: O'Brien–Fleming-type boundary, P Type: Pocock-type boundary, IT: Information time, HR: Hazard ratio.

Table 2.12: Conditional bias at stage 2 for futility ($\lambda = 0.5$, Scenario 4)

Scenario	HR	logHR	MLE	S2E	CMAE	CMUE	CUMVUE	WE1	WE2
OF Type	1.0	0.000	−0.037	0.030	0.021	0.019	0.015	−0.023	−0.024
35% IT	0.9	−0.105	−0.025	0.039	0.008	0.006	0.003	−0.020	−0.020
	0.7	−0.357	0.001	0.067	0.008	0.007	0.005	0.001	0.000
	0.5	−0.693	0.007	0.065	0.007	0.007	0.007	0.007	0.007
OF Type	1.0	0.000	−0.089	0.042	0.025	0.021	0.015	−0.023	−0.026
50% IT	0.9	−0.105	−0.050	0.049	0.025	0.021	0.015	−0.014	−0.016
	0.7	−0.357	−0.010	0.057	0.008	0.006	0.002	−0.004	−0.005
	0.5	−0.693	0.010	0.052	0.011	0.011	0.010	0.011	0.011
OF Type	1.0	0.000	−0.160	0.055	0.036	0.029	0.015	−0.003	−0.010
70% IT	0.9	−0.105	−0.099	0.079	0.039	0.031	0.017	0.002	−0.004
	0.7	−0.357	−0.019	0.064	0.018	0.014	0.004	0.000	−0.003
	0.5	−0.693	0.005	0.045	0.007	0.006	0.005	0.005	0.005
P Type	1.0	0.000	−0.063	0.037	0.022	0.021	0.018	−0.029	−0.030
35% IT	0.9	−0.105	−0.040	0.044	0.017	0.015	0.012	−0.023	−0.024
	0.7	−0.357	−0.010	0.065	0.005	0.004	0.001	−0.010	−0.010
	0.5	−0.693	0.001	0.057	0.002	0.002	0.001	0.001	0.001
P Type	1.0	0.000	−0.109	0.045	0.026	0.023	0.018	−0.019	−0.022
50% IT	0.9	−0.105	−0.069	0.055	0.022	0.019	0.013	−0.019	−0.021
	0.7	−0.357	−0.008	0.072	0.018	0.016	0.011	0.000	−0.001
	0.5	−0.693	0.002	0.055	0.004	0.003	0.003	0.002	0.002
P Type	1.0	0.000	−0.173	0.056	0.034	0.028	0.015	−0.002	−0.009
70% IT	0.9	−0.105	−0.108	0.062	0.038	0.032	0.018	0.003	−0.003
	0.7	−0.357	−0.021	0.060	0.020	0.015	0.005	0.001	−0.002
	0.5	−0.693	0.006	0.037	0.008	0.007	0.006	0.007	0.006

OF Type: O’Brien–Fleming-type boundary, P Type: Pocock-type boundary, IT: Information time, HR: Hazard ratio.

Table 2.13: Conditional bias at stage 2 for efficacy ($\lambda = 2$, Scenario 2)

Scenario	HR	logHR	MLE	S2E	CMAE	CMUE	CUMVUE	WE1	WE2
OF Type	1.0	0.000	0.007	0.022	0.006	0.007	0.007	0.006	0.006
35% IT	0.9	−0.105	0.011	0.093	0.009	0.010	0.010	0.011	0.011
	0.7	−0.357	0.012	0.248	0.005	0.006	0.008	0.012	0.012
	0.5	−0.693	0.030	0.353	0.006	0.006	0.008	0.025	0.026
OF Type	1.0	0.000	0.014	0.019	0.011	0.011	0.013	0.014	0.014
50% IT	0.9	−0.105	0.010	0.099	0.003	0.004	0.007	0.010	0.010
	0.7	−0.357	0.027	0.240	−0.001	0.002	0.006	0.019	0.020
	0.5	−0.693	0.076	0.320	0.008	0.010	0.013	0.034	0.036
OF Type	1.0	0.000	0.017	0.027	0.005	0.008	0.013	0.014	0.015
70% IT	0.9	−0.105	0.021	0.108	−0.004	0.000	0.009	0.012	0.014
	0.7	−0.357	0.072	0.216	−0.002	0.003	0.015	0.025	0.030
	0.5	−0.693	0.162	0.257	0.016	0.019	0.025	0.036	0.040
P Type	1.0	0.000	0.012	0.012	0.005	0.006	0.008	0.011	0.011
35% IT	0.9	−0.105	0.019	0.091	0.007	0.008	0.010	0.017	0.017
	0.7	−0.357	0.035	0.224	0.003	0.005	0.007	0.030	0.030
	0.5	−0.693	0.071	0.313	0.009	0.010	0.011	0.037	0.038
P Type	1.0	0.000	0.014	0.015	0.003	0.005	0.008	0.011	0.012
50% IT	0.9	−0.105	0.023	0.090	0.002	0.004	0.009	0.018	0.019
	0.7	−0.357	0.059	0.216	0.004	0.006	0.011	0.033	0.035
	0.5	−0.693	0.116	0.280	0.010	0.011	0.013	0.032	0.033
P Type	1.0	0.000	0.017	0.022	−0.002	0.001	0.009	0.011	0.013
70% IT	0.9	−0.105	0.030	0.098	−0.008	−0.002	0.009	0.013	0.016
	0.7	−0.357	0.097	0.199	0.002	0.008	0.019	0.031	0.035
	0.5	−0.693	0.192	0.235	0.017	0.020	0.025	0.034	0.037

OF Type: O'Brien–Fleming-type boundary, P Type: Pocock-type boundary, IT: Information time, HR: Hazard ratio.

Table 2.14: Conditional bias at stage 2 for futility ($\lambda = 2$, Scenario 4)

Scenario	HR	logHR	MLE	S2E	CMAE	CMUE	CUMVUE	WE1	WE2
OF Type	1.0	0.000	−0.037	0.042	0.017	0.016	0.012	−0.025	−0.026
35% IT	0.9	−0.105	−0.017	0.113	0.017	0.015	0.012	−0.012	−0.013
	0.7	−0.357	0.001	0.257	0.009	0.008	0.006	0.002	0.002
	0.5	−0.693	0.002	0.375	0.003	0.003	0.003	0.003	0.003
OF Type	1.0	0.000	−0.088	0.078	0.020	0.017	0.010	−0.028	−0.030
50% IT	0.9	−0.105	−0.054	0.135	0.015	0.012	0.005	−0.024	−0.026
	0.7	−0.357	−0.006	0.264	0.010	0.008	0.005	−0.002	−0.003
	0.5	−0.693	0.003	0.392	0.003	0.003	0.003	0.003	0.003
OF Type	1.0	0.000	−0.156	0.138	0.029	0.022	0.009	−0.011	−0.017
70% IT	0.9	−0.105	−0.096	0.191	0.034	0.027	0.013	−0.003	−0.009
	0.7	−0.357	−0.014	0.283	0.021	0.017	0.008	0.003	0.001
	0.5	−0.693	0.005	0.397	0.006	0.006	0.005	0.006	0.005
P Type	1.0	0.000	−0.067	0.061	0.014	0.012	0.010	−0.036	−0.038
35% IT	0.9	−0.105	−0.039	0.129	0.015	0.014	0.011	−0.025	−0.026
	0.7	−0.357	−0.007	0.257	0.008	0.007	0.005	−0.006	−0.006
	0.5	−0.693	0.004	0.373	0.005	0.005	0.004	0.005	0.005
P Type	1.0	0.000	−0.109	0.091	0.020	0.018	0.013	−0.026	−0.028
50% IT	0.9	−0.105	−0.070	0.152	0.016	0.013	0.008	−0.024	−0.026
	0.7	−0.357	−0.012	0.270	0.011	0.009	0.005	−0.006	−0.007
	0.5	−0.693	0.004	0.387	0.006	0.005	0.005	0.005	0.005
P Type	1.0	0.000	−0.169	0.151	0.028	0.022	0.010	−0.010	−0.016
70% IT	0.9	−0.105	−0.109	0.193	0.025	0.019	0.006	−0.011	−0.017
	0.7	−0.357	−0.020	0.285	0.020	0.016	0.006	0.001	−0.002
	0.5	−0.693	0.003	0.395	0.005	0.004	0.003	0.004	0.003

OF Type: O'Brien–Fleming-type boundary, P Type: Pocock-type boundary, IT: Information time, HR: Hazard ratio.

unbiased regardless of the shape of the Weibull distribution. In addition, the conditional coverage probability of the CUMVUE was higher than or equal to that of the S2E. In Scenario 3, the CMAE reduced the conditional bias of the MLE regardless of the setting used. The CMAE also improved the conditional coverage probability of the MLE. On the contrary, the CMUE tended to markedly overcorrect the conditional bias of the MLE. In Scenario 4, the CUMVUE was unbiased, similar to that in Scenario 2. The conditional coverage probability of the S2E was lower than 95% but that of the CUMVUE, although conservatively, was over 95%. Our simulation results suggest that the stage 2 estimator was not universally useful because 1) the estimator cannot be calculated when the study terminated early, 2) the magnitude of the conditional bias for the estimator was sometimes larger than that for the CUMVUE, and 3) the estimator was valid only when the hazard was constant over time.

2.6 Application

We applied the bias-adjusted estimators to published data from a clinical breast cancer trial. The CLEOPATRA study of pertuzumab in combination with trastuzumab in patients with HER2-positive metastatic breast cancer was a randomized, parallel-controlled clinical trial.[1,44] In total, 808 patients were randomized to receive pertuzumab plus trastuzumab and docetaxel (402 patients) or trastuzumab and docetaxel (406 patients). The study was planned to include one interim analysis for overall survival time. An interim analysis was conducted in this study to allow the possibility of early stopping for efficacy. The planned hazard ratio was 0.75, corresponding to a log hazard ratio of -0.288. The required sample size and death rate to detect 80% power were 800 and 385 patients, respectively. The efficacy analysis for overall survival was planned at the time of the final progression-free survival analysis. An O'Brien–Fleming-type alpha-spending function was used to control the cumulative significance level, 0.05. The accrual time was 2.5 years. The median duration of the follow-up at the interim and final data cutoff was 19.3 and 49.5 months, respectively. One interim analysis for the overall survival was conducted when 165 events were observed to correspond to 43% information time, and the estimated hazard ratio did not cross the stopping boundary of the alpha-spending function for the interim analysis of overall survival. The MLE for the hazard ratio calculated in the final analysis was 0.68, corresponding to a log hazard ratio of -0.386. The CMAE, CMUE, CUMVUE, WE1, and WE2 were applied to this study to estimate the log hazard ratio. The Fisher information at stage 2 was calculated approximately as $389/4$ by using the number of events at stage 2. The bootstrap conditional confidence intervals of each estimator could not be estimated here because the individual patient data from the study could not be accessed. Table 2.15 shows the hazard ratio for each estimator. All the hazard ratios estimated by the conditional bias-adjusted estimators were lower than those of the MLE. As expected, WE1 and WE2 were between the MLE and each of the conditional bias-adjusted estimators. This result indicates that the estimated hazard ratio from the study without termination tends to underestimate the difference in the treatment effect. Our simulations suggest that the CUMVUE is a minimum-biased estimator in the study without early termination. Therefore, in terms of bias reduction, we recommend adding the value of 0.646, which is the estimate of the hazard ratio from the CUMVUE for the adjusted value.

Table 2.15: Hazard ratio using the bias-adjusted estimators

MLE	CMAE	CMUE	CUMVUE	WE1	WE2
0.680	0.644	0.645	0.648	0.664	0.665

2.7 Summary

In this study, we evaluated the performance of the CMAE, CMUE, CUMVUE, and WE through simulation studies in scenarios where one interim analysis is conducted and the primary endpoint is survival outcome. The performance of the conditional bias-adjusted estimator depended on the scenarios studied. It is difficult to develop a "completely" conditional unbiased estimator because the true hazard ratio is unknown;[27,47] therefore, using the estimator with the smallest conditional bias and conditional mean-squared error regardless of the true hazard ratio is recommended. Our simulation results indicate that selecting the conditional bias-adjusted estimator depending on the scenarios is reasonable. If a trial terminates at the interim analysis, the use of the CMAE is recommended. Otherwise, the CUMVUE is the appropriate bias-adjusted estimator. In Scenario 1, the conditional bias of the CMAE was somewhat larger than that of the CMUE. However, in practice, this occurrence may be rare because the stopping probability for efficacy is relatively low in situations where the hazard ratio is near 1.0. The reason for the low conditional coverage probability of the CMUE as found in Scenario 1 was that the conditional mean-squared error for the CMUE was relatively large.In Scenario 3, the reason for the overcorrection of the CMUE might be that the stopping probability for futility was larger than that for efficacy. Further, the denominator of Equation (ES 8) in the case of Scenario 3 became smaller than that of Equation (3) in the case of Scenario 1.

Chapter 3

Extension of conditional estimation using prior information

3.1 Introduction

Many medical researchers have discussed the bias when the trial stops early for efficacy.[2–4, 28, 51] Chapter 2 compared the performance of existing bias-adjusted estimators in settings in which the trial does or does not stop for efficacy at the interim analysis. Hence, the use of the CMAE is recommended in the former case because the conditional bias and coverage probability are better than the others.

The existing conditional bias-adjusted estimators can reduce this bias sufficiently when the trial has continued. Therefore, we focus on those cases in which the trial was stopped early for efficacy or futility. In particular, we propose a new bias-adjusted estimator using prior information on the treatment effect before the start of the trial when the trial stops for efficacy or futility.

If the interim analysis with both efficacy and futility boundaries is planned, the maximum information time, Fisher information, variance, and stopping boundary in each stage differ from those in the trial with the interim analysis only for the efficacy boundary. Thus, we assumed three patterns: planned only for efficacy and stopped for efficacy, planned for both efficacy and futility and stopped for efficacy, and planned for both efficacy and futility and stopped for futility.

