Statistics in Medicine

WILEY

Statistics in Medicine

RESEARCH ARTICLE OPEN ACCESS

Adjustment of Conditional Bias in Hazard Ratios for Group Sequential Testing of Progression-Free Survival and Overall Survival

Shoki Izumi[1,2] | Shogo Nomura[3] | Yutaka Matsuyama[4]

[1]Department of Biostatistics, Division of Health Sciences and Nursing, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan | [2]Section of Biostatistics, Department of Clinical Data Science, Clinical Research & Education Promotion Division, National Center of Neurology and Psychiatry, Tokyo, Japan | [3]Department of Biostatistics and Bioinformatics, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan | [4]Department of Biostatistics, School of Public Health, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

Correspondence: Shogo Nomura (snomura@m.u-tokyo.ac.jp)

Received: 21 April 2024 | **Revised:** 13 January 2025 | **Accepted:** 21 April 2025

Funding: The authors received no specific funding for this work.

Keywords: conditional estimation | group sequential design | hierarchical procedure

ABSTRACT

In confirmatory randomized controlled trials of patients with metastatic cancer, progression-free survival (PFS) and overall survival (OS) are often used as multiple primary endpoints. The overall hierarchical strategy is a typical multiplicity adjustment method that analyzes PFS once and performs an interim OS analysis at the time of PFS analysis using an alpha-spending function—only if the statistical significance of PFS is demonstrated. A subsequent final OS analysis is conducted if the interim OS analysis does not result in early stopping for efficacy. In this study, we focused on the adjustment of conditional bias (CB) in hazard ratio estimates for OS in both interim and final analyses when a trial applied the overall hierarchical strategy. As CB-adjusting estimators for a single primary endpoint may have limited performance, we extended the conditional mean-adjusted estimator to the case of an overall hierarchical strategy. Motivated by an actual oncology trial, we evaluated the performance of the proposed estimators through a simulation study. In the case of early stopping for efficacy, the CB of the proposed estimator was smaller than that of the existing methods with comparable root mean squared error.

1 | Introduction

In confirmatory randomized controlled trials (RCTs) for patients with metastatic cancer, progression-free survival (PFS) and overall survival (OS) are often used as multiple primary endpoints [1, 2]. Although an increase in OS is the ultimate goal of cancer treatment, PFS is available earlier and captures direct activity on cancer cells. This motivates the current strategy that first tests PFS and applies a group sequential test for OS in a hierarchical manner. One possible approach would be to analyze PFS once and perform an interim OS analysis at the time of PFS analysis using an alpha-spending function, only if the statistical significance of PFS is demonstrated. According to Glimm, Maurer, and Bretz [3], this multiplicity adjustment strategy is called the *overall hierarchical* strategy. The family-wise error rate is strictly controlled in a strong sense due to the closure principle, and the other statistical performance of a more general hierarchical testing strategy has been discussed elsewhere [3].

In this study, we focused on a bias-reduced estimator of the hazard ratio (HR) for OS when a trial applied the overall hierarchical strategy. The FLAURA study [4, 5], a confirmatory RCT for first-line *EGFR* mutated advanced non-small cell lung cancer patients comparing osimertinib (a third-generation epidermal growth factor receptor inhibitor [EGFRi]) versus a first-generation EGFRi, is a typical example. This trial used an overall hierarchical strategy, which resulted in regulatory approval based on a remarkable gain in PFS, together with the statistical significance of OS in the final analysis. The HRs for OS were 0.63 in an interim analysis (not significant) and 0.80 in the final analysis, based on a boundary calculated from the O'Brien-Fleming type alpha spending function. In the label of U.S. approval, only the latter value was reported as an HR for OS [6]. From a statistical viewpoint, the HR for OS in the final analysis is a conditional estimate, given that the trial continues until the final analysis following an overall hierarchical strategy (i.e., PFS significance and OS insignificance in the interim analysis), thus rendering it conditionally biased [7]. Note that if the study stopped due to efficacy in terms of OS, the reported HR for OS in the interim analysis also suffers from the same type of bias conditional on PFS significance. The direction of the conditional bias (CB) requires caution. CB of treatment effect for PFS is in the direction of overestimation (i.e., smaller HR) when there is a significance in PFS. This trend is consistent for interim OS when there is a significant difference in interim OS driven by significance in PFS. The direction of CB in the final OS may go in either direction because the insignificance of interim OS leads to underestimation in the treatment effect (i.e., larger HR) of the final OS. Selected reporting similar to that in the FLAURA study is not rare in actual cancer drug development programs [8–11], and this may have a greater impact on pricing and reimbursement decisions after drug approval in some countries. Additionally, bias adjustment is also an important issue in adaptive trial designs to produce reliable treatment effect estimates, as discussed in Robertson et al., who reviewed bias adjustment in adaptive designs, including group sequential, sample size re-estimation, multi-arm multi-stage, response-adaptive randomization, and adaptive enrichment designs [12, 13].

Statistical methodologies for adjusting for the above-mentioned CB in HR have been proposed [14–20]; however, the underlying situation is limited to RCTs with a single primary endpoint. This study proposes a multi-step version of the conditional mean-adjusted estimator (CMAE), which is an existing CB-adjusted estimator. We then aimed to extend it to situations in which an overall hierarchical strategy was used. After reviewing the existing CB-adjusted estimators for group sequential design in Section 2, the motivating situation and our proposal are detailed in Section 3. Simulation studies follow in Section 4, which evaluate the performance of an ordinally reported maximum likelihood estimator (MLE), existing CB-adjusted estimators, and our proposed method. An additional simulation study evaluates the proposed estimator with the mis-specified parameter in Section 5. An illustration is shown in Section 6. The discussion is provided in Section 7.

2 | Existing CB-Adjusted Estimators

2.1 | Motivating Trials and Notations

To provide a brief overview of the existing CB-adjusted estimators for a single primary endpoint, we first motivate an RCT that sets the OS as the primary endpoint. The trial plans only one interim analysis to consider stop for efficacy. $\widehat{\theta}_l$ is denoted as an MLE of log HR at the l-th analysis ($l = 1$: interim analysis, $l = 2$: final analysis) using a Cox regression model. HR is defined as the ratio of the hazard function of the control group to that of the experimental group. If the null hypothesis $H_0 : \theta \leq 0$ is rejected against the alternative hypothesis $H_1 : \theta > 0$, the experimental group is declared to be significantly superior to the control group. For a sufficiently large sample size, $\widehat{\theta}_l$ asymptotically follows a bivariate normal distribution: [21]

$$\begin{pmatrix} \widehat{\theta}_1 \\ \widehat{\theta}_2 \end{pmatrix} \sim MVN\left(\begin{pmatrix} \theta \\ \theta \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_2^2 \\ \sigma_2^2 & \sigma_2^2 \end{pmatrix} \right)$$

where σ_l^2 is the inverse of the observed information I_l obtained from partial likelihood for θ at the l-th analysis. Let $Z_l = \widehat{\theta}_l \sqrt{I_l}$ be the standardized test statistic and an interval $\left(b_l \sqrt{I_l}, \infty \right)$ be the rejection region; thus, $R_l = \left(b_l, \infty \right)$ is the rejection region for $\widehat{\theta}_l$. Let M represent the stopping stage

$$M = \begin{cases} 1, \text{if } \widehat{\theta}_1 \in R_1 \\ 2, \text{otherwise} \end{cases}$$

and let m be its realized value. The null hypothesis H_0 is rejected at the l-th analysis if $\widehat{\theta}_l \in R_l$.

For n-dimensional vectors $\boldsymbol{a} = \left(a_1, \ldots, a_n \right)^T$, $\boldsymbol{b} = \left(b_1, \ldots, b_n \right)^T$, and $\boldsymbol{x} = \left(x_1, \ldots, x_n \right)^T$, we abbreviate multiple integrals over multiple intervals $\left[a_1, b_1 \right], \ldots, \left[a_n, b_n \right]$ as follows:

$$\int_{\boldsymbol{a}}^{\boldsymbol{b}} f(\boldsymbol{x}) d\boldsymbol{x} \coloneqq \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} f\left(x_1, \ldots, x_n \right) dx_n \ldots dx_1$$

Let $\phi_n(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\Sigma})$ be a probability density function of the multivariate normal distribution with expectation $\boldsymbol{\theta} \in \mathbb{R}^n$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$, and let $L_n(\boldsymbol{a}, \boldsymbol{b}; \boldsymbol{\theta}, \boldsymbol{\Sigma})$ be the multiple integral of $\phi_n(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\Sigma})$ over the multiple intervals $\left[a_1, b_1 \right], \ldots, \left[a_n, b_n \right]$ given by

$$\phi_n(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\Sigma}) \coloneqq \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\theta})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\theta}) \right)$$

$$L_n(\boldsymbol{a}, \boldsymbol{b}; \boldsymbol{\theta}, \boldsymbol{\Sigma}) \coloneqq \int_{\boldsymbol{a}}^{\boldsymbol{b}} \phi_n(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\Sigma}) d\boldsymbol{x}$$

For univariate normal distribution, the density function and the cumulative distribution function are denoted as $\phi(x) = \phi_1(x; 0, 1)$ and $\Phi(x) = L_1(-\infty, x; 0, 1)$, respectively.

2.2 | CB-Adjusted Estimators

2.2.1 | CMAE

The CB of the MLE $\hat{\theta}_m$ can be written as the difference between the conditional expectation of $\hat{\theta}_m$, which can be written by the expectation of truncated normal distribution, and the true treatment effect θ. For $m = 1$ and $m = 2$, respectively, the CB is represented by the following formulas: [22]

$$E\left(\hat{\theta}_1 | M = 1\right) - \theta = \sigma_1 \frac{\phi\left(\frac{\theta - b_1}{\sigma_1}\right)}{\Phi\left(\frac{\theta - b_1}{\sigma_1}\right)} \tag{1}$$

$$E\left(\hat{\theta}_2 | M = 2\right) - \theta = -\left(\frac{\sigma_2^2}{\sigma_1}\right) \frac{\phi\left(\frac{b_1 - \theta}{\sigma_1}\right)}{\Phi\left(\frac{b_1 - \theta}{\sigma_1}\right)}$$

Although an unbiased estimator can be obtained if we know the true value of θ, the value of θ is unknown. Thus, one type of CB-adjusted estimator is the CMAE proposed by Troendle and Yu [15]. The CMAE, which we refer to as $\hat{\theta}_m^{CMAE}$, is rooted in the mean-adjusted estimator of Whitehead [23], considering unconditional bias correction. For any value of m (= 1, 2), CMAE is defined as the solution of the formula describing that $E\left(\hat{\theta}_M | M = m\right)$ is equal to the observed value of $\hat{\theta}_m$. Here, the mean parameter θ is substituted by an unknown quality $\tilde{\theta}$. $\hat{\theta}_1^{CMAE}$ is the solution to the following equation (in $\tilde{\theta}$):

$$\tilde{\theta} + \sigma_1 \frac{\phi\left(\frac{\tilde{\theta} - b_1}{\sigma_1}\right)}{\Phi\left(\frac{\tilde{\theta} - b_1}{\sigma_1}\right)} - \hat{\theta}_1 = 0 \tag{2}$$

Similarly, $\hat{\theta}_2^{CMAE}$ is the solution to

$$\tilde{\theta} - \left(\frac{\sigma_2^2}{\sigma_1}\right) \frac{\phi\left(\frac{b_1 - \tilde{\theta}}{\sigma_1}\right)}{\Phi\left(\frac{b_1 - \tilde{\theta}}{\sigma_1}\right)} - \hat{\theta}_2 = 0$$

According to previous simulation studies [14–16, 18, 22, 24, 25], the CMAE tended to have a smaller CB than the MLE when a two-sided group sequential test was used. This trend was also observed when $M = 2$, regardless of the sidedness of the group sequential test. Conversely, when a one-sided group sequential test is adopted and $M = 1$, the CMAE may overcorrect the CB [20, 25]. That is, $\hat{\theta}_1^{CMAE}$ may have the CB in the opposite direction compared to the CB of $\hat{\theta}_1$: $E\left(\hat{\theta}_1 - \theta | M = 1\right) > 0 > E\left(\hat{\theta}_1^{CMAE} - \theta | M = 1\right)$. In some cases, the degree of overcorrection can be substantial. This is probably the reason why the simulation study by Shimura et al. [22] used an estimator, $\hat{\theta}_1^{CMAE,simple}$, given by

$$\hat{\theta}_1^{CMAE,simple} := \hat{\theta}_1 - \sigma_1 \frac{\phi\left(\frac{\hat{\theta}_1 - b_1}{\sigma_1}\right)}{\Phi\left(\frac{\hat{\theta}_1 - b_1}{\sigma_1}\right)}$$

This idea is similar to the simple estimator of Guo and Liu [26], which corrects the unconditional bias for binary outcomes by estimating the unconditional bias through the substitution of the MLE for the parameter of interest. The previous simulation studies [19, 22] suggested that, although $\hat{\theta}_1^{CMAE,simple}$ tends not to overcorrect CB, $\hat{\theta}_1^{CMAE,simple}$ corrects CB insufficiently when $\hat{\theta}_1$ has large CB. Such situations include an early stop with the small true effect, or with a small information fraction when the O'Brien–Flemming type alpha spending function is used [19, 22]. Theoretically, $\hat{\theta}_1^{CMAE,simple}$ was shown to asymptotically leave positive CB (Supporting Information File S.1). An intuitive explanation is that because CB is a strictly monotonically decreasing function with θ, substituting the overestimated MLE for the true parameter θ might lead to underestimation of CB, resulting in a remaining CB.

2.2.2 | Penalized MLE

For a trial stopped early for benefit, Marschner et al. [20] proposed a penalized MLE (pMLE), whose support does not include values implying opposite directions. One may be concerned with the negative value of the CB-adjusted estimates (CMAE and other estimators detailed in Supporting Information File S.9), even though an early stop due to efficacy is demonstrated, that is, $M = 1$. To address this problem, the tuning parameter $\lambda \in [0, 1]$ is introduced to the estimating process of conditional MLE, which can be transformed to CMAE [27]. Conditional MLE with λ can be deformed to the modified version of Equation (2) of CMAE:

$$\tilde{\theta} + \lambda\sigma_1 \frac{\phi\left(\frac{\tilde{\theta} - b_1}{\sigma_1}\right)}{\Phi\left(\frac{\tilde{\theta} - b_1}{\sigma_1}\right)} - \hat{\theta}_1 = 0 \tag{3}$$

Let the solution of (3), which is a class of CB-adjusted estimator with parameter λ, be represented by $P\left(\lambda, \hat{\theta}_1\right)$. $P\left(0, \hat{\theta}_1\right)$ is the MLE and $P\left(1, \hat{\theta}_1\right)$ is the CMAE. Additionally, $P\left(\lambda, \hat{\theta}_1\right)$ strictly monotonically decreases with λ and strictly monotonically increases with $\hat{\theta}_1$ (Supporting Information File S.2); that is, a larger value of λ indicates stronger correction. Then, λ^* is chosen to satisfy $P\left(\lambda^*, b_1\right) = 0$; as a result, $P\left(\lambda^*, \hat{\theta}_1\right) > 0$ holds. $P\left(\lambda^*, \hat{\theta}_1\right)$ is the pMLE, denoted by $\hat{\theta}_1^{pMLE}$, and its value is between MLE and CMAE.