3.2 Remaining conditional bias of CMAE

Again, the CMAE is calculated by reducing the bias from $\hat{\theta}_{MLE,1}$ by using the following formula if the trial is terminated early for efficacy:

$$\hat{\theta}_{CMAE,1} = \hat{\theta}_{MLE,1} - B(\hat{\theta}_{MLE,1}|\sigma_1, a_1, M = 1). \tag{3.1}$$

We note that the parameter of interest is $\exp(\theta)$, although the bias correction is on the MLE of θ. As the true θ becomes zero, $B(\theta|\sigma_1, a_1, M = 1)$ becomes large. In the case that a relatively large treatment effect $\hat{\theta}_{MLE,1}$ is observed, $B(\hat{\theta}_{MLE,1}|\sigma_1, a_1, M = 1)$ will be large and there would be less bias correction. Hence, the remaining bias is nonnegligible.

3.3 Structure of the proposed estimator

As described in Section 1.4, if θ is given, the conditional bias can be obtained exactly. If $B(\theta|\sigma_1, a_1, M = 1)$ can be estimated without bias, the complete unbiased estimator can also be calculated. The CMAE uses the MLE as an alternative to θ; however, the remaining conditional bias is nonnegligible, especially when the trial is stopped early, even if the true treatment effect is not large. This comes from the underestimation of the magnitude of $B(\theta|\sigma_1, a_1, M = 1)$ because the MLE itself includes the bias. Therefore, improving the estimator imputed to $B(\theta|\sigma_1, a_1, M = 1)$ may help reduce the remaining conditional bias.

Thompson[46] developed a "shrunken estimator," which represents the MLE shrunk toward θ_0 by multiplying a shrinking parameter by the MLE if we believe the prior parameter θ_0 is close to the true value θ. Based on the shrunken estimator, we proposed a weighted estimator, defined as the weight average of the MLE and prior information, to reduce the bias of the estimate for $B(\theta|\sigma_1, a_1, M = 1)$ as follows:

$$\hat{\theta}_* = c\hat{\theta}_{MLE,1} + (1 - c)\theta_0, \tag{3.2}$$

where θ_0 is a prior information parameter and c $(0 \leq c \leq 1)$ is a weight parameter. The WCMAE uses $\hat{\theta}_*$ to estimate $B(\theta|\sigma_1, a_1, M = 1)$. That is, the WCMAE is defined as follows:

$$\hat{\theta}_{WCMAE,1} = \hat{\theta}_{MLE,1} - B(\hat{\theta}_*|\sigma_1, a_1, M = 1). \tag{3.3}$$

This is a modification of the CMAE that uses $\hat{\theta}_{MLE,1}$ to estimate $B(\theta|\sigma_1, a_1, M = 1)$. We obtain the CMAE as a special case where c is 1 in Equation (5).

As described in Section 3.1, we focus on a trial that stopped early for efficacy or futility. Therefore, the proposed estimator for efficacy (futility) stopping might be appropriate when the observed MLE tends to be larger (smaller) than the planned effect size. In this case, Equation (5) would become a conservative estimate and the use of the WCMAE would be expected to reduce the overestimation of the effect size by the MLE.

3.4 Specification of the parameters

3.4.1 Prior information parameter θ_0

We must set the prior information parameter θ_0 and weight parameter c to use Equation (6). In our research, θ_0 is set to the log hazard ratio used to calculate the sample size. The rationale of this approach is as follows:

- All confirmatory clinical trials with the interim analysis determine the sample size before the start of the trial.
- The methods for calculating the sample size are predefined in the protocols.[37] Although the rationale for calculating the sample size may be subjective, the use of a prespecified effect size is relatively objective because θ_0 cannot change in a statistical analysis.
- A randomized clinical trial is planned to detect the predetermined effect size under the alternative hypothesis. Therefore, it would be natural to set the effect size as the prior parameter θ_0 from the perspective of the hypothesis testing.
- The effect size used to determine the sample size is considered to be based on all the available prior information.

The methodology used to determine θ_0 is the same as general approaches of determining the sample size in randomized clinical trials. The results of a previous clinical trial for the same, similar, and rival drugs are considered to be the prior information. Statistically, a point estimate or lower 95% confidence limit of the effect size is used.[52] Pocock[35] combined historical control data with the sample size calculation. Moreover, a random effects model in a meta-analysis[43] or in a network meta-analysis could be applied when several phase III randomized trials may be available, taking into account their variability.[16]

3.4.2 Weight parameter c

For the parameter of c, Thompson[46] proposed c as a function of the MLE to minimize the mean-squared error of the shrunken estimator in Equation (5). The mean-squared error of the shrunken estimator is defined as

$$E\left[\left(\hat{\theta}_* - \theta\right)^2\right] = E\left[\left\{c(\hat{\theta}_{MLE,1} - \theta_0) - (\theta - \theta_0)\right\}^2\right],$$

and c is derived by minimizing the mean-squared error as

$$c = \frac{(\theta - \theta_0)^2}{(\theta - \theta_0)^2 + \sigma_1^2}.$$

Thompson[46] replaced the unknown true parameter θ in c by the MLE:

$$\hat{c} = \frac{(\hat{\theta}_{MLE,1} - \theta_0)^2}{(\hat{\theta}_{MLE,1} - \theta_0)^2 + \sigma_1^2}.$$

Therefore, $\hat{\theta}_*$ can be rewritten as

$$\hat{\theta}_* = \frac{(\hat{\theta}_{MLE,1} - \theta_0)^2}{(\hat{\theta}_{MLE,1} - \theta_0)^2 + \sigma_1^2}\hat{\theta}_{MLE,1} + \left\{1 - \frac{(\hat{\theta}_{MLE,1} - \theta_0)^2}{(\hat{\theta}_{MLE,1} - \theta_0)^2 + \sigma_1^2}\right\}\theta_0. \tag{3.4}$$

The prior θ_0 may be unreliable when the discrepancy between the MLE and θ_0 is large. Therefore, there is a risk of overcorrection by shrinking the MLE toward θ_0. From Equation (7), the larger $\hat{\theta}_{MLE,1} - \theta_0$ becomes, the larger the weight of the MLE is. From this point of view, the use of Equation (7) is expected to avoid the overcorrection of the MLE. On the contrary, when the MLE and θ_0 are close to each other, the setting of θ_0 is relatively reliable and the risk of overcorrection is not serious. In this case, the weight of θ_0 should be large.

3.5 Analytical bias of the proposed estimator

Based on Sections 1.4, 3.3, and 3.4, we can analytically evaluate the conditional bias of the MLE, CMAE, and WCMAE. The formula and derivation of the conditional biases are shown below.

3.5.1 The CMAE

For simplicity, $B(\theta|\sigma_1, a_1, M = 1)$ is expressed as $B(\theta)$. Using the delta method, the conditional bias of the CMAE as hazard ratio is

$$\begin{aligned}
E\left[\exp(\hat{\theta}_{CMAE,1}) - \exp(\theta)\right] &= E\left[\exp\{\hat{\theta}_{MLE,1} - B(\hat{\theta}_{MLE,1})\}\right] - \exp(\theta) \\
&\approx \exp\left\{E\left[\hat{\theta}_{MLE,1} - B(\hat{\theta}_{MLE,1})\right]\right\} - \exp(\theta) \\
&= \exp\left\{E\left[\hat{\theta}_{MLE,1}\right] - E\left[B(\hat{\theta}_{MLE,1})\right]\right\} - \exp(\theta) \\
&= \exp\left\{\theta + B(\theta) - E\left[B(\hat{\theta}_{MLE,1})\right]\right\} - \exp(\theta) \\
&\approx \exp\left\{\theta + B(\theta) - B\left(E\left[\hat{\theta}_{MLE,1}\right]\right)\right\} - \exp(\theta) \\
&= \exp\left\{\theta + B(\theta) - B\left(\theta + B(\theta)\right)\right\} - \exp(\theta).
\end{aligned}$$

3.5.2 The WCMAE

The conditional bias of the WCMAE as hazard ratio is

$$
\begin{aligned}
E\left[\exp(\hat{\theta}_{WCMAE,1}) - \exp(\theta)\right] &= E\left[\exp\{\hat{\theta}_{MLE,1} - B(\hat{\theta}_*)\}\right] - \exp(\theta) \\
&\approx \exp\left\{E\left[\hat{\theta}_{MLE,1} - B(\hat{\theta}_*)\right]\right\} - \exp(\theta) \\
&= \exp\left\{E\left[\hat{\theta}_{MLE,1}\right] - E\left[B(\hat{\theta}_*)\right]\right\} - \exp(\theta) \\
&= \exp\left\{\theta + B(\theta) - E\left[B(\hat{\theta}_*)\right]\right\} - \exp(\theta) \\
&\approx \exp\left\{\theta + B(\theta) - B\left(E\left[\hat{\theta}_*\right]\right)\right\} - \exp(\theta) \\
&= \exp\left\{\theta + B(\theta) - B\left(E\left[\hat{c}\hat{\theta}_{MLE,1} + (1-\hat{c})\theta_0\right]\right)\right\} - \exp(\theta) \\
&\approx \exp\left\{\theta + B(\theta) - B\left(c(\theta + B(\theta)) + (1-c)\theta_0\right)\right\} - \exp(\theta),
\end{aligned}
$$

where c is defined as below:

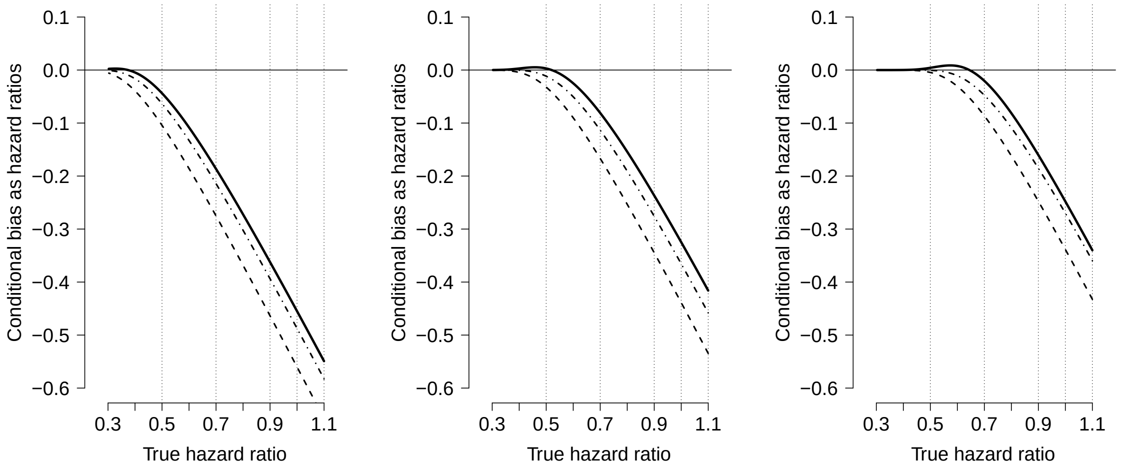
$$
c = \frac{(E\left[\hat{\theta}_{MLE,1}\right] - \theta_0)^2}{(E\left[\hat{\theta}_{MLE,1}\right] - \theta_0)^2 + \sigma_1^2} = \frac{(\theta + B(\theta) - \theta_0)^2}{(\theta + B(\theta) - \theta_0)^2 + \sigma_1^2}.
$$

In Figures 3.1 and 3.3, the conditional bias of the MLE, CMAE, and WCMAE are shown when the O'Brien–Fleming-type boundary for efficacy is used with the information time of 35%, 50%, and 70% and the trial stops for efficacy. Figures 3.2 and 3.4 show the conditional bias when the O'Brien–Fleming-type boundaries for both efficacy and futility are used and the trial stops for efficacy. Figures 3.5 and 3.6 show the conditional bias when the O'Brien–Fleming-type boundaries for both efficacy and futility are used and the trial stops for futility. The parameter of $\exp(\theta_0)$ was set to 0.7 (moderate effect size) and 0.5 (enthusiastic effect size), respectively.

Figures 3.1 and 3.2 show the conditional bias for efficacy when $\exp(\theta_0) = 0.7$

(Moderate Scenario) and the O'Brien–Fleming-type boundary is used. The difference in the treatment effect is overestimated (underestimated) if the bias is negative (positive). The MLE had the largest conditional bias. The maximum bias for the MLE was over -0.6 as hazard ratios when the information time was 35%. The conditional bias for the MLE converged to zero as the true hazard ratio approached 0.3. In addition, the bias decreased as the information time at the interim analysis increased from 35% to 70%. The CMAE reduced the conditional bias of the MLE. However, the bias reduction of the CMAE was insufficient because the MLE, used to estimate $B(\theta|\sigma_1, a_1, M = 1)$, tends to be less than the true value. The WCMAE was better than the CMAE when the true hazard ratio ≥ 0.7. In Figure 3.1 (c), the WCMAE was better than the CMAE when the true hazard ratio ≥ 0.65, whereas the absolute conditional bias of the WCMAE and CMAE was close to each other when the hazard ratio < 0.65. The WCMAE occasionally overcorrected the bias of the MLE by 0.035 as hazard ratios when the true hazard ratio < 0.65. However, the overcorrection was close to zero and the positive bias of the WCMAE became negligible when the true hazard ratio became smaller. As defined in Section 3.4.2, the WCMAE uses the shrunken estimator weighted by the MLE and θ_0 to estimate $B(\theta|\sigma_1, a_1, M = 1)$. In Figure 3.1, the overcorrection was restricted because the weight of the MLE relatively increased if the prior information was far from the true hazard ratio.