Other CB-adjusted estimators are summarized in Supporting Information File S.9.

3 | Adjustment of CB in RCTs Applying an Overall Hierarchical Strategy

3.1 | Design and Notation

We considered an RCT (allocation ratio 1:1) setting PFS and OS as the multiple primary endpoints. The trial planned to perform an interim OS analysis at the same time as the main PFS analysis. No interim analyses were planned for PFS. Interim OS analysis will be performed if the PFS is statistically significant. Similarly, a final OS analysis will be performed if the interim OS analysis recommends a continuation. This overall hierarchical strategy never inflates the type I error rate, owing to the closure testing principle [3].

Let k represent the primary endpoints ($k = 1$: PFS, $k = 2$: OS) and l represent the analysis stage ($l = 1$: main PFS analysis or interim OS analysis, $l = 2$: final OS analysis). θ_{kl} denotes log HR comparing the control group with the experimental group. We also assumed that the true value of θ_{kl} remained unchanged across the analysis stages (i.e., proportionality of hazards assumption), and we denoted the true value as θ_k and $H_{k0} : \theta_k \leq 0$ as the null hypothesis for endpoint k. We consider applying a Cox regression model for estimating θ_{kl}, and let the maximum partial likelihood estimate be $\hat{\theta}_{kl}$. With a sufficiently large sample size, $\left(\hat{\theta}_{11}, \hat{\theta}_{21}, \hat{\theta}_{22}\right)$ follow the trivariate normal distribution,

$$\begin{pmatrix} \hat{\theta}_{11} \\ \hat{\theta}_{21} \\ \hat{\theta}_{22} \end{pmatrix} \sim MVN\left(\begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_2 \end{pmatrix}, \boldsymbol{\Sigma}\right)$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11}^2 & Cov\left[\hat{\theta}_{11}, \hat{\theta}_{21}\right] & Cov\left[\hat{\theta}_{11}, \hat{\theta}_{22}\right] \\ Cov\left[\hat{\theta}_{11}, \hat{\theta}_{21}\right] & \sigma_{21}^2 & \sigma_{22}^2 \\ Cov\left[\hat{\theta}_{11}, \hat{\theta}_{22}\right] & \sigma_{22}^2 & \sigma_{22}^2 \end{pmatrix}$$

where σ_{kl}^2 is the inverse of the observed information I_{kl} obtained from partial likelihood for θ_k at the l-th analysis. According to Gou and Xi [28], we assumed that $Cov\left[\hat{\theta}_{11}, \hat{\theta}_{2l}\right] \coloneqq \rho \cdot \min\left(\sigma_{11}, \sigma_{2l}\right)^2$, where ρ is defined as $Corr\left[\hat{\theta}_{12}, \hat{\theta}_{22}\right]$, which equals the Pearson correlation coefficient of PFS and OS given several assumptions (see Section S.4.2 in Supporting Information File), which is hard to verify in actual RCTs. Let $R_{kl} = \left(b_{kl}, \infty\right)$ be the rejection region for $\hat{\theta}_{kl}$; stopping stage M is defined as:

$$M = \begin{cases} 1, \text{if } \hat{\theta}_{11} \in R_{11}, \hat{\theta}_{21} \in R_{21} \\ 2, \text{if } \hat{\theta}_{11} \in R_{11}, \hat{\theta}_{21} \notin R_{21} \end{cases}$$

In Section 1, we illustrate a motivating example of the FLAURA study. The labeling itself can be driven by the final OS significance. When one concerns a conditional bias due to the significance of final OS testing, $\hat{\theta}_{22} \in R_{22}$ should be added for $\hat{\theta}_{22}$. CB correction in this case is detailed in Section S.5 in the Supporting Information File.

3.2 | Extension of CMAE

Considering $M = m$, let $m^* = I(k = 1) + I(k = 2)m$, where $I\left(k = k'\right)$ is the indicator function taking 1 when $k = k'$ and 0 when $k \neq k'$ (m^* keeps 1 irrespective of m when $k = 1$). Let $B\left(k, m, \theta_1, \theta_2\right)$ represent CB of $\hat{\theta}_{km^*}$ when stopping stage $M = m$:

$$B\left(k, m, \theta_1, \theta_2\right) \coloneqq \frac{\int_{\boldsymbol{a}}^{\boldsymbol{b}} x_{I(k=1)+I(k=2)(m+1)} \phi_3\left(\boldsymbol{x}; \left(\theta_1, \theta_2, \theta_2\right)^T, \boldsymbol{\Sigma}\right) \mathrm{d}\boldsymbol{x}}{L_3\left(\boldsymbol{a}, \boldsymbol{b}; \left(\theta_1, \theta_2, \theta_2\right)^T, \boldsymbol{\Sigma}\right)} - \theta_k$$

where $\boldsymbol{a} = \left(b_{11}, b_{21}, -\infty\right)^T$ and $\boldsymbol{b} = (\infty, \infty, \infty)^T$ for $m = 1$ and $\boldsymbol{a} = \left(b_{11}, -\infty, -\infty\right)^T$ and $\boldsymbol{b} = \left(\infty, b_{21}, \infty\right)^T$ for $m = 2$. This is in the form of subtracting the parameter θ_k from the first term of the expected value of the truncated normal distribution. Based on this expression, the extension of $\hat{\theta}_m^{CMAE}$ may be defined as the solution to the following simultaneous equations of $\left(\tilde{\theta}_1, \tilde{\theta}_2\right)$:

$$\begin{cases} \tilde{\theta}_2 + B\left(2, m, \tilde{\theta}_1, \tilde{\theta}_2\right) - \hat{\theta}_{2m} = 0 \\ \tilde{\theta}_1 + B\left(1, m, \tilde{\theta}_1, \tilde{\theta}_2\right) - \hat{\theta}_{11} = 0 \end{cases} \tag{4}$$

The solution of $\tilde{\theta}_2$ can be regarded as a CB-adjusted estimator of OS log HR, and we denote it by $\hat{\theta}_{2m}^{MCMAE}$. Note that a solution for $\tilde{\theta}_2$ requires a post hoc adjustment of $\tilde{\theta}_1$. For most trials following an overall hierarchical strategy, an intermediate report, including the PFS estimate, $\hat{\theta}_{11}$, may be submitted to medical conferences and/or journals as well as new drug applications for regulatory agencies regardless of interim OS estimates, $\hat{\theta}_{21}$. Nevertheless, acceptance of such submission can be highly dependent on either or both OS interim and/or final analysis results simply because the oncology community always focuses on OS. These considerations imply that, in medical reports or labels, there would be no doubts about the existence of conditional bias $\hat{\theta}_{11}$ driven by $\hat{\theta}_{2m}$.

Similar to $\hat{\theta}_1^{CMAE}$, $\hat{\theta}_{2m}^{MCMAE}$ would become an overcorrected CB-adjusted estimator (shown by the simulation study in Section 4). Thus, we next considered the extension of $\hat{\theta}_1^{CMAE,simple}$, substituting MLE $\left(\hat{\theta}_{11}, \hat{\theta}_{2m}\right)$ for $\left(\theta_1, \theta_2\right)$:

$$\hat{\theta}_{2m}^{MCMAE,simple} \coloneqq \hat{\theta}_{2m} - B\left(2, m, \hat{\theta}_{11}, \hat{\theta}_{2m}\right)$$

Similar to the $\hat{\theta}_1^{CMAE,simple}$, this estimator corrected CB insufficiently when $\hat{\theta}_{2m}$ has a large CB (Section 4). Additionally, $\hat{\theta}_{21}^{MCMAE,simple}$ under $\rho = 0$ was shown to be equivalent to $\hat{\theta}_1^{CMAE,simple}$ (Supporting Information File S.6).