Figures 3.3 and 3.4 show the conditional bias for efficacy when $\exp(\theta_0) = 0.5$ (Enthusiastic Scenario). This scenario corresponds to the case where the planned effect size was estimated enthusiastically. The conditional bias became entirely larger than that in the Moderate Scenario since the required events, sample sizes, and Fisher information at the interim analysis decreased. In Figures 3.3 (a) and 3.3 (b), the WCMAE with $\exp(\theta_0)$ of 0.5 was the best of all the estimators regardless of the true hazard ra-

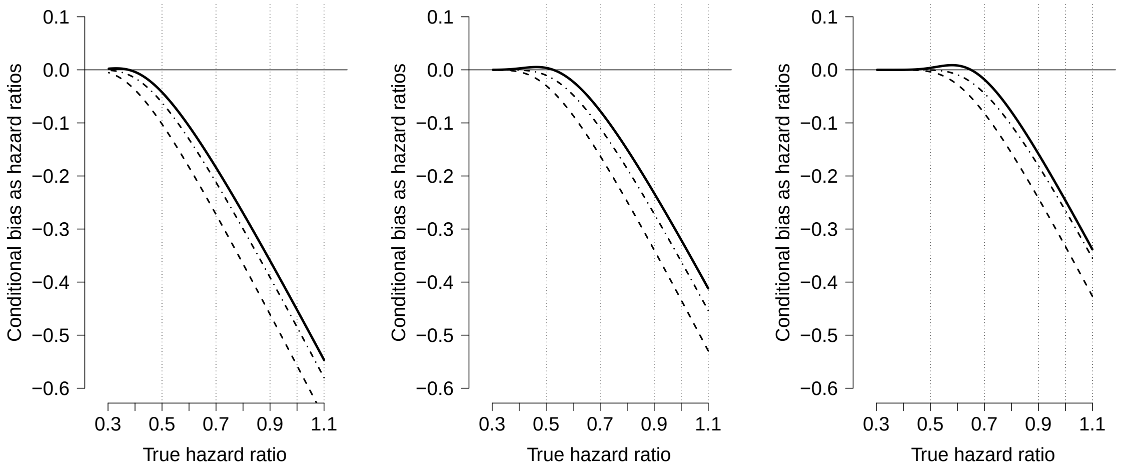


(a) 35% IT

(b) 50% IT

(c) 70% IT

Figure 3.1: Conditional bias of the MLE (dashed line), CMAE (dot-dashed line), and WCMAE (solid line) when the O'Brien–Fleming-type boundary was used with the information time of 35%, 50%, and 70% (Planned for efficacy, Stopped for efficacy). IT means information time. $\exp(\theta_0)$ was set to 0.7. The vertical reference lines correspond to the true hazard ratios of 0.5, 0.7, 0.9, 1.0, and 1.1, respectively.



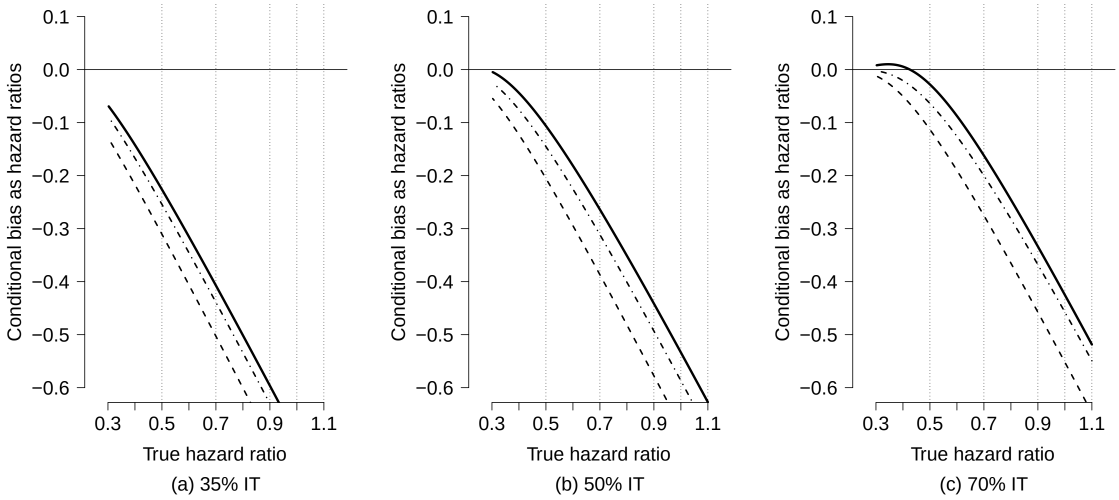
(a) 35% IT

(b) 50% IT

(c) 70% IT

Figure 3.2: Conditional bias of the MLE (dashed line), CMAE (dot-dashed line), and WCMAE (solid line) when the O'Brien–Fleming-type boundary was used with the information time of 35%, 50%, and 70% (Planned for efficacy and futility, Stopped for efficacy). IT means information time. $\exp(\theta_0)$ was set to 0.7. The vertical reference lines correspond to the true hazard ratios of 0.5, 0.7, 0.9, 1.0, and 1.1, respectively.

tio. The WCMAE overcorrected when the true hazard ratio was <0.4 in Figure 3.3 (c). Figures 3.1 (c) and 3.3 (c) suggest that the overcorrection occurred if the true hazard ratio was substantially smaller than $\exp(\theta_0)$. As defined in Section 3.4.1, θ_0 is set to the log hazard ratio used to calculate the sample size. If the prespecified effect size θ_0 was large, the conditional bias of the MLE tended to be large and the shrunken estimator was suitable.



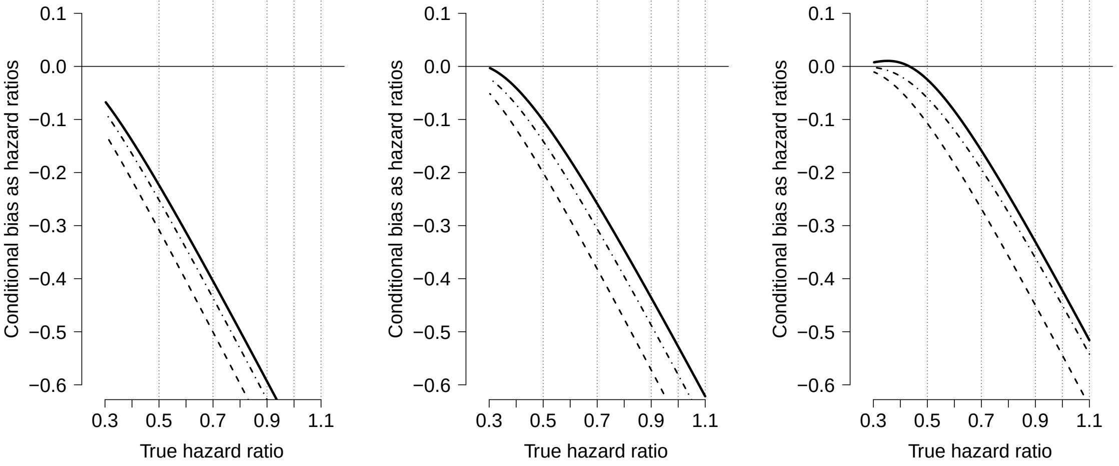
(a) 35% IT

(b) 50% IT

(c) 70% IT

Figure 3.3: Conditional bias of the MLE (dashed line), CMAE (dot-dashed line), and WCMAE (solid line) when the O'Brien–Fleming-type boundary was used with the information time of 35%, 50%, and 70% (Planned for efficacy, Stopped for efficacy). IT means information time. $\exp(\theta_0)$ was set to 0.5. The vertical reference lines correspond to the true hazard ratios of 0.5, 0.7, 0.9, 1.0, and 1.1, respectively.

In Figures 3.5 and 3.6, the conditional biases for futility when $\exp(\theta_0) = 0.7$ (Moderate Scenario) and $\exp(\theta_0) = 0.5$ (Enthusiastic Scenario) are presented, respectively. The MLE had the largest positive conditional bias. This result indicates that the use of the MLE tends to underestimate the treatment effect when the trial stopped for futility at the interim analysis. The maximum bias for the MLE was over 0.8 when $\exp(\theta_0) = 0.7$ and the information time was 35%. The conditional bias of the WCMAE was the small-



(a) 35% IT

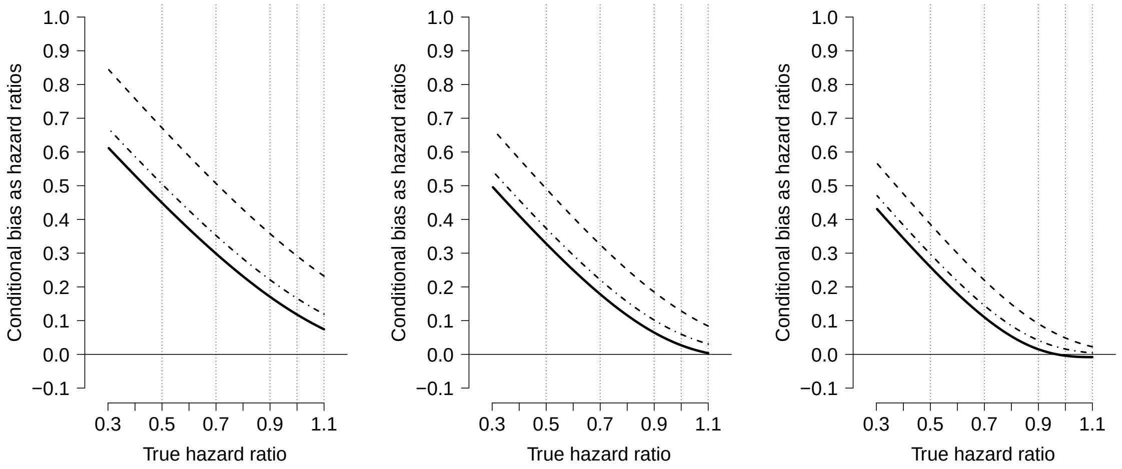
(b) 50% IT

(c) 70% IT

Figure 3.4: Conditional bias of the MLE (dashed line), CMAE (dot-dashed line), and WCMAE (solid line) when the O’Brien–Fleming-type boundary was used with the information time of 35%, 50%, and 70% (Planned for efficacy and futility, Stopped for efficacy). IT means information time. $\exp(\theta_0)$ was set to 0.5. The vertical reference lines correspond to the true hazard ratios of 0.5, 0.7, 0.9, 1.0, and 1.1, respectively.

est in all the estimators regardless of the true hazard ratio. Interestingly, the overestimation of the WCMAE when the trial stopped for futility was less serious than when the trial stopped for efficacy because of the larger bias of the MLE.

The conditional bias when the Pocock-type boundary was used is shown below. The results based on the O’Brien–Fleming-type boundary had a larger absolute conditional bias than those based on the Pocock-type. In Figures 3.7 to 3.10, although the performance of the WCMAE was close to that of the CMAE when the true hazard ratio was 1.1 and the information time was 70%, the stopping probability for efficacy at the interim analysis was relatively small. In addition, the other results for the Pocock-type boundary were similar to those for the O’Brien–Fleming-type. As a result, the WCMAE was better than the CMAE in terms of the conditional bias and the advantage of using the shrunken estimator was emphasized compared with the MLE.

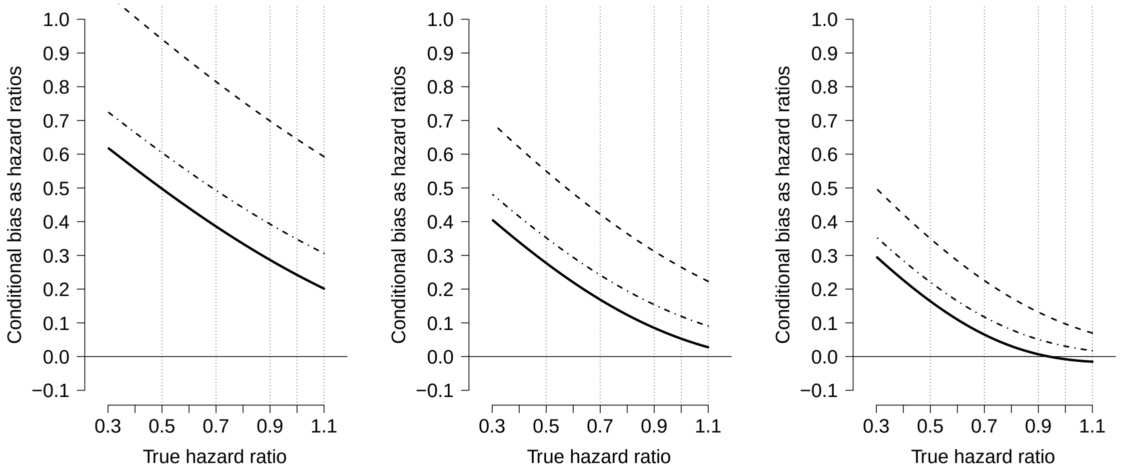


(a) 35% IT

(b) 50% IT

(c) 70% IT

Figure 3.5: Conditional bias of the MLE (dashed line), CMAE (dot-dashed line), and WCMAE (solid line) when the O'Brien–Fleming-type boundary was used with the information time of 35%, 50%, and 70% (Planned for efficacy and futility, Stopped for futility). IT means information time. $\exp(\theta_0)$ was set to 0.7. The vertical reference lines correspond to the true hazard ratios of 0.5, 0.7, 0.9, 1.0, and 1.1, respectively.

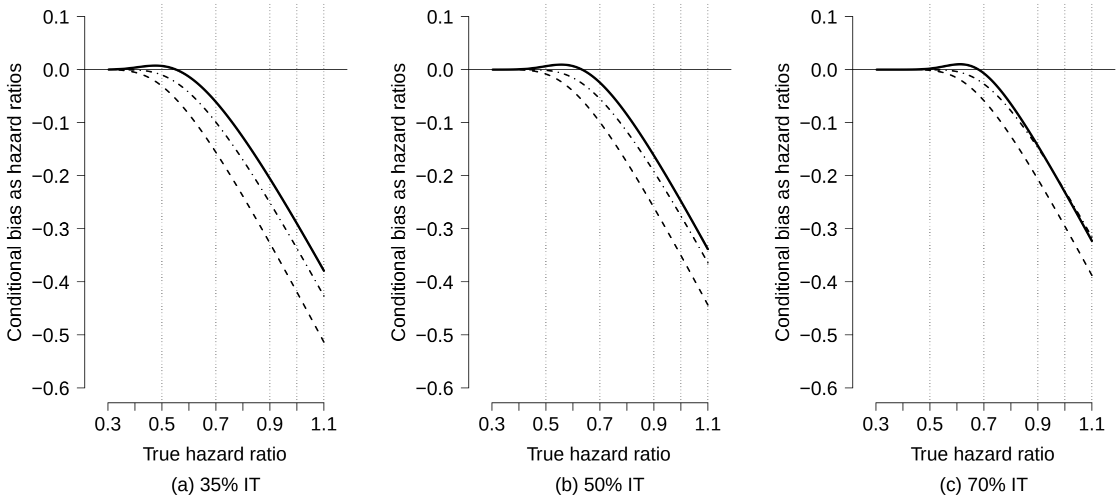


(a) 35% IT

(b) 50% IT

(c) 70% IT

Figure 3.6: Conditional bias of the MLE (dashed line), CMAE (dot-dashed line), and WCMAE (solid line) when the O'Brien–Fleming-type boundary was used with the information time of 35%, 50%, and 70% (Planned for efficacy and futility, Stopped for futility). IT means information time. $\exp(\theta_0)$ was set to 0.5. The vertical reference lines correspond to the true hazard ratios of 0.5, 0.7, 0.9, 1.0, and 1.1, respectively.

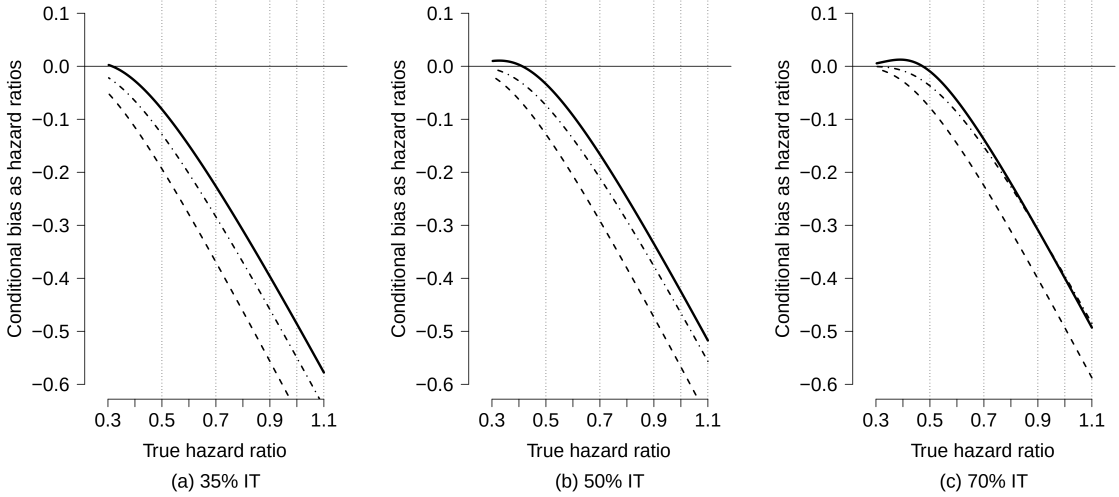


(a) 35% IT

(b) 50% IT

(c) 70% IT

Figure 3.7: Conditional bias of the MLE (dashed line), CMAE (dot-dashed line), and WCMAE (solid line) when the Pocock-type boundary was used with the information time of 35%, 50%, and 70% (Planned for efficacy, Stopped for efficacy). IT means information time. $\exp(\theta_0)$ was set to 0.7. The vertical reference lines correspond to the true hazard ratios of 0.5, 0.7, 0.9, 1.0, and 1.1, respectively.

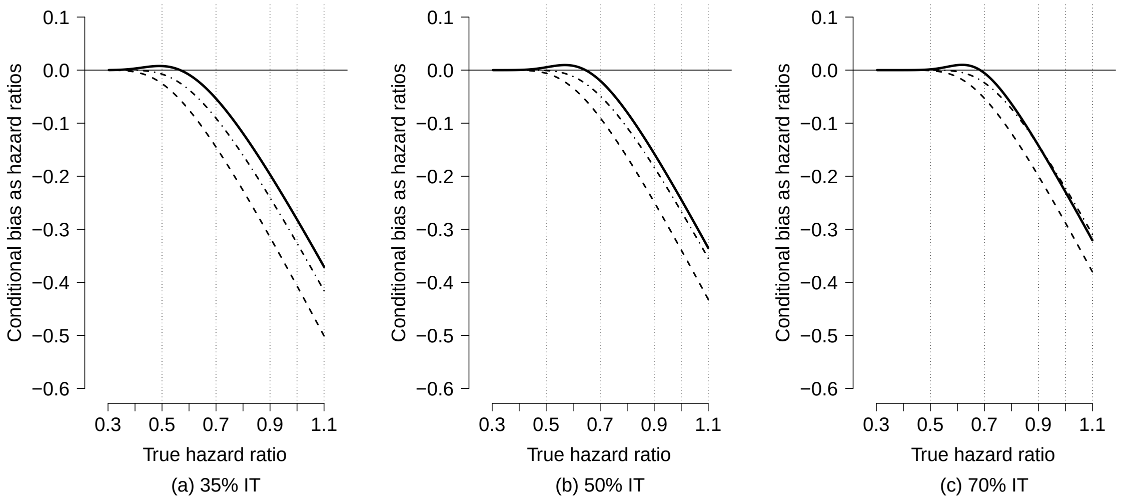


(a) 35% IT

(b) 50% IT

(c) 70% IT

Figure 3.8: Conditional bias of the MLE (dashed line), CMAE (dot-dashed line), and WCMAE (solid line) when the Pocock-type boundary was used with the information time of 35%, 50%, and 70% (Planned for efficacy, Stopped for efficacy). IT means information time. $\exp(\theta_0)$ was set to 0.5. The vertical reference lines correspond to the true hazard ratios of 0.5, 0.7, 0.9, 1.0, and 1.1, respectively.

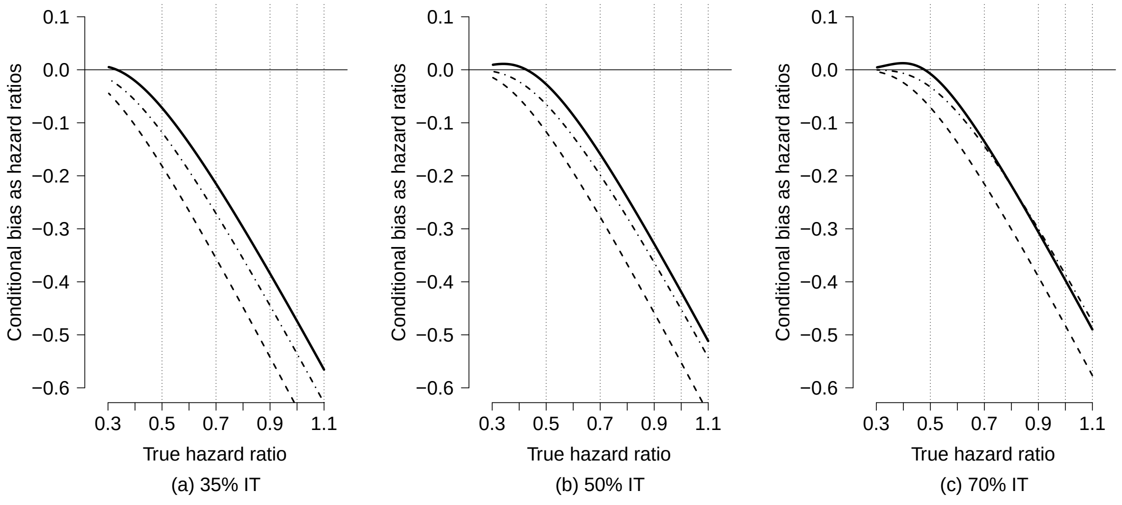


(a) 35% IT

(b) 50% IT

(c) 70% IT

Figure 3.9: Conditional bias of the MLE (dashed line), CMAE (dot-dashed line), and WCMAE (solid line) when the Pocock-type boundary was used with the information time of 35%, 50%, and 70% (Planned for efficacy and futility, Stopped for efficacy). IT means information time. $\exp(\theta_0)$ was set to 0.7. The vertical reference lines correspond to the true hazard ratios of 0.5, 0.7, 0.9, 1.0, and 1.1, respectively.

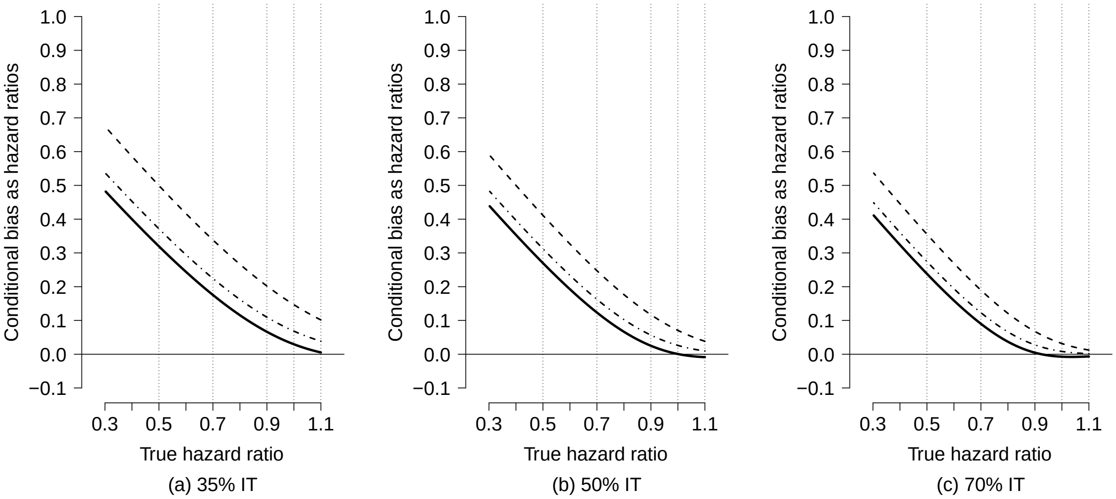


(a) 35% IT

(b) 50% IT

(c) 70% IT

Figure 3.10: Conditional bias of the MLE (dashed line), CMAE (dot-dashed line), and WCMAE (solid line) when the Pocock-type boundary was used with the information time of 35%, 50%, and 70% (Planned for efficacy and futility, Stopped for efficacy). IT means information time. $\exp(\theta_0)$ was set to 0.5. The vertical reference lines correspond to the true hazard ratios of 0.5, 0.7, 0.9, 1.0, and 1.1, respectively.

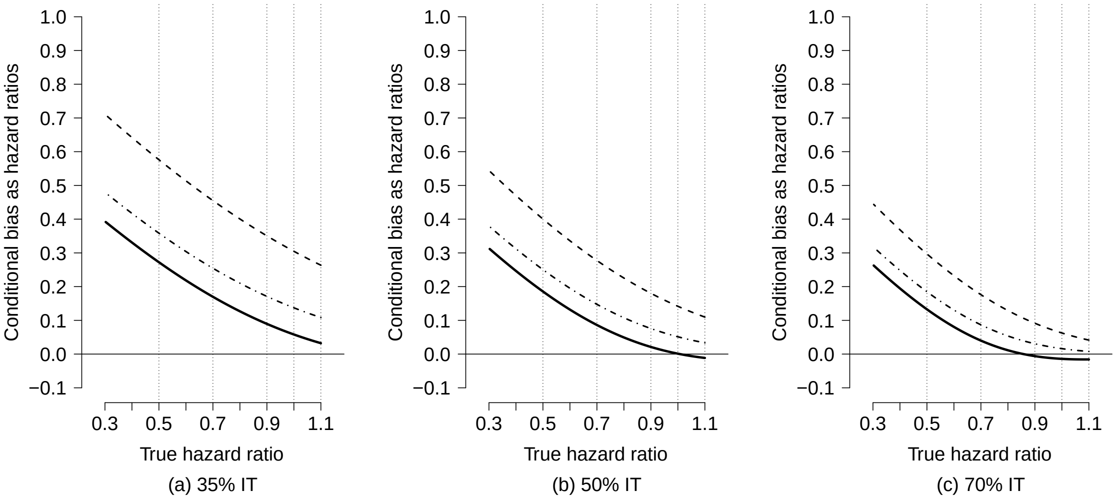


(a) 35% IT

(b) 50% IT

(c) 70% IT

Figure 3.11: Conditional bias of the MLE (dashed line), CMAE (dot-dashed line), and WCMAE (solid line) when the Pocock-type boundary was used with the information time of 35%, 50%, and 70% (Planned for efficacy and futility, Stopped for futility). IT means information time. $\exp(\theta_0)$ was set to 0.7. The vertical reference lines correspond to the true hazard ratios of 0.5, 0.7, 0.9, 1.0, and 1.1, respectively.



(a) 35% IT

(b) 50% IT

(c) 70% IT

Figure 3.12: Conditional bias of the MLE (dashed line), CMAE (dot-dashed line), and WCMAE (solid line) when the Pocock-type boundary was used with the information time of 35%, 50%, and 70% (Planned for efficacy and futility, Stopped for futility). IT means information time. $\exp(\theta_0)$ was set to 0.5. The vertical reference lines correspond to the true hazard ratios of 0.5, 0.7, 0.9, 1.0, and 1.1, respectively.

We note that the decision making of early stopping is determined by statistical hypothetical testing and the bias correction does not affect whether the trial stops. In Section 3.6, we therefore discuss the conditional mean-squared error for practical situations in detail.