3.2.1 | Iteration of Correcting CB

In order to balance under-correction of $\hat{\theta}_{2m}^{MCMAE,simple}$ and over-correction of $\hat{\theta}_{2m}^{MCMAE}$, we propose an iterated correction method driven by the idea of repeated correction based on $\hat{\theta}_1^{CMAE,simple}$. $\hat{\theta}_1^{CMAE,simple}$ is the estimator substituting $\hat{\theta}_1$ for θ to approximate CB (1). Given the results that $\hat{\theta}_1^{CMAE,simple}$ estimates θ better than $\hat{\theta}_1$, as in some scenarios in the previous simulation study [19, 22], substituting $\hat{\theta}_1^{CMAE,simple}$ for θ might lead to a better approximation of CB (1). This method is explained in Section "Repeated Correction for Single Primary Endpoint" of the Appendix A. Let $\tau \in \mathbb{N}_{\geq 1}$ be the number of repetition and $\hat{\theta}_1^{CMAE(\tau)}$ be the CB-adjusted estimator at the τ-th repetition. It was found that a large τ leads to a strong bias correction. Thus, a number of repetitions would exist that correct CB with some moderate degree between the overcorrected $\hat{\theta}_1^{CMAE}$ and insufficiently corrected $\hat{\theta}_1^{CMAE,simple}$ (or $\hat{\theta}_1$).

Given the above property of the iterated adjustment method for single-endpoint situations, we next consider an extension to multiple primary endpoint cases, following the notation in Section 3.1. Let $\hat{\theta}_{2m}^{MCMAE(\tau)}$ be the τ-th CB-adjusted estimator. It is obtained by the following steps until τ reaches the upper limit, $\tau_{\max} \in \mathbb{N}_{\geq 1}$.

Step 1. Set $\tau = 1$.

Step 2. Calculate $B\left(2, m, \hat{\theta}_{11}^{MCMAE(\tau-1)}, \hat{\theta}_{2m}^{MCMAE(\tau-1)}\right)$ and $B\left(1, m, \hat{\theta}_{11}^{MCMAE(\tau-1)}, \hat{\theta}_{2m}^{MCMAE(\tau-1)}\right)$, where the initial values are set as $\left(\hat{\theta}_{11}^{MCMAE(0)}, \hat{\theta}_{2m}^{MCMAE(0)}\right) = \left(\hat{\theta}_{11}, \hat{\theta}_{2m}\right)$.

Step 3. Update the CB-corrected estimators at the τ-th iteration:

$$\hat{\theta}_{2m}^{MCMAE(\tau)} := \hat{\theta}_{2m} - B\left(2, m, \hat{\theta}_{11}^{MCMAE(\tau-1)}, \hat{\theta}_{2m}^{MCMAE(\tau-1)}\right)$$

$$\hat{\theta}_{11}^{MCMAE(\tau)} := \hat{\theta}_{11} - B\left(1, m, \hat{\theta}_{11}^{MCMAE(\tau-1)}, \hat{\theta}_{2m}^{MCMAE(\tau-1)}\right)$$

Step 4. Increment τ by 1, and if τ is greater than $\tau_{\max}$, stop the iteration steps. Otherwise, repeat Step 2.

The integral computation in Step 2 uses the method proposed by Kan and Robotti [29]. Here, $\hat{\theta}_{2m}^{MCMAE(1)}$ is identical to $\hat{\theta}_{2m}^{MCMAE,simple}$. Within the simulation study we conducted, $\hat{\theta}_{21}^{MCMAE(\tau_{\max})}$ with a large value of τ_{max} was close to $\hat{\theta}_{21}^{MCMAE}$. As illustrated in Section 4, the performance of $\hat{\theta}_{21}^{MCMAE(100)}$ was similar to $\hat{\theta}_{21}^{MCMAE}$. Here, $\hat{\theta}_{21}^{MCMAE(\tau)}$ was shown not to amplify the overestimation of MLE because $B(k, 1, \theta_1, \theta_2) > 0$ holds for $k = 1, 2$ under $\rho \geq 0$ (Supporting Information File S.7).

3.2.2 | Determination of $\tau_{\max}$

For $M = 1$, as $\tau_{\max}$ becomes larger, $\hat{\theta}_{21}^{MCMAE(\tau_{\max})}$ is observed to have smaller values, which indicates that $\hat{\theta}_{21}^{MCMAE(\tau_{\max})}$ is a monotone decreasing function of $\tau_{\max}$. This examination implies that our proposed iterative estimator $\hat{\theta}_{21}^{MCMAE(\tau_{\max})}$ with a large value of τ_{max} may result in an excessive adjustment, which means that $\hat{\theta}_{21}^{MCMAE(\tau_{\max})}$ can be negative even though the trial rejects H_{k0}.

This can be avoided by incorporating the idea of pMLE into $\hat{\theta}_{21}^{MCMAE(\tau_{\max})}$. The resulting estimator, $\hat{\theta}_{21}^{MCMAE(\tau_{\max}^*, w^*)}$, uses two additional parameters $(\tau_{\max}^*, w^*)$, as detailed in Section "Details of $\hat{\theta}_{21}^{MCMAE(\tau_{\max}^*, w^*)}$" in the Appendix A. τ_{max}^* was set as the value of $\tau_{\max}$ that corrects CB to the degree that it does not contradict the results of hypothesis testing. w^* was introduced to enable $\hat{\theta}_{21}^{MCMAE(\tau_{\max}^*, w^*)}$ to take the value between $\hat{\theta}_{21}^{MCMAE(\tau_{\max}^*)}$ and $\hat{\theta}_{21}^{MCMAE(\tau_{\max}^*+1)}$. Thereby, w^* set the infimum of $\hat{\theta}_{21}^{MCMAE(\tau_{\max}^*, w^*)}$ as zero, which is the infimum of the values implying the superiority of the experimental group.

4 | Simulation Study 1

The performance of the proposed method was assessed through simulation studies motivated by a trial that applies an overall hierarchical procedure for multiple primary endpoints (PFS and OS). The trial design was the same as that described in Section 3.1.

The existing methods detailed in Section 2 (CMAE, simplified version of CMAE, and pMLE) and the MLE were used for OS estimators as benchmark, although they disregarded the impact of hierarchical testing on PFS and OS. Note that $\hat{\theta}_m^{CMAE}$ was calculated for $M = 1, 2$ (Sections 4.2.1–4.2.2), and $\hat{\theta}_1^{CMAE,simple}$ and $\hat{\theta}_1^{pMLE}$ were calculated only for $M = 1$ (Section 4.2.1). We also used other existing CB-adjusted estimators for a single primary outcome (Section S.9 in the Supporting Information File).

For the proposed method, we evaluated the proposed estimator $\hat{\theta}_{21}^{MCMAE(\tau_{\max}^*, w^*)}$ for $M = 1$ (Section 4.2.1), $\hat{\theta}_{2m}^{MCMAE(\tau_{max})}$ with different values of τ_{max}(= 1 or 100) for $M = 1, 2$ (Sections 4.2.1–4.2.2), and $\tau_{max} = 5$ for $M = 2$ (Section 4.2.2).

The R-code for the simulation is available upon request from the authors.

4.1 | Settings

We used four evaluation metrics in the scale of HR (i.e., scale of $\exp(\theta)$), of which a large value represents a large effect of the experimental group. The metrics included CB, percentage CB (%CB), empirical standard error (EmpSE), and root mean squared error (RMSE), which were calculated as follows:

$$\text{CB} = \hat{E}\left[\exp\left(\hat{\theta}_{2M}\right) | M = m\right] - \exp(\theta_2),$$

$$\%\text{CB} = \frac{CB}{\exp(\theta_2)},$$

$$\text{EmpSE} = \sqrt{\frac{n_{sim}}{n_{sim}-1}\hat{E}\left[\left(\exp\left(\hat{\theta}_{2M}\right) - \hat{E}\left[\exp\left(\hat{\theta}_{2M}\right) | M = m\right]\right)^2 | M = m\right]},$$

and

$$\text{RMSE} = \sqrt{\hat{E}\left[\left(\exp\left(\hat{\theta}_{2M}\right) - \exp(\theta_2)\right)^2 | M = m\right]}.$$

The one-sided significance level was set at 0.025. The Fisher information was approximated by the total number of events divided by four [30]. For the test of PFS, we set the number of events with an expected HR $\exp(-\theta_{10})$ of 0.65 and a power of 0.9, resulting in the total required number of events being 227. The O'Brien-Fleming type or Pocock type alpha spending function [31] was applied to compute the stopping boundaries for OS analysis. The number of events for the interim and final OS analyses was set to achieve a power of 0.8 with an expected HR $\exp(-\theta_{20})$ of 0.7. The timing of the OS interim analysis was an information fraction (IF) of 0.25, 0.5, or 0.75. The true parameters ρ, θ_1, and θ_2 were chosen from $\{0, 0.5, 0.9\}$ for ρ, $\{0.85, 0.6\}$ for $\exp(-\theta_1)$, and $\{0.9, 0.65\}$ for $\exp(-\theta_2)$. Under these settings, $\left(\hat{\theta}_{11}, \hat{\theta}_{21}, \hat{\theta}_{22}\right)$ were generated from a multivariate normal distribution. In each of the 72 scenarios, we obtained 3000 estimates satisfying the condition $M = 1$ and another 3000 estimates satisfying $M = 2$.

4.2 | Results

4.2.1 | Scenario With Early Termination for Benefit

Table 1 shows the results for $M = 1$ with Pocock and O'Brien-Fleming type alpha spending functions, IF of 0.5, HR $\exp(-\theta_1)$ of 0.85 for PFS, and HR $\exp(-\theta_2)$ of 0.65 for OS. We first consider the independent scenarios where true $\rho = 0$.

TABLE 1 | Results of simulation study 1 in $M = 1$, IF = 0.5, $\exp(-\theta_1) = 0.85$, and $\exp(-\theta_2) = 0.65$.

Alpha spending function	**Pocock**			**O'Brien-Fleming**		
ρ	**0**	**0.5**	**0.9**	**0**	**0.5**	**0.9**
$P\left(\widehat{\theta}_{11} \in R_{11}\right)$	0.231	0.231	0.231	0.231	0.231	0.231
$P\left(\widehat{\theta}_{11} \in R_{11}, \widehat{\theta}_{21} \in R_{21}\right)$	0.15	0.192	0.221	0.066	0.108	0.149
$\widehat{\theta}_{21}$	0.17 (11%)/0.2/0.26	0.24 (16%)/0.23/0.34	0.3 (19%)/0.23/0.38	0.38 (24%)/0.18/0.42	0.41 (26%)/0.2/0.45	0.43 (28%)/0.21/0.48
$\widehat{\theta}_1^{CMAE,simple}$	0.072 (5%)/0.25/0.26	0.16 (10%)/0.28/0.33	0.23 (15%)/0.28/0.36	0.23 (15%)/0.24/0.33	0.26 (17%)/0.26/0.37	0.3 (19%)/0.27/0.4
$\widehat{\theta}_1^{CMAE}$	−0.17 (−11%)/0.54/0.56	−0.0076 (−0%)/0.51/0.51	0.13 (9%)/0.44/0.46	−0.25 (−16%)/0.66/0.7	−0.15 (−10%)/0.66/0.67	−0.056 (−4%)/0.63/0.64
$\widehat{\theta}_1^{pMLE}$	−0.0023 (−0%)/0.32/0.32	0.11 (7%)/0.34/0.36	0.2 (13%)/0.32/0.38	0.025 (2%)/0.38/0.38	0.083 (5%)/0.4/0.41	0.14 (9%)/0.4/0.42
$\widehat{\theta}_{21}^{MCMAE,simple}$	0.072 (5%)/0.25/0.26	0.13 (8%)/0.27/0.3	0.17 (11%)/0.26/0.31	0.23 (15%)/0.24/0.33	0.24 (15%)/0.25/0.34	0.26 (17%)/0.26/0.37
$\widehat{\theta}_{21}^{MCMAE(100)}$	−0.12 (−8%)/0.5/0.52	−0.22 (−14%)/0.48/0.53	−0.25 (−16%)/0.55/0.6	−0.16 (−10%)/0.62/0.64	−0.3 (−19%)/0.56/0.63	−0.31 (−20%)/0.62/0.69
$\widehat{\theta}_{21}^{MCMAE}$	−0.16 (−10%)/0.5/0.52	−0.23 (−15%)/0.5/0.55	−0.27 (−17%)/0.57/0.62	−0.21 (−14%)/0.59/0.63	−0.31 (−20%)/0.57/0.65	−0.32 (−21%)/0.64/0.71
$\widehat{\theta}_{21}^{MCMAE(\tau^*_{max}, w^*)}$	−0.023 (−1%)/0.33/0.33	0.021 (1%)/0.32/0.32	0.055 (4%)/0.3/0.31	−0.0065 (−0%)/0.38/0.38	−0.0076 (−0%)/0.35/0.35	0.022 (1%)/0.36/0.36
$\left(\tau^*_{max}, w^*\right)$	(3, 0.84)	(3, 0.22)	(3, 0.05)	(6, 0.26)	(5, 0.16)	(4, 0.81)

Note: Rows represent the estimators, each column represents the alpha spending functions and the true correlation ρ, and each cell contains CB(%CB)/EmpSE/RMSE. $P\left(\widehat{\theta}_{11} \in R_{11}\right)$ = probability of significance of test for PFS, $P\left(\widehat{\theta}_{11} \in R_{11}, \widehat{\theta}_{21} \in R_{21}\right)$ = probability of $M = 1$, $\widehat{\theta}_{21}$ = MLE for OS (Section 2.1), $\widehat{\theta}_1^{CMAE,simple}$ = simplified version of CMAE(Section 2.2.1), $\widehat{\theta}_1^{CMAE}$ = conditional mean-adjusted estimator (CMAE) (Section 2.2.1), $\widehat{\theta}_1^{pMLE}$ = penalized MLE (Section 2.2.2), $\widehat{\theta}_{21}^{MCMAE,simple}$ = extension of CMAE with iteration number of one (Section 3.2), $\widehat{\theta}_{21}^{MCMAE(100)}$ = extension of CMAE with iteration number of 100 (Section 3.2.1), $\widehat{\theta}_{21}^{MCMAE}$ = extension of CMAE by Equation (4) (Section 3.2), $\widehat{\theta}_{21}^{MCMAE(\tau^*_{max}, w^*)}$ = extension of CMAE with iteration number of τ^*_{max} and weight w^* determined by the similar way to the pMLE (Section 3.2.2).

In this scenario, the significance of PFS does not introduce the CB of MLE ($\widehat{\theta}_{21}$) for OS; thus, only the group sequential test for OS caused a positive CB of the MLE. As shown in the second and fifth columns in Table 1, the %CB of $\widehat{\theta}_1^{CMAE,simple}$ was about half of MLE ($\widehat{\theta}_{21}$). For CMAE, MCMAE ($\widehat{\theta}_{21}^{MCMAE}$), and the repeated correction version of MCMAE ($\widehat{\theta}_{21}^{MCMAE(100)}$), there was a trend of overcorrection, together with larger EmpSE and RMSE. Compared to these and other estimators, the %CB of pMLE was far smaller, with a relatively smaller EmpSE and RMSE. These trends were highlighted when a larger selection bias was introduced by a more stringent boundary using the O'Brien-Fleming type alpha spending function (see the second row of Table 1). For both spending functions, our proposed MCMAE $\left(\tau^*_{\max}, w^*\right)$ showed almost the same performance as pMLE.