3.6 Simulation study

3.6.1 Data generation and scenarios

In this section, we compare the performances of the MLE and conditional bias-adjusted estimators. Our simulation scenarios are motivated by clinical trials using two-stage GSDs. The trial characteristics such as the planned hazard ratio, sample size, number of events, and information time at the interim analysis are sourced from a previously published article.[31]

We assumed a randomized, parallel two-group comparison trial with one interim analysis in our simulation study. Two types of interim analyses (only the efficacy boundary and both the efficacy and the futility boundaries) were considered. The Lan–DeMets alpha-spending function with the Pocock- or O'Brien–Fleming-type boundaries was used. We assumed the information time for the first interim analysis to be 35%, 50%, and 70%. The experimental drug is more efficacious compared with the reference drug if the hazard ratio is less than 1 and the log hazard ratio is less than 0. Overall type I and type II errors were set to 0.05 and 0.20, respectively. We considered two scenarios for the planned effect size. The planned hazard ratio $\exp(\theta_0)$, which is used to calculate the number of events and sample size, was set to 0.7 as the "Moderate Scenario" and 0.5 as the "Enthusiastic Scenario," respectively. The accrual and follow-up times were set to three and five years, respectively. We considered the Weibull distribution with the hazard function defined as

$$\frac{\lambda t^{\lambda-1}}{\exp(\theta G)},$$

where λ is a shape parameter of the Weibull distribution, t is time from enrollment, and G is an indicator that becomes 1 for the experimental drug and 0 for the reference drug. We set λ to 0.5, 1, and 2. The hazard ratios for each shape parameter $\exp(\theta)$ were 1.1, 1.0, 0.9, 0.7, and 0.5. The Cox proportional hazards model was used to estimate the MLE at the interim analysis. Note that in GSD with the Cox regression analysis, the actual information time of survival data used in practice is based on the variance. The estimators including the proposed method need the actual information time at the interim and final analyses. However, when the trial stopped at the interim analysis, the actual information time of the final analysis, a function of the variance, is not available at the interim analysis and is just estimated by the planned number of events. We used the estimated information time of the final analysis as the actual information time to apply the conditional methods. To calculate the conditional bias and mean-squared error, we set the number of times that the estimate could be obtained to 5,000 for each condition.

We evaluated the conditional bias and mean-squared error for the MLE and bias-adjusted estimators as the hazard ratio. In the simulations, the parameter of interest was $\exp(\theta)$. The conditional bias was calculated as the average of the difference between the hazard ratio calculated by each bias-adjusted estimator and the true hazard ratio by the stopping stage. The conditional bias in stage 1 is defined as

$$\frac{1}{s_1}\sum_{k=1}^{s_1+s_2}\left\{\exp(\hat{\theta}_k)-\exp(\theta)\right\}D(M=1),$$

where $D(\bullet)$ is an indicator function that becomes 1 when $M = 1$ and s_i is the number of trials stopped in stage i. Note that s_1 is set to 5,000 but $s_1 + s_2$ is greater than or equal to 5,000. The conditional mean-squared error in stage 1 is defined as

$$\frac{1}{s_1}\sum_{k=1}^{s_1+s_2}\left\{\exp(\hat{\theta}_k)-\exp(\theta)\right\}^2 D(M=1).$$

All simulation studies were performed with R version 3.1.1.[36]

3.6.2 Results

In Tables 3.1 and 3.2, the interim analysis was planned only for efficacy. Table 1 presents the results of the conditional mean-squared error as hazard ratios for $\exp(\theta_0) = 0.7$ with $\lambda = 1.0$ (Moderate Scenario) and Table 3.2 for $\exp(\theta_0) = 0.5$ with $\lambda = 1.0$ (Enthusiastic Scenario). Each table consists of six rows: rows 1, 2, and 3 show the O'Brien–Fleming-type boundary at the first interim analysis with 35%, 50%, and 70% information time, respectively and rows 4, 5, and 6 show the Pocock-type boundary at the first interim analysis with 35%, 50%, and 70% information time, respectively. The columns contain the true hazard ratio, expected probability that the trial stops for efficacy (futility) at the interim analysis, expected probability that the null hypothesis is rejected (accepted) by the final analysis, ratio of the trial that stopped at the interim analysis in the simulation, and simulation results of the compared estimators.

Table 3.1 presents the evaluation results of the conditional mean-squared error for efficacy when $\exp(\theta_0) = 0.7$ and $\lambda = 1.0$ (Moderate Scenario). The conditional mean-squared error for all the estimators decreased as the true hazard ratio approached 0.5. Additionally, the mean-squared error became smaller when the information time at the first interim analysis increased because the variance of the estimators in stage 1 decreased. The MLE had the largest mean-squared error by 0.455 when the true hazard ratio was 1.1. The WCMAE had the smallest conditional mean-squared error of all the estimators when the true hazard ratio was more than or equal to 0.7. The main reason for these results would come from the decrease in the conditional bias of the WCMAE. The mean-squared error for the WCMAE tended to be large when the O'Brien–Fleming-type boundary was used and the true hazard ratio was 0.5. This is because the overcorrection of the conditional bias occurred in Figure 3.3. However, if HR=0.5, the maximum mean-squared error for the WCMAE was lower than 0.02 and was not serious. The conditional mean-squared error for the Pocock-type boundary was smaller than the O'Brien–Fleming-type boundary. In particular, the mean-squared error for all the estimators was lower than 0.16 regardless of the true hazard ratio when the information time was 70%.

Table 3.2 shows the conditional mean-squared error when $\exp(\theta_0) = 0.5$ and $\lambda = 1.0$ (Enthusiastic Scenario). The conditional mean-squared error for $\exp(\theta_0)$ of 0.5 was larger than that for $\exp(\theta_0)$ of 0.7. The mean-squared error for the MLE when $\exp(\theta_0) = 0.5$ was more than twice when $\exp(\theta_0) = 0.7$. The conditional mean-squared error for all the estimators decreased as the true hazard ratio approached 0.5. In addition, the mean-squared error became smaller when the information time at the first interim analysis increased. The WCMAE had the smallest mean-squared error in most of the scenarios because the conditional bias of the WCMAE was much smaller than that of the MLE and CMAE when $\exp(\theta_0) = 0.5$. In particular, when the information

Table 3.1: Conditional mean-squared error as hazard ratios (Planned for efficacy, Stopped for efficacy, $\exp(\theta_0) = 0.7$, $\lambda = 1.0$).

Scenario	HR	EP1	EP2	$s_1/(s_1+s_2)$	MLE	CMAE	WCMAE
OF Type 35% IT	1.1	0.000	0.010	0.000	0.455	0.355	0.317
	1.0	0.001	0.050	0.001	0.331	0.249	0.218
	0.9	0.004	0.181	0.003	0.229	0.164	0.141
	0.7	0.050	0.800	0.044	0.083	0.051	0.041
	0.5	0.400	0.999	0.368	0.015	0.010	0.009
OF Type 50% IT	1.1	0.001	0.011	0.001	0.299	0.217	0.180
	1.0	0.006	0.050	0.005	0.202	0.139	0.111
	0.9	0.022	0.181	0.020	0.127	0.081	0.062
	0.7	0.220	0.800	0.203	0.033	0.018	0.014
	0.5	0.814	0.999	0.800	0.006	0.010	0.014
OF Type 70% IT	1.1	0.004	0.011	0.004	0.192	0.132	0.119
	1.0	0.019	0.050	0.018	0.119	0.076	0.066
	0.9	0.073	0.180	0.070	0.064	0.037	0.030
	0.7	0.513	0.800	0.504	0.011	0.009	0.008
	0.5	0.978	0.999	0.973	0.007	0.010	0.012
P Type 35% IT	1.1	0.008	0.013	0.009	0.269	0.185	0.149
	1.0	0.024	0.050	0.023	0.181	0.117	0.090
	0.9	0.063	0.174	0.063	0.111	0.067	0.049
	0.7	0.331	0.800	0.325	0.029	0.018	0.015
	0.5	0.847	0.999	0.831	0.008	0.014	0.019
P Type 50% IT	1.1	0.009	0.012	0.009	0.200	0.134	0.119
	1.0	0.031	0.050	0.031	0.126	0.079	0.067
	0.9	0.094	0.174	0.094	0.070	0.040	0.032
	0.7	0.500	0.800	0.491	0.014	0.012	0.011
	0.5	0.961	1.000	0.960	0.007	0.012	0.015
P Type 70% IT	1.1	0.010	0.011	0.010	0.153	0.102	0.106
	1.0	0.039	0.050	0.041	0.089	0.054	0.055
	0.9	0.134	0.177	0.136	0.046	0.026	0.025
	0.7	0.672	0.800	0.661	0.008	0.010	0.008
	0.5	0.994	0.999	0.994	0.007	0.009	0.010

EP1: Expected probability to reject the null hypothesis at the interim analysis, EP2: Expected probability to reject the null hypothesis in whole the trial, $s_1/(s_1+s_2)$: Ratio of the trial that stopped at the interim analysis in the simulation for each scenario, OF Type: O'Brien–Fleming-type boundary, P Type: Pocock-type boundary, IT: Information time, HR: Hazard ratio.

time was 70% and HR $\geq$ 0.7, the WCMAE reduced the mean-squared error of the MLE by half. The conditional mean-squared error for the WCMAE was larger than the CMAE when the O'Brien–Fleming boundary was used with 70% information time and HR=0.5. This result comes from the large variance of the WCMAE rather than the overcorrection since the conditional bias for the WCMAE in Figure 3.3 was smaller than that of the CMAE in the same conditions. However, the increase in the mean-squared error for the WCMAE was not serious. The results in Tables 3.3 and 3.4 when the boundaries for both efficacy and futility were used were similar to those in Tables 3.1 and 3.2, respectively.

Tables 3.5 and 3.6 show the conditional mean-squared error for futility when $\exp(\theta_0) =$ 0.7 (Moderate Scenario) and $\exp(\theta_0) = 0.5$ (Enthusiastic Scenario) and $\lambda = 1.0$. Contrary to Table 3.1, the conditional mean-squared error for all the estimators increased as the true hazard ratio became smaller. The WCMAE had the smallest mean-squared error when the true hazard ratio $\leq$ 0.7. The mean-squared error of the WCMAE was sometimes larger than that of the MLE and CMAE when HR $>$ 0.7, although the conditional bias of the WCMAE for futility was the smallest in Section 3.5. This finding indicated that the large mean-squared error of the WCMAE was caused by the large variance of the WCMAE rather than the overcorrection of the conditional bias. However, in Tables 3.5 and 3.6, the maximum difference of the mean-squared error between the WCMAE and CMAE was less than 0.012 and was not large.

The results for the cases in which $\lambda = 0.5$ and $\lambda = 2.0$ were similar to the case in which $\lambda = 1.0$ (data not shown). These results indicated that the same discussions of the exponential distribution could be applied to the Weibull distributions regardless of the shape parameter under the proportional hazard assumption.

From these results and Section 3.5, if θ_0 is planned enthusiastically, the WCMAE

Table 3.2: Conditional mean-squared error as hazard ratios (Planned for efficacy, Stopped for efficacy, $\exp(\theta_0) = 0.5$, $\lambda = 1.0$).

Scenario	HR	EP1	EP2	$s_1/(s_1+s_2)$	MLE	CMAE	WCMAE
OF Type	1.1	0.000	0.024	0.000	0.898	0.793	0.757
35% IT	1.0	0.001	0.050	0.000	0.716	0.624	0.593
	0.9	0.002	0.103	0.001	0.555	0.476	0.449
	0.7	0.009	0.357	0.006	0.298	0.244	0.226
	0.5	0.050	0.800	0.034	0.121	0.092	0.083
OF Type	1.1	0.003	0.024	0.002	0.678	0.534	0.479
50% IT	1.0	0.006	0.050	0.003	0.522	0.399	0.352
	0.9	0.012	0.102	0.007	0.390	0.291	0.253
	0.7	0.052	0.357	0.035	0.182	0.122	0.101
	0.5	0.220	0.800	0.170	0.059	0.035	0.029
OF Type	1.1	0.009	0.024	0.008	0.450	0.317	0.258
70% IT	1.0	0.019	0.050	0.017	0.329	0.221	0.174
	0.9	0.040	0.102	0.035	0.228	0.146	0.111
	0.7	0.161	0.356	0.147	0.086	0.048	0.035
	0.5	0.513	0.800	0.485	0.020	0.016	0.020
P Type	1.1	0.014	0.025	0.014	0.610	0.448	0.373
35% IT	1.0	0.024	0.050	0.024	0.467	0.333	0.272
	0.9	0.040	0.099	0.038	0.344	0.237	0.190
	0.7	0.117	0.346	0.109	0.160	0.102	0.080
	0.5	0.331	0.800	0.311	0.050	0.032	0.030
P Type	1.1	0.017	0.025	0.017	0.470	0.328	0.261
50% IT	1.0	0.031	0.050	0.031	0.348	0.234	0.182
	0.9	0.057	0.099	0.056	0.246	0.157	0.119
	0.7	0.182	0.347	0.174	0.098	0.056	0.043
	0.5	0.500	0.800	0.471	0.025	0.021	0.025
P Type	1.1	0.020	0.024	0.020	0.361	0.241	0.186
70% IT	1.0	0.039	0.050	0.040	0.256	0.162	0.121
	0.9	0.078	0.101	0.077	0.172	0.104	0.075
	0.7	0.266	0.351	0.257	0.058	0.033	0.027
	0.5	0.672	0.800	0.660	0.014	0.019	0.026

EP1: Expected probability to reject the null hypothesis at the interim analysis, EP2: Expected probability to reject the null hypothesis in whole the trial, $s_1/(s_1+s_2)$: Ratio of the trial that stopped at the interim analysis in the simulation for each scenario, OF Type: O'Brien–Fleming-type boundary, P Type: Pocock-type boundary, IT: Information time, HR: Hazard ratio.

Table 3.3: Conditional mean-squared error as hazard ratios (Planned for efficacy and futility, Stopped for efficacy, $\exp(\theta_0) = 0.7$, $\lambda = 1.0$).

Scenario	HR	EP1	EP2	$s_1/(s_1+s_2)$	MLE	CMAE	WCMAE
OF Type	1.1	0.000	0.010	0.000	0.450	0.351	0.313
35% IT	1.0	0.001	0.050	0.001	0.328	0.247	0.216
	0.9	0.004	0.182	0.003	0.227	0.163	0.139
	0.7	0.051	0.800	0.045	0.082	0.051	0.041
	0.5	0.408	0.999	0.374	0.015	0.010	0.009
OF Type	1.1	0.001	0.010	0.001	0.289	0.210	0.174
50% IT	1.0	0.006	0.050	0.005	0.195	0.134	0.107
	0.9	0.022	0.182	0.021	0.121	0.077	0.059
	0.7	0.228	0.800	0.213	0.030	0.016	0.013
	0.5	0.828	0.999	0.799	0.006	0.010	0.013
OF Type	1.1	0.004	0.010	0.004	0.187	0.129	0.118
70% IT	1.0	0.019	0.050	0.018	0.115	0.073	0.064
	0.9	0.075	0.182	0.072	0.063	0.037	0.030
	0.7	0.531	0.800	0.518	0.011	0.009	0.008
	0.5	0.982	0.999	0.978	0.006	0.009	0.011
P Type	1.1	0.008	0.012	0.008	0.257	0.176	0.143
35% IT	1.0	0.024	0.050	0.024	0.171	0.111	0.087
	0.9	0.066	0.178	0.064	0.103	0.062	0.046
	0.7	0.355	0.800	0.355	0.025	0.016	0.014
	0.5	0.875	0.998	0.866	0.008	0.014	0.019
P Type	1.1	0.009	0.012	0.009	0.190	0.128	0.116
50% IT	1.0	0.031	0.050	0.031	0.118	0.074	0.064
	0.9	0.098	0.179	0.097	0.065	0.037	0.030
	0.7	0.531	0.800	0.526	0.012	0.011	0.010
	0.5	0.972	0.999	0.968	0.008	0.012	0.015
P Type	1.1	0.009	0.011	0.009	0.146	0.096	0.103
70% IT	1.0	0.039	0.050	0.040	0.084	0.051	0.053
	0.9	0.139	0.180	0.138	0.042	0.023	0.023
	0.7	0.697	0.800	0.691	0.008	0.010	0.008
	0.5	0.996	0.999	0.995	0.006	0.008	0.009

EP1: Expected probability to reject the null hypothesis at the interim analysis, EP2: Expected probability to reject the null hypothesis in whole the trial, $s_1/(s_1+s_2)$: Ratio of the trial that stopped at the interim analysis in the simulation for each scenario, OF Type: O'Brien–Fleming-type boundary, P Type: Pocock-type boundary, IT: Information time, HR: Hazard ratio.

Table 3.4: Conditional mean-squared error as hazard ratios (Planned for efficacy and futility, Stopped for efficacy, $\exp(\theta_0) = 0.5$, $\lambda = 1.0$).

Scenario	HR	EP1	EP2	$s_1/(s_1+s_2)$	MLE	CMAE	WCMAE
OF Type	1.1	0.000	0.023	0.000	0.903	0.802	0.760
35% IT	1.0	0.001	0.050	0.000	0.720	0.629	0.591
	0.9	0.002	0.103	0.001	0.559	0.481	0.449
	0.7	0.009	0.359	0.005	0.298	0.245	0.223
	0.5	0.051	0.800	0.033	0.122	0.093	0.082
OF Type	1.1	0.003	0.023	0.002	0.661	0.519	0.450
50% IT	1.0	0.006	0.050	0.004	0.508	0.387	0.329
	0.9	0.012	0.103	0.008	0.375	0.276	0.229
	0.7	0.053	0.359	0.038	0.175	0.117	0.091
	0.5	0.228	0.800	0.180	0.054	0.032	0.025
OF Type	1.1	0.009	0.023	0.008	0.447	0.317	0.281
70% IT	1.0	0.019	0.050	0.016	0.326	0.221	0.192
	0.9	0.041	0.103	0.034	0.226	0.145	0.121
	0.7	0.167	0.359	0.143	0.085	0.047	0.036
	0.5	0.531	0.800	0.482	0.019	0.016	0.015
P Type	1.1	0.014	0.025	0.014	0.577	0.421	0.348
35% IT	1.0	0.024	0.050	0.023	0.438	0.309	0.250
	0.9	0.041	0.101	0.040	0.320	0.217	0.171
	0.7	0.124	0.355	0.120	0.143	0.089	0.066
	0.5	0.355	0.800	0.335	0.043	0.028	0.026
P Type	1.1	0.016	0.024	0.017	0.442	0.307	0.275
50% IT	1.0	0.031	0.050	0.031	0.324	0.215	0.188
	0.9	0.058	0.101	0.059	0.226	0.143	0.121
	0.7	0.193	0.355	0.181	0.086	0.049	0.038
	0.5	0.531	0.800	0.508	0.021	0.020	0.018
P Type	1.1	0.019	0.024	0.020	0.344	0.229	0.240
70% IT	1.0	0.039	0.050	0.040	0.243	0.154	0.160
	0.9	0.079	0.102	0.079	0.160	0.095	0.097
	0.7	0.278	0.356	0.274	0.053	0.031	0.027
	0.5	0.697	0.800	0.689	0.013	0.018	0.015

EP1: Expected probability to reject the null hypothesis at the interim analysis, EP2: Expected probability to reject the null hypothesis in whole the trial, $s_1/(s_1+s_2)$: Ratio of the trial that stopped at the interim analysis in the simulation for each scenario, OF Type: O'Brien–Fleming-type boundary, P Type: Pocock-type boundary, IT: Information time, HR: Hazard ratio.

Table 3.5: Conditional mean-squared error as hazard ratios (Planned for efficacy and futility, Stopped for futility, $\exp(\theta_0) = 0.7$, $\lambda = 1.0$).

Scenario	HR	EP1	EP2	$s_1/(s_1+s_2)$	MLE	CMAE	WCMAE
OF Type	1.1	0.501	0.990	0.491	0.109	0.086	0.084
35% IT	1.0	0.347	0.950	0.334	0.130	0.085	0.076
	0.9	0.203	0.818	0.201	0.164	0.094	0.078
	0.7	0.030	0.200	0.031	0.276	0.148	0.115
	0.5	0.001	0.001	0.001	0.464	0.270	0.217
OF Type	1.1	0.787	0.990	0.781	0.047	0.054	0.058
50% IT	1.0	0.624	0.950	0.621	0.046	0.044	0.045
	0.9	0.416	0.818	0.413	0.055	0.039	0.036
	0.7	0.070	0.200	0.069	0.118	0.065	0.049
	0.5	0.001	0.001	0.001	0.246	0.146	0.115
OF Type	1.1	0.942	0.990	0.943	0.032	0.040	0.044
70% IT	1.0	0.842	0.950	0.842	0.025	0.030	0.034
	0.9	0.643	0.818	0.633	0.024	0.024	0.025
	0.7	0.126	0.200	0.126	0.055	0.031	0.023
	0.5	0.001	0.001	0.001	0.152	0.093	0.072
P Type	1.1	0.767	0.988	0.758	0.062	0.069	0.075
35% IT	1.0	0.617	0.950	0.600	0.061	0.057	0.059
	0.9	0.429	0.822	0.418	0.070	0.050	0.047
	0.7	0.094	0.200	0.095	0.131	0.071	0.053
	0.5	0.002	0.002	0.003	0.259	0.149	0.114
P Type	1.1	0.905	0.988	0.898	0.039	0.048	0.053
50% IT	1.0	0.785	0.950	0.781	0.033	0.038	0.042
	0.9	0.585	0.821	0.570	0.033	0.029	0.030
	0.7	0.124	0.200	0.125	0.071	0.039	0.030
	0.5	0.001	0.001	0.002	0.175	0.104	0.080
P Type	1.1	0.970	0.989	0.967	0.030	0.036	0.040
70% IT	1.0	0.898	0.950	0.891	0.023	0.029	0.033
	0.9	0.725	0.820	0.718	0.019	0.021	0.024
	0.7	0.158	0.200	0.157	0.042	0.023	0.018
	0.5	0.001	0.001	0.001	0.130	0.079	0.061

EP1: Expected probability to accept the null hypothesis at the interim analysis, EP2: Expected probability to accept the null hypothesis in whole the trial, $s_1/(s_1+s_2)$: Ratio of the trial that stopped at the interim analysis in the simulation for each scenario, OF Type: O'Brien–Fleming-type boundary, P Type: Pocock-type boundary, IT: Information time, HR: Hazard ratio.

Table 3.6: Conditional mean-squared error as hazard ratios (Planned for efficacy and futility, Stopped for futility, $\exp(\theta_0) = 0.5$, $\lambda = 1.0$).

Scenario	HR	EP1	EP2	$s_1/(s_1+s_2)$	MLE	CMAE	WCMAE
OF Type	1.1	0.425	0.977	0.403	1.018	0.736	0.672
35% IT	1.0	0.347	0.950	0.326	0.989	0.669	0.596
	0.9	0.268	0.897	0.257	1.066	0.708	0.627
	0.7	0.124	0.641	0.119	0.986	0.531	0.427
	0.5	0.030	0.800	0.030	1.109	0.554	0.426
OF Type	1.1	0.713	0.977	0.660	0.323	0.306	0.306
50% IT	1.0	0.624	0.950	0.572	0.303	0.264	0.258
	0.9	0.517	0.897	0.488	0.310	0.247	0.233
	0.7	0.272	0.641	0.254	0.315	0.197	0.167
	0.5	0.070	0.200	0.070	0.393	0.207	0.158
OF Type	1.1	0.903	0.977	0.892	0.162	0.181	0.192
70% IT	1.0	0.842	0.950	0.825	0.135	0.147	0.155
	0.9	0.750	0.897	0.730	0.113	0.116	0.121
	0.7	0.458	0.641	0.440	0.104	0.077	0.072
	0.5	0.126	0.200	0.122	0.158	0.088	0.067
P Type	1.1	0.698	0.975	0.665	0.462	0.438	0.436
35% IT	1.0	0.617	0.950	0.596	0.416	0.367	0.357
	0.9	0.521	0.899	0.499	0.377	0.305	0.290
	0.7	0.297	0.645	0.290	0.387	0.253	0.217
	0.5	0.094	0.200	0.092	0.441	0.236	0.178
P Type	1.1	0.855	0.976	0.840	0.204	0.222	0.233
50% IT	1.0	0.785	0.950	0.769	0.177	0.185	0.194
	0.9	0.689	0.899	0.673	0.156	0.152	0.155
	0.7	0.416	0.645	0.411	0.149	0.106	0.095
	0.5	0.124	0.200	0.125	0.211	0.118	0.091
P Type	1.1	0.943	0.976	0.944	0.131	0.150	0.162
70% IT	1.0	0.898	0.950	0.891	0.109	0.125	0.135
	0.9	0.822	0.898	0.813	0.087	0.096	0.103
	0.7	0.540	0.644	0.528	0.075	0.063	0.061
	0.5	0.158	0.200	0.159	0.112	0.064	0.049

EP1: Expected probability to accept the null hypothesis at the interim analysis, EP2: Expected probability to accept the null hypothesis in whole the trial, $s_1/(s_1+s_2)$: Ratio of the trial that stopped at the interim analysis in the simulation for each scenario, OF Type: O'Brien–Fleming-type boundary, P Type: Pocock-type boundary, IT: Information time, HR: Hazard ratio.

could improve the conditional bias and mean-squared error of the CMAE regardless of the true hazard ratio. Additionally, for θ_0 in the moderate setting, the WCMAE showed better performance than the CMAE, especially when the information time at the interim analysis was small.

3.7 Application

We applied the bias-adjusted estimators to the published data. The trial described in Oza et al.[31] is a randomized, open-label, phase II trial comparing (a) oral Ridaforolimus with (b) progestin or investigator choice chemotherapy in patients presenting with metastatic or recurrent endometrial cancer who had progressive disease following one or two lines of chemotherapy and no hormonal therapy. The primary endpoint was progression-free survival assessed by independent radiologic review. The planned log hazard ratio was -0.405, corresponding to a hazard ratio of 0.667, and the required number of events to detect approximately 80% power was 121. One interim analysis for efficacy was planned at 48% of the final number of events. Although no specific alpha-spending function was described in the article or protocol, we assumed that an O'Brien–Fleming- or Pocock-type boundary was used to control the cumulative one-sided significance level, 0.10. In total, 130 patients were randomized to (a) Ridaforolimus (64 patients) or (b) progestin or chemotherapy (66 patients). The accrual time was 48 months. One interim analysis for progression-free survival was conducted and the estimated hazard ratio crossed the stopping boundary. The MLE for the hazard ratio calculated at the interim analysis was 0.53, corresponding to a log hazard ratio of -0.635. The CMAE and WCMAE were applied to this study to estimate the hazard ratio. The number of events at the interim analysis was 58. The Fisher information in stage 1 was approximated as 58/4. Table 3.7 presents the hazard ratio for each estimator.

The CMAE and WCMAE estimated a smaller difference in the treatment effect than the MLE regardless of the alpha-spending function. If the O'Brien–Fleming-type boundary was used, the bias-correction by the CMAE and WCMAE was larger than that when the Pocock-type boundary was used. The required events of 58 was standard for a phase II trial but not large for phase III. Substantial conditional bias may be large in this example because the smaller the number of events at the interim analysis, the smaller the Fisher information is. The planned hazard ratio of 0.667 was relatively close to the Moderate Scenario in our simulation setting. Although the WCMAE might overcorrect depending on the true hazard ratio, our simulations suggest that the WCMAE is a minimally biased estimator in studies with early termination. This result indicates that the estimated hazard ratio from studies with termination tends to slightly overestimate the treatment effects. Therefore, we recommend the use of the value 0.675 or 0.617, which is the estimate of the hazard ratio from the WCMAE, as the adjusted value.

Table 3.7: Hazard ratio using bias-adjusted estimators by boundary

Boundary	MLE	CMAE	$\exp(\theta_0)$	$\exp(\hat{\theta}_*)$	WCMAE
O'Brien–Fleming-type	0.53	0.622	0.667	0.604	0.675
Pocock-type	0.53	0.580	0.667	0.604	0.617

3.8 Summary

In this study, we proposed and evaluated the performance of the WCMAE through analytical and simulation studies. In the scenarios, one interim analysis for efficacy and futility was conducted and the trial stopped. In the situation where the trial is stopped early, information about the treatment effect will be limited and the motivation to borrow prior information becomes large. The WCMAE is an intuitive estimator because it is constructed with a prespecified difference in the treatment effect and MLE. We used the MLE to define the weight for shrinkage. This property is desirable in practical settings because we can calculate the WCMAE without any additional parameters. This would be appropriate when there is no specific information about the accuracy of the prior information. The WCMAE can also be applied to the MLE reported in the literature. The value of θ_0, such as the planned hazard ratio, can be easily captured from medical articles. Indeed, Sato et al.[37] and Gosho et al.,[14] who reviewed 238 articles published in *The New England Journal of Medicine*, reported an increase in the calculations of the power and sample size.