For correlated scenarios (true $\rho = 0.5$ or 0.9), %CB of MLE ($\widehat{\theta}_{21}$) increased due to the selection bias stemming from the PFS significance. The trend of overcorrection of the CMAE was observed on a case-by-case basis, although this trend remained for the MCMAE and its repeated correction version. Except for CMAE, as ρ got larger, the magnitude of CB increased. Unlike in $\rho = 0$, %CB of pMLE was non-negligible. For MCMAE $\left(\tau^*_{\max}, w^*\right)$, %CB was smaller than pMLE with comparable EmpSE to pMLE and roughly comparable RMSE to $\widehat{\theta}_{21}^{MCMAE,simple}$.

The performance of MCMAE $\left(\tau^*_{\max}, w^*\right)$ depends on which selection was stronger among an earlier stop due to PFS analysis or OS interim analysis. When the former selection was stronger, MCMAE $\left(\tau^*_{\max}, w^*\right)$ performed better. This is simply because a larger ρ introduced a larger CB, which is shown in Table 1. Situations where the latter selection gets stronger include a case where the true treatment effect on OS is small (i.e., $\exp(-\theta_2) = 0.9$), as shown in Table 2. As a whole, ρ did not substantially alter the results in Table 2. Although the %CB of the CMAE was far larger than that in Table 1, a similar trend was observed. These results were also seen in the other scenarios with a large effect on PFS ($\exp(-\theta_1) = 0.6$) (Tables S1 and S2 in Supporting Information File, including additional existing methods listed in Section S10). Nevertheless, the performance of MCMAE $\left(\tau^*_{\max}, w^*\right)$ was generally comparable to pMLE. Tables S3 and S4 present the results, including those of the additional existing methods. Additionally, as the IF decreased, the CB of the MLE based on the OS interim analysis increased (results not shown).

4.2.2 | Scenario With Continuation to the Final Analysis

Table 3 presents the results for $M = 2$ with the same scenario as in Table 1: Pocock and O'Brien-Fleming type alpha spending functions, IF of 0.5, HR $\exp(-\theta_1)$ of 0.85 for PFS, and HR $\exp(-\theta_2)$ of 0.65 for OS. Looking at the row of MLE ($\widehat{\theta}_{22}$), the CB of MLE increased with respect to ρ. That is, the PFS test introduced the CB in the direction of overestimation. Meanwhile, as shown in the independent scenario ($\rho = 0$) of the same row, MLE had negative value of CB. In other words, the group sequential test caused CB in the direction of underestimation. Specifically, for the correlated scenarios ($\rho = 0.5$ or 0.9), existing methods that did not consider PFS analysis tended to have CB toward overestimation, as illustrated in CMAE. Unlike in $M = 1$, MCMAE $\left(\tau_{max}\right)$ did not show monotonic behavior with respect to τ_{max}. $\widehat{\theta}_{22}^{MCMAE(100)}$ could have too large estimates although the possibility of calculation error accumulated through large number of iterations could not be excluded. When using an O'Brien-Fleming type alpha spending function, the CB due to the OS interim analysis was less than when using the Pocock type alpha spending function. The RMSE of the CB-corrected estimates was not smaller than that of the MLE. This was similar to other scenarios (Tables S5, S6, S9–S12, S14, and S16 in the Supporting

TABLE 2 | Results of simulation study 1 in $M = 1$, IF = 0.5, $\exp(-\theta_1) = 0.85$, and $\exp(-\theta_2) = 0.9$.

Alpha spending function	**Pocock**		
ρ	**0**	**0.5**	**0.9**
$P\left(\widehat{\theta}_{11} \in R_{11}\right)$	0.231	0.231	0.231
$P\left(\widehat{\theta}_{11} \in R_{11}, \widehat{\theta}_{21} \in R_{21}\right)$	0.014	0.032	0.05
$\widehat{\theta}_{21}$	0.44 (40%)/0.1/0.45	0.46 (41%)/0.11/0.47	0.45 (41%)/0.11/0.47
$\widehat{\theta}_1^{CMAE,simple}$	0.3 (27%)/0.14/0.33	0.32 (29%)/0.15/0.36	0.32 (28%)/0.15/0.35
$\widehat{\theta}_1^{CMAE}$	−0.23 (−21%)/0.52/0.57	−0.15 (−14%)/0.53/0.55	−0.19 (−17%)/0.53/0.57
$\widehat{\theta}_1^{pMLE}$	0.16 (14%)/0.21/0.27	0.19 (17%)/0.22/0.3	0.18 (16%)/0.22/0.29
$\widehat{\theta}_{21}^{MCMAE,simple}$	0.3 (27%)/0.14/0.33	0.31 (28%)/0.15/0.34	0.3 (27%)/0.14/0.34
$\widehat{\theta}_{21}^{MCMAE(100)}$	−0.13 (−12%)/0.49/0.51	−0.21 (−19%)/0.43/0.48	−0.25 (−23%)/0.48/0.54
$\widehat{\theta}_{21}^{MCMAE}$	−0.18 (−16%)/0.44/0.47	−0.22 (−20%)/0.44/0.49	−0.26 (−24%)/0.49/0.56
$\widehat{\theta}_{21}^{MCMAE\left(\tau^*_{max}, w^*\right)}$	0.14 (12%)/0.21/0.25	0.15 (14%)/0.19/0.25	0.15 (14%)/0.19/0.25
$\left(\tau^*_{max}, w^*\right)$	(3, 0.84)	(3, 0.22)	(3, 0.05)

Note: Rows represent the estimators, each column represents the alpha spending functions and the true correlation ρ, and each cell contains CB(%CB)/EmpSE/RMSE. $P\left(\widehat{\theta}_{11} \in R_{11}\right)$ = probability of significance of test for PFS, $P\left(\widehat{\theta}_{11} \in R_{11}, \widehat{\theta}_{21} \in R_{21}\right)$ = probability of $M = 1$, $\widehat{\theta}_{21}$ = MLE for OS (Section 2.1), $\widehat{\theta}_1^{CMAE,simple}$ = simplified version of CMAE (Section 2.2.1), $\widehat{\theta}_1^{CMAE}$ = conditional mean-adjusted estimator (CMAE) (Section 2.2.1), $\widehat{\theta}_1^{pMLE}$ = penalized MLE (Section 2.2.2), $\widehat{\theta}_{21}^{MCMAE,simple}$ = extension of CMAE with iteration number of one (Section 3.2), $\widehat{\theta}_{21}^{MCMAE(100)}$ = extension of CMAE with iteration number of 100 (Section 3.2.1), $\widehat{\theta}_{21}^{MCMAE}$ = extension of CMAE by Equation (4) (Section 3.2), $\widehat{\theta}_{21}^{MCMAE\left(\tau^*_{max}, w^*\right)}$ = extension of CMAE with iteration number of τ^*_{max} and weight w^* determined by the similar way to the pMLE (Section 3.2.2).

TABLE 3 | Results of simulation study 1 in $M = 2$, IF = 0.5, $\exp(-\theta_1) = 0.85$, and $\exp(-\theta_2) = 0.65$.