Chapter 4

Application of conditional bias-adjusted estimators to actual oncology clinical trials

4.1 Introduction

As shown in Chapters 2 and 3, several researchers have proposed bias-adjusted estimators such as the CMAE and WCMAE to address the issue.[40,47] However, in published clinical trials, the bias-adjusted estimators have been rarely reported, suggesting the naïve estimates by the MLE is nonchalantly presented in data monitoring committee (DMC) for interim analysis.[55] To emphasize the need for considering overestimation of hazard ratio, we applied the bias-adjusted estimators to early-terminated oncology clinical trials published in major journals.

However, in published clinical trials, a bias-adjusted estimator, such as the hazard ratio (HR), has been rarely reported, suggesting that naïve estimates obtained using the maximum likelihood method may be nonchalantly presented to the data monitoring committee (DMC) at the interim analysis.[55] Presently, we applied the bias-adjusted estimators to early-terminated oncology clinical trials published in major journals to emphasize the need for considering HR overestimation.

4.2 Search strategy for systematic review

This review was restricted to oncology clinical trials using GSD with a pre-planned interim analysis. We included parallel group randomized clinical trials that were halted for efficacy considerations. We identified trials published each year from 2013 to 2017 by a search of MEDLINE and EMBASE. The search was restricted to 11 scientific journals: *NEJM, JCO, Cancer discovery, Lancet, Lancet Oncology, JAMA, JAMA Oncology, CA Cancer Journal for Clinicians, Annals of Internal Medicine, Nature Reviews Clinical Oncology,* and *British Medical Journal*. A free text search employed relevant keywords, which included

"*hazard ratio*" AND ("*interim analys(i/e)s*" OR "*group sequential*" OR "*two stage*" OR "*stop*" "*stopping*" OR "*terminate*" OR "*termination*" OR "*halt*" OR "*close*" OR "*continue*" OR "*continuation*" OR "*prematurely*" OR "*independent data monitoring*" OR "*data and safety monitoring board*" OR "*DSMB*" OR "*Brien-Fleming*" OR "*Pocock*" OR "*'Lan-DeMets*" OR "*Fisher information*" OR "*boundar(y/ies)*"). These keywords were used in combination with additional eligibility filters, namely publication year ("*2013-2017*") and check tags ("*humans*" AND "*English*" AND ("*article*" OR "*article in press*"), ("*new england journal of medicine*" OR "*jama - journal of the american medical association*" OR "*the lancet*" OR "*british medical journal*" OR "*annals of in-*

ternal medicine" OR "*journal of clinical oncology*" OR "*nature reviews cancer*" OR "*cancer cell*" OR "*nature reviews clinical oncology*" OR "*cancer discovery*" OR "*jama oncology*" OR "*ca cancer journal for clinicians*" OR "*the lancet oncology*")).

We excluded articles mainly describing statistical methodologies, subgroup analysis, retrospective analysis, meta-analysis, non-inferiority clinical trials, and post-hoc analysis. In addition, trials that had been stopped but did not meet the efficacy criteria for termination were also excluded. Duplicate articles were eliminated by comparing the registration number from three registration databases: clinical trial.gov, UMIN, and ISRCTN Registry. For each article, the eligibility was independently reviewed twice by three biostatisticians.

4.3 Results and summary

A total of 198 abstracts were screened for eligibility. Of these, we excluded 98 articles that were obviously not eligible in this phase. The full text of the remaining 101 articles was assessed to identify 19 eligible clinical trials to apply the bias-adjusted estimators. If the information needed to calculate the bias-adjusted estimates (alpha-spending function, number of events, and planned HR) was missing, we complemented the information by referring to the protocols, statistical analysis plan, and related articles. Of the 19 eligible trials, two (No. 80 and No. 117) each included two efficacy endpoints and one planned interim analysis. However, these two trials met the termination criteria only for one endpoint. There were no trials that used any bias-adjusted estimators.

Figure 4.1 shows the naïve HR, HRs adjusted by the CMAE and WCMAE, endpoints, and number of events at interim analysis in each trial. The experimental treatment is more efficacious than the reference treatment if the HR < 1. The most common type of the endpoint for the interim analysis was progression-free survival time (PFS: 79% of the trials), followed by overall survival time (16%) and disease-free survival time (5%). The naïve HR, number of events at interim analysis, and information time ranged from 0.2 to 0.71, 58 to 540, and 48% to 82%, respectively. The endpoint was progression-free survival time in all the trials with the naïve HR ≤ 0.6. The adjusted HRs by the CMAE and WCMAE were higher than the naïve HR.

As shown in Figure 4.1, the difference between the naïve HR and adjusted HRs depended on the number of events at the interim analysis and the naïve HR itself. The bias-adjusted estimators in the large trials, such as No. 51 (243 events) and No. 80 (414 events), were similar to the naïve HR. However, the bias-adjusted estimators in the small trials, such as No. 88 (58 events), were highly disparate from the naïve HR. The importance of the number of events is easy to interpret; the data about the treatment effect from the small trials is generally limited and the data at the interim analysis is smaller than that at the final analysis. Thus, the positive result at the interim analysis tends to exaggerate the treatment effect when the trials are halted for efficacy reasons. DMC members and stakeholders of clinical trial sponsors should carefully interpret and report the results from such trials when the number of events at interim analysis is quite small. On the contrary, even if the number of events at interim analysis was small, the difference between naïve and bias-adjusted HRs became small in the trials that showed a large treatment effect (e.g., No. 53 and No. 140). For a treatment that showed an extremely positive effect, such as 0.20 (No. 78) and 0.22 (No. 140), the risk of the overestimation of naïve HR would be avoided.

If the estimated treatment effect is not overly large, the naïve HR should be inter-

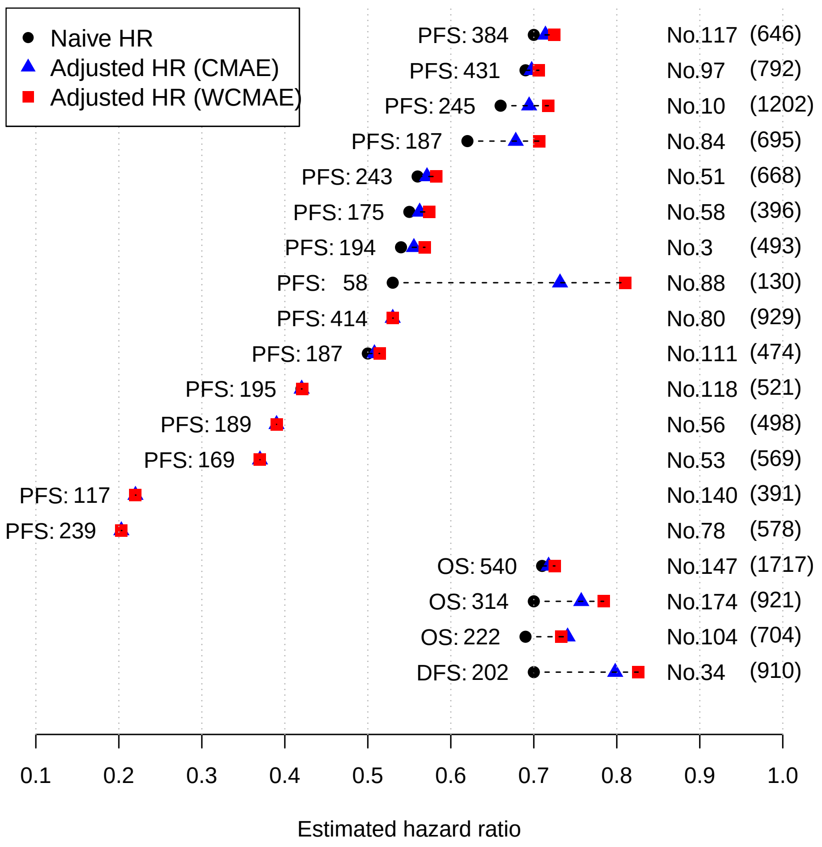


Figure 4.1: Naïve and bias-adjusted hazard ratios with the endpoints in 19 eligible oncology clinical trials. The trials were sorted by endpoints and naïve HR. The values adjacent to the endpoints and study numbers represent the number of events at interim analysis and sample sizes in each trial, respectively

preted with skepticism. It is difficult to control the result before starting the trial, but researchers can determine the number of events at interim analysis in advance. As the number of events at interim analysis increases, the magnitude of the overestimation decreases. Increasing the number of events at interim analysis when planning clinical trials may help reduce the overestimation. We assume that large number of events have unintentionally contributed to suppress the conditional bias even in clinical trials that did not use the conditional bias-adjusted estimators.

Chapter 5

Discussion and conclusion

It is difficult to develop a “completely” conditional unbiased estimator because the true hazard ratio is unknown;[27,47] therefore, it is recommended to use an estimator with less conditional bias regardless of the unknown true hazard ratio and prior information. In the trial for novel drugs, the prior information about θ would often be insufficient because no rival or similar trials are conducted. However, even in these situations, an unreasonably large or small θ_0 would not be used. Therefore, it is important to assess the performance of the WCMAE in practical simulation settings. Our analytical results indicated that the WCMAE comprehensively improves the CMAE in terms of the conditional bias, although the performance of the conditional bias-adjusted estimator depends on the true difference in the treatment effect. The results also indicated that the WCMAE somewhat overcorrects the negative conditional bias, reducing it more than necessary when the true conditional bias of the MLE is small. This bias regards the difference in the treatment effect as smaller than is true, which is less serious for a confirmatory trial than is negative bias from the perspective of a regulating authority, as it leads to a conservative result. It is therefore better to evaluate the stability of the adjusted estimator by comparing it with other estimates, including the MLE.

Other types of shrinkage estimators have been proposed in multi-armed trials. Lindley's estimator shrinks the MLE of the selected treatment toward the sample mean of all the treatments.[25] Hwang showed that Lindley's shrinkage estimator had preferable properties in terms of the mean-squared error when the most promising arms was selected in single-stage multi-armed trials with $k(k \geq 4)$ experimental treatments.[19] Carreras and Brannath considered an extension of Hwang's result to the problem of overestimation in adaptive two-stage designs with the selection of a single treatment arm.[9] The situation they considered corresponds to the case where the trial does not stop at the first interim analysis in two-stage designs. An estimator using not only the first-stage data but also the second-stage data was developed by Bowden et al.[6] Their methods use the results of the other arms in the first stage when the trial does not stop. Contrary to multi-armed and continued trials, little information on the treatment effect can be used in two-armed and stopped trials. Additionally, GSDs are frequently two-armed trials. Therefore, the proposed method using prior information would be useful in practical situations.

For the WCMAE, we did not include the results for the confidence intervals even though they are interpreted in terms of not only the point estimates but also the confidence intervals. An exact conditional confidence interval was proposed by Ohman Strickland and Casella.[30] However, as pointed out by those authors as well as Fan and DeMets,[13] the conditional interval suffers from two undesirable properties when $\hat{\theta}_{MLE,1}$ is close to a_m and the trial stops early. Firstly, the confidence interval becomes extremely wide. For example, if $\hat{\theta}_{MLE,1}$ goes to a_m, the width of the confidence interval converges to infinity. Secondly, inconsistency between the confidence interval and result of the hypothesis testing arises. As mentioned above, obtaining a completely unbiased estimator is difficult when imprecise prior information is given. As an alternative approach, Pepe et al.[32] and Shimura et al.[39] used a nonparametric bootstrap confidence interval.

The WCMAE is calculated by adopting the frequentist approach. On the contrary, $\hat{\theta}_*$ can also be interpreted as the expectation of the posterior distribution in the Bayesian paradigm.[9, 10, 19, 20] As discussed in Section 3.4.1, the planned effect size is determined by using information such as the results of previous clinical trials of the drug, nonclinical trials, and clinical trials of rival drugs. This information can be regarded as informative prior information for estimating the difference in the treatment effects. As described in Section 1.4, $(\hat{\theta}_{MLE,1}, \hat{\theta}_{MLE,2})$ follows a bivariate normal distribution. Assuming a normal distribution with mean θ_0 and variance $1/(1-c)$, like the prior distribution, the distribution of the MLE and prior distributions are conjugate distributions and easy to combine. In this case, the posterior distribution of θ after stopping at the interim analysis is denoted as $N(\hat{\theta}_*, \sigma_*^2)$, where

$$\sigma_*^2 = \left(\frac{1}{c} + \frac{1}{1-c}\right)^{-1}. \tag{5.1}$$

For multi-armed trials with $k(k \geq 4)$ experimental treatments, Hwang proved that Lindley's shrinkage estimator had smaller Bayes risk under squared loss than the MLE of the most promising arms for all independent and identically distributed prior normal distributions.[19] Although Thompson's shrunken estimator was not discussed in their studies, the analytical and simulation results presented in Sections 3.5 and 3.6.2 showed the better performances of the WCMAE in terms of the conditional bias and mean-squared estimator. The variance of the prior distribution has been discussed under the paradigm of the Bayesian GSD.[15, 20, 43]

One limitation of our study is that we cannot guarantee that the planned difference in the treatment effects truly reflects the actual difference, as the prior information for planning group sequential trials may be uncertain for rare diseases for which too few patients can be recruited. If the prior information is insufficient, the use of θ_0 for the prior information may be unreasonable. Moreover, we evaluated the performance of the WCMAE in trials with only one interim analysis. It may be meaningful to compare the overall bias between the conditional bias-adjusted estimators, including the WCMAE, in two-stage designs. In addition, conditional bias-adjusted estimators exist in adaptive designs.[7, 22] An extension of the WCMAE to adaptive designs may also be needed. As a trial design becomes more complex, the building of conditional estimators is mathematically more difficult. Regarding this, Bebu et al.,[5] Bowden and Glimm,[7] and Kimani et al.[22] discussed the conditional estimation in the adaptive design.