Alpha spending function	**Pocock**			**O'Brien-Fleming**		
ρ	**0**	**0.5**	**0.9**	**0**	**0.5**	**0.9**
$P\left(\widehat{\theta}_{11} \in R_{11}\right)$	0.231	0.231	0.231	0.231	0.231	0.231
$P\left(\widehat{\theta}_{11} \in R_{11}, \widehat{\theta}_{21} \notin R_{21}\right)$	0.081	0.039	0.01	0.165	0.123	0.082
$\widehat{\theta}_{22}$	−0.12 (−8%)/0.14/0.18	−0.049 (−3%)/0.13/0.14	0.089 (6%)/0.12/0.15	−0.049 (−3%)/0.17/0.18	0.038 (2%)/0.16/0.16	0.18 (11%)/0.12/0.21
$\widehat{\theta}_2^{CMAE}$	0.032 (2%)/0.24/0.24	0.16 (11%)/0.25/0.3	0.44 (29%)/0.25/0.51	0.044 (3%)/0.25/0.26	0.17 (11%)/0.25/0.3	0.39 (26%)/0.22/0.45
$\widehat{\theta}_{22}^{MCMAE,simple}$	−0.027 (−2%)/0.19/0.19	0.038 (2%)/0.18/0.19	0.13 (8%)/0.16/0.2	0.015 (1%)/0.22/0.22	0.079 (5%)/0.21/0.22	0.18 (12%)/0.17/0.24
$\widehat{\theta}_{22}^{MCMAE(5)}$	0.031 (2%)/0.23/0.24	0.043 (3%)/0.24/0.24	0.09 (6%)/0.24/0.25	0.044 (3%)/0.25/0.25	0.038 (2%)/0.25/0.25	0.097 (6%)/0.26/0.28
$\widehat{\theta}_{22}^{MCMAE(100)}$	0.012 (1%)/0.23/0.23	−0.13 (−8%)/0.64/0.65	6.6×10^8 (4.28×10^{10}%)/3.6×10^{10}/3.6×10^{10}	0.034 (2%)/0.25/0.25	−0.14 (−9%)/0.41/0.43	−0.16 (−10%)/0.55/0.57
$\widehat{\theta}_{22}^{MCMAE}$	0.078 (5%)/0.84/0.84	−0.084 (−5%)/0.31/0.32	−0.19 (−12%)/0.48/0.51	0.048 (3%)/0.25/0.26	−0.11 (−7%)/0.35/0.36	−0.18 (−12%)/0.52/0.55

Note: Rows represent the estimators, each column represents the alpha spending functions and the true correlation ρ, and each cell contains CB(%CB)/EmpSE/RMSE. $P\left(\widehat{\theta}_{11} \in R_{11}\right)$ = probability of significance of test for PFS, $P\left(\widehat{\theta}_{11} \in R_{11}, \widehat{\theta}_{21} \notin R_{21}\right)$ = probability of $M = 2$, $\widehat{\theta}_{22}$ = MLE for OS (Section 2.1), $\widehat{\theta}_2^{CMAE}$ = conditional mean-adjusted estimator (Section 2.2.1), $\widehat{\theta}_{21}^{MCMAE,simple}$ = extension of CMAE with iteration number of one (Section 3.2), $\widehat{\theta}_{21}^{MCMAE(100)}$ = extension of CMAE with iteration number of 100 (Section 3.2.1), $\widehat{\theta}_{21}^{MCMAE(100)}$ = extension of CMAE with iteration number of 100 (Section 3.2.1), $\widehat{\theta}_{22}^{MCMAE}$ = extension of CMAE by Equation (4) (Section 3.2).

Information File, including additional existing methods listed in Section S10). Nevertheless, $\widehat{\theta}_{22}^{MCMAE,simple}$ could have smaller CB and RMSE than MLE and the existing methods when the CB due to OS interim analysis was large, as in the cases with an IF of 0.75 and $\exp(-\theta_2)$ of 0.65 (Tables S13 and S15), or when the CB due to OS interim analysis was small and the CB due to PFS analysis was large, as in the cases with an IF of 0.25 and $\exp(-\theta_1)$ of 0.85 (Tables S7 and S8).

4.3 | High Level Summary of the Simulation Results

For $M = 1$, when the effect of PFS was small, the CB of MLE and existing methods increased with ρ. The proposed method ($\widehat{\theta}_{21}^{MCMAE(\tau^*_{max},w^*)}$) corrected CB due to the significance of PFS and resulted in a smaller CB. For $M = 2$, MLE had the least RMSE in many scenarios, partly because the direction of CB due to the significance of PFS and insignificance of OS interim analysis were opposite and tended to cancel each other. Nevertheless, in some scenario, $\widehat{\theta}_{22}^{MCMAE,simple}$ could have smaller CB and RMSE than those of MLE and the existing methods.

5 | Simulation Study 2

We conducted simulation studies 2A and 2B to evaluate the impact of misspecification of ρ.

5.1 | Simulation Study 2A

Simulation study 2A was conducted under the same situation as simulation study 1. We additionally evaluated $\widehat{\theta}_{21}^{MCMAE(\tau^*_{\max},w^*)}$ for $M = 1$ with ρ specified from $\{0, 0.1, 0.3, 0.5, 0.7, 0.9\}$, irrespective of true ρ. It was denoted by $\widehat{\theta}_{21}^{MCMAE\left(\rho,\tau^*_{\max,\rho},w^*_\rho\right)}$. For $M = 2$, we evaluated $\widehat{\theta}_{21}^{MCMAE,simple}$ with ρ specified from $\{0, 0.1, 0.3, 0.5, 0.7, 0.9\}$, irrespective of true ρ, and it was denoted by $\widehat{\theta}_{22}^{MCMAE,simple(\rho)}$.

5.1.1 | Settings

The settings were the same as in simulation study 1.

5.1.2 | Results

5.1.2.1 | Scenario With Early Termination for Benefit. Table S3, which shows the results of simulation study 1 for $M = 1$, IF = 0.5, $\exp(-\theta_1) = 0.85$, $\exp(-\theta_2) = 0.65$, also includes the results of simulation study 2A in the last six rows. The last six rows in the second column represents CB(%CB)/EmpSE/RMSE of $\widehat{\theta}_{21}^{MCMAE\left(\rho,\tau^*_{\max,\rho},w^*_\rho\right)}$ with ρ specified from $\{0, 0.1, 0.3, 0.5, 0.7, 0.9\}$ with Pocock type alpha spending function under true $\rho = 0$. With wrongly specified $\rho \geq 0.1$, the degree of overcorrection increased. Little overcorrection was observed with specified $\rho = 0.9$ under true $\rho = 0.5$, and no overcorrection under true $\rho = 0.9$. EmpSE decreased with specified ρ across all combination of true ρ, $\exp(-\theta_1)$ and $\exp(-\theta_2)$ (Tables S1–S4).

5.1.2.2 | Scenario With Continuation to the Final Analysis. For $M = 2$, Table S11 includes the results of simulation study 2A. Rows 12–17 represent CB(%CB)/EmpSE/RMSE of $\widehat{\theta}_{22}^{MCMAE,simple(\rho)}$ with ρ specified from $\{0, 0.1, 0.3, 0.5, 0.7, 0.9\}$. Similar to Table S3 for $M = 1$, the degree of overcorrection increased with wrongly specified $\rho \geq 1$, and no overcorrection was observed under true $\rho = 0.9$. EmpSE did not necessarily decrease with specified ρ, probably because $\widehat{\theta}_{22}^{MCMAE,simple(\rho)}$ does not change the fixed number of iteration of one depending specified ρ as $\widehat{\theta}_{21}^{MCMAE\left(\rho,\tau^*_{\max,\rho},w^*_\rho\right)}$ does.