Other limitation in the applicability of this study originates from the proportional hazard model assumption. For example, the development of molecular targeted anticancer drugs is becoming popular, such as kinase inhibitors and immune checkpoint inhibitors. In these therapies, proportionality assumption can not be applied because of the delayed effect or subgroup-specific efficacy. Precision medicine is an approach that allows doctors to select treatments that are most likely to be effective based on genetic, environmental, and lifestyle factors. Targeted therapy, such as antibodies and tyrosine kinase inhibitors, provides the foundation of precision medicine. The use of “Real world data” may also allow us to obtain new approaches for precision medicine. The precision medicine is based on the idea that the treatment effect is different by each population. In the statistical point of view, this means that there is an interaction between the covariate and treatment;therefore, the proportional hazard assumption may not be valid. Actually, some recent clinical trials showed the delayed separation in the survival curve between the experimental and reference groups.[49] These results suggest that the effect size expressed by the hazard ratio changes depending on time and the proportional hazard assumption is not valid for the trials. Many of the statistical methods including fixed sample design and GSD are based on the assumption. Even the fixed sample design, the power may seriously decrease if a method assuming the proportional assumption was used when the assumption was not valid. The problem of the non-proportional situation is also problematic in GSD. As described in Section 1.4, the GSD is based the canonical distribution for the sequence of the test statistics. Therefore, in the case that the proportional assumption is not valid, adopting the GSD would be in appropriate and the bias-adjusted estimator could not be used. Extension of the GSD to the non-proportional situation is needed because that is considered to be clinically meaningful.

The issues of this study are concluded as follows: When planning an interim analysis, we recommend using the CUMVUE for trials that do not terminate early. On the contrary, we recommend the use of the proposed estimator for trials that are terminated early for efficacy or futility.

Acknowledgments

I genuinely thank Professor Yoshihiro Arakawa, who, in addition to his constant guidance and advice, also provided a research environment and numerous opportunities. His kind comments and suggestions regarding matters of a personal nature are also deeply appreciated. This study would be incomplete without his encouragement and guidance.

I am immeasurably indebted to Associate Professor Kazumasa Yamagishi, Associate Professor Ken-ichi Koike, and Lecturer Ayumi Takayashiki, who consistently supported and encouraged me throughout this study. They constituted my doctoral committee members and provided many helpful comments and suggestions.

I express my appreciation for Professor Masahiko Gosho and Associate Professor Kazushi Maruo, University of Tsukuba. Had it not been for their unfailing positivity and guidance, I would not have accomplished what is set forth here, for I learned from them not only the facts related to my study but also a positive and sincere approach towards life in general.

I wish to thank Professor Yukiko Wagatsuma and Assistant Professor Mizuho Fukushige, whose meticulous support was enormously helpful to me.

I wish to thank Mr. Katsuhiko Sawada, Director of the Data Science Department at Taiho Pharmaceutical Co., Ltd., who provided substantial support and guidance.

I thank all my friends who participated in our many discussions, particularly, Dr. Akihiro Hirakawa, Mr. Keisuke Tada, Mr. Ryota Ishii, Ms. Ayako Sato, Mr. Tadashi Hirooka, Dr. Akira Fukushima, Mr. Yasunobu Yamaguchi, Mr. Takahiro Tanase, Mr. Yuji Kunitomi, Mr. Takatoshi Yonekura, and all members of the Gosho and Wagatsuma laboratories.

Finally, I would like to express my gratitude to my family for their moral support and warm encouragement.

March 2019

Bibliography

[1] Baselga J, Cortés J, Kim SB, Im SA, Hegg R, Im YH, Roman L, Pedrini JL, Pienkowski T, Knott A, Clark E, Benyunes MC, Ross G, Swain SM. Pertuzumab plus trastuzumab plus docetaxel for metastatic breast cancer. N Engl J Med 2012;366:109–119.

[2] Bassler D, Briel M, Montori VM, Lane M, Glasziou P, Zhou Q, Heels-Ansdell D, Walter SD, Guyatt GH, STOPIT-2 Study Group AT. Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. J Am Med Assoc 2010;303:1180–1187.

[3] Bassler D, Ferreira-Gonzalez I, Briel M, Cook DJ, Devereaux PJ, Heels-Ansdell D, Kirpalani H, Meade MO, Montori VM, Rozenberg A, Schünemann HJ, Guyatt GH. Systematic reviewers neglect bias that results from trials stopped early for benefit. J Clin Epidemiol 2007;60:869–873.

[4] Bassler D, Montori VM, Briel M, Glasziou P, Guyatt G. Early stopping of randomized clinical trials for overt efficacy is problematic. J Clin Epidemiol 2008;61:241–246.

[5] Bebu I, Luta G, Dragalin V. Likelihood inference for a two-stage design with treatment selection. Biometrical J 2010;52:811–822.

[6] Bowden J, Brannath W, Glimm E. Empirical Bayes estimation of the selected treatment mean for two-stage drop-the-loser trials: a meta-analytic approach. Stat Med 2014;33:388–400.

[7] Bowden J, Glimm E. Conditionally unbiased and near unbiased estimation of the selected treatment mean for multistage drop-the-losers trials. Biometrical J 2014;56:332–349.

[8] Bryant J, Day R. Incorporating toxicity considerations into the design of two-stage phase II clinical trials. Biometrics 1995;51:1372–1383.

[9] Carreras, M, Brannath W. Shrinkage estimation in two-stage adaptive designs with midtrial treatment selection. Stat Med 2013;32:1677–1690.

[10] Chow SC, Chang M. Adaptive design methods in clinical trials. Boca Raton: CRC Press; 2011. 195 p.

[11] Dávalos A, Alvarez-Sabín J, Castillo J, Díez-Tejedor E, Ferro J, Martínez-Vila E, Serena J, Segura T, Cruz VT, Masjuan J. Citicoline in the treatment of acute ischaemic stroke: an international, randomised, multicentre, placebo-controlled study (ICTUS trial). The Lancet 2012;380:349–357.

[12] Emerson SS, Fleming TR. Parameter estimation following group sequential hypothesis testing. Biometrika 1990;77:875–892.

[13] Fan X, DeMets DL. Conditional and unconditional confidence intervals following a group sequential test. J Biopharm Stat 2006;16:107–122.

[14] Gosho M, Sato Y, Nagashima K, Takahashi S. Trends in study design and the statistical methods employed in a leading general medicine journal. J Clin Pharm Ther 2017: DOI: 10.1111/JCPT.12605.

[15] Grossman J, Parmar MK, Spiegelhalter DJ, Freedman LS. A unified method for monitoring and analysing controlled trials. Stat Med 1994;13:1815–1826.

[16] Gsponer T, Gerber F, Bornkamp B, Ohlssen D, Vandemeulebroecke M, Schmidli H. A practical guide to Bayesian group sequential designs. Pharm Stat 2014;13:71–80.

[17] Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. Nat Biotechnol 2014;32:40–51.

[18] Hesketh P, Gralla R, Webb R, Ueno W, DelPrete S, Bachinsky M, Dirlam N, Stack C, Silberman S. Randomized phase II study of the neurokinin 1 receptor antagonist CJ-11,974 in the control of cisplatin-induced emesis. J Clin Oncol 1999;17:338–338.

[19] Hwang JT. Empirical Bayes estimation for the means of the selected populations. Sankhya Ser B 1993;55:285–304.

[20] Jennison C, Turnbull BW. Group sequential methods with applications to clinical trials. Boca Raton: CRC Press; 2000. 49 p.

[21] Kim K, DeMets DL. Confidence intervals following group sequential tests in clinical trials. Biometrics 1987;43:857–864.

[22] Kimani PK, Todd S, Stallard N. Conditionally unbiased estimation in phase II/III clinical trials with early stopping for futility. Stat Med 2013;32:2893–2910.

[23] Koopmeiners JS, Feng Z, Pepe MS. Conditional estimation after a two-stage diagnostic biomarker study that allows early termination for futility. Stat Med 2012;31:420–435.

[24] Lan KG, DeMets DL. Discrete sequential boundaries for clinical trials. Biometrika 1983;70:659–663.

[25] Lindley DV. Discussion of Professor Stein's paper "Confidence sets for the mean of a multivariate normal distribution". J R Stat Soc Series B Stat Methodol 1962;24:265–296.

[26] Liu A, Hall W, Yu KF, Wu C. Estimation following a group sequential test for distributions in the one-parameter exponential family. Stat Sin 2006;16:165.

[27] Liu A, Troendle JF, Yu KF, Yuan VW. Conditional maximum likelihood estimation following a group sequential test. Biometrical J 2004;46:760–768.

[28] Montori VM, Devereaux PJ, Adhikari NKJ, Burns KEA, Eggert CH, Briel M, Lacchetti C, Leung TW, Darling E, Bryant DM, Bucher HC, Schünemann HJ, Meade MO, Cook DJ, Erwin PJ, Sood A, Sood R, Lo B, Thompson CA, Zhou Q, Mills E, Guyatt GH. Randomized trials stopped early for benefit: a systematic review. J Am Med Assoc 2005;294:2203.

[29] O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics 1979;35:549–556.

[30] Ohman–Strickland PA, Casella G. Conditional inference following group sequential testing. Biometrical J 2003;45:515–526.

[31] Oza AM, Pignata S, Poveda A, McCormack M, Clamp A, Schwartz B, Cheng J, Li X, Campbell K, Dodion P, Haluska FG. Randomized phase II trial of ridaforolimus in advanced endometrial carcinoma. J Clin Oncol 2015;33:3576–3582.

[32] Pepe MS, Feng Z, Longton G, Koopmeiners J. Conditional estimation of sensitivity and specificity from a phase 2 biomarker study allowing early termination for futility. Stat Med 2009;28:762–779.

[33] Pinheiro JC, DeMets DL. Estimating and reducing bias in group sequential designs with Gaussian independent increment structure. Biometrika 1997;84:831–845.

[34] Pocock SJ. Group sequential methods in the design and analysis of clinical trials. Biometrika 1977;64:191–199.

[35] Pocock SJ. The combination of randomized and historical controls in clinical trials. J Chronic Dis 1976;29:175–188.

[36] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria 2012. ISBN 3-900051-07-0.

[37] Sato Y, Gosho M, Nagashima K, Takahashi S, Ware JH, Laird NM. Statistical methods in the journal: an update. N Engl J Med 2017;376:1086–1087.

[38] Schaid DJ, Wieand S, Therneau TM. Optimal two-stage screening designs for survival comparisons. Biometrika 1990;77:507–513.

[39] Shimura M, Gosho M, Hirakawa A. Comparison of conditional bias-adjusted estimators for interim analysis in clinical trials with survival data. Stat Med 2017;36:2067–2080.

[40] Shimura M, Maruo K, Gosho M. Conditional estimation using prior information in 2stage group sequential designs assuming asymptotic normality when the trial terminated early. Pharm Stat 2018;17:400–413.

[41] Shimura M, Nomura S, Wakabayashi M, Maruo K, Gosho M. Are bias-adjusted estimators unnecessary to reduce overestimation of treatment effect? : A literature review of oncology clinical trials that stopped early for efficacy. (submitted).

[42] Simon R. Optimal two-stage designs for phase II clinical trials. Control Clin Trials 1989;10:1–10.

[43] Spiegelhalter DJ, Freedman LS, Parmar MK. Bayesian approaches to randomized trials. J R Stat Soc Ser A Stat Soc 1994:357–416.

[44] Swain SM, Baselga J, Kim SB, Ro J, Semiglazov V, Campone M, Ciruelos E, Ferrero JM, Schneeweiss A, Heeson S, Clark E, Ross G, Benyunes MC, Cortés J. Pertuzumab, trastuzumab, and docetaxel in HER2-positive metastatic breast cancer. N Engl J Med 2015;372:724–734.

[45] Thall PF, Simon R, Ellenberg SS. Two-stage selection and testing designs for comparative clinical trials. Biometrika 1988;75:303–310.

[46] Thompson JR. Some shrinkage techniques for estimating the mean. J Am Stat Assoc 1968;63:113–122.

[47] Troendle JF, Yu KF. Conditional estimation following a group sequential clinical trial. Commun Stat Theory Methods 1999;28:1617–1634.

[48] U.S. Department of Health and Human Services, Food and Drug Administration. Guidance for Industry. Adaptive Design Clinical Trials for Drugs and Biologics. Draft Guidance. https://www.fda.gov/downloads/Drugs/.../Guidances/UCM201790.pdf. Accessed September 25, 2017.

[49] Uno H, Claggett B, Tian L, Inoue E, Gallo P, Miyata T, Schrag D, Takeuchi M, Uyama Y, Zhao L, Skali H, Solomon S, Jacobus S, Hughes M, Packer M, Wei LJ. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. J Clin Oncol 2014;32:2380–2385.

[50] Van den Berghe G, Wouters P, Weekers F, Verwaest C, Bruyninckx F, Schetz M, Vlasselaers D, Ferdinande P, Lauwers P, Bouillon R. Intensive insulin therapy in critically ill patients. N Engl J Med 2001;345:1359–1367.

[51] Viele K, McGlothlin A, Broglio K. Interpretation of clinical trials that stopped early. J Am Med Assoc 2016;315:1646–1647.

[52] Wang SJ, Hung H, O'Neill RT. Adapting the sample size planning of a phase III trial based on phase II data. Pharm Stat 2006;5:85–97.

[53] Whitehead J. On the bias of maximum likelihood estimation following a sequential test. Biometrika 1986;73:573–581.

[54] Whitehead J. The design and analysis of sequential clinical trials. Chichester: John Wiley & Sons; 1997. 138 p.

[55] Zhang JJ, Blumenthal GM, He K, Tang S, Cortazar P, Sridhara R. Overestimation of the effect size in group sequential trials. Clin Cancer Res 2012;18:4872–4876.

[56] Zhong H, Prentice RL. Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. Biostatistics 2008;9:621–634.