5.2 | Simulation Study 2B

We conducted simulation study 2B to evaluate the impact of misspecification of ρ under wide range of θ_1 and θ_2. For $M = 1$, we evaluated $\widehat{\theta}_{21}^{MCMAE\left(\rho,\tau^*_{\max,\rho},w^*_\rho\right)}$ with ρ specified from $\{0, 0.1, 0.3, 0.5, 0.7, 0.9\}$ irrespective of true ρ, and compared it with $\widehat{\theta}_1^{pMLE}$. For $M = 2$, we evaluated $\widehat{\theta}_{22}^{MCMAE,simple(\rho)}$ with ρ specified from $\{0, 0.1, 0.3, 0.5, 0.7, 0.9\}$ irrespective of true ρ, and compared it with $\widehat{\theta}_2^{CMAE}$. We chose $\widehat{\theta}_2^{CMAE}$ as benchmark because $\widehat{\theta}_2^{CMAE}$ performs well under $M = 2$, unlike $\widehat{\theta}_1^{CMAE}$ under $M = 1$.

5.2.1 | Settings

The three evaluation metrics defined in simulation study 1 were used: absolute CB, EmpSE, and RMSE. The differences in each matric between $\widehat{\theta}_{21}^{MCMAE\left(\rho,\tau^*_{\max,\rho},w^*_\rho\right)}$ and $\widehat{\theta}_1^{pMLE}$ were calculated for $M = 1$, and those between $\widehat{\theta}_{22}^{MCMAE,simple(\rho)}$ and $\widehat{\theta}_2^{CMAE}$ were calculated for $M = 2$. True ρ was chosen from $\{0, 0.5, 0.9\}$. True $\exp(-\theta_1)$ and $\exp(-\theta_2)$ was respectively chosen from 0.5 to 1.0 by 0.05. The Pocock type alpha spending function [31] was applied to compute the stopping boundaries for OS analysis. The timing of the OS interim analysis was an IF of 0.5. The other settings were the same as in those in the simulation study 1.

5.2.2 | Results

5.2.2.1 | Scenario With Early Termination for Benefit. The results of simulation study 2 for $M = 1$ are shown in Figures 1, 2 and S7–S13 in Supporting Information File.

Figure 1 shows the differences in absolute CB between $\widehat{\theta}_{21}^{MCMAE\left(\rho,\tau^*_{\max,\rho},w^*_\rho\right)}$ and $\widehat{\theta}_1^{pMLE}$ for $M = 1$ under true $\rho = 0.5$. Each panel represents ρ for $\widehat{\theta}_{21}^{MCMAE\left(\rho,\tau^*_{\max,\rho},w^*_\rho\right)}$, specified from $\{0, 0.1, 0.3, 0.5, 0.7, 0.9\}$. Each cell in each sub-figures represents the differences in absolute CB. Blue cell represents the absolute CB of $\widehat{\theta}_{21}^{MCMAE\left(\rho,\tau^*_{\max,\rho},w^*_\rho\right)}$ being smaller than that of pMLE. The absolute value of CB of $\widehat{\theta}_{21}^{MCMAE\left(\rho,\tau^*_{\max,\rho},w^*_\rho\right)}$ was nearly equal to or smaller than that of pMLE across all θ_1, θ_2 and specified ρ for $\widehat{\theta}_{21}^{MCMAE\left(\rho,\tau^*_{\max,\rho},w^*_\rho\right)}$.

Figure S7 showed the differences in absolute CB under true $\rho = 0$. When ρ was truly zero, and true $\exp(-\theta_2)$ was smaller than 0.7, then $\widehat{\theta}_{21}^{MCMAE\left(\rho,\tau^*_{\max,\rho},w^*_\rho\right)}$ with positive ρ specified had larger degree

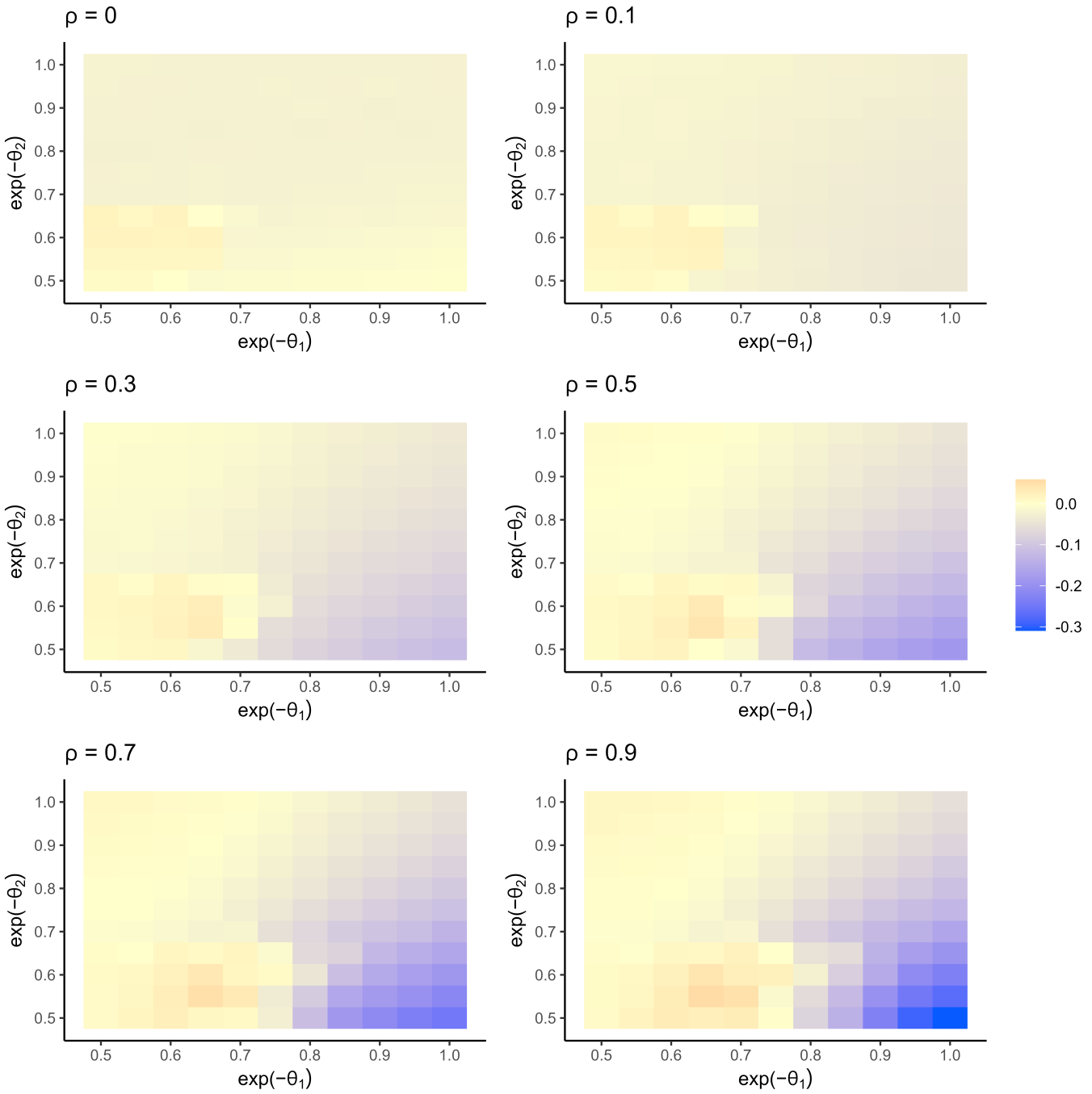

FIGURE 1 | Results of simulation study 2 for differences in absolute CB between $\widehat{\theta}_{21}^{MCMAE\left(\rho,\tau^*_{\max,\rho},w^*_\rho\right)}$ and $\widehat{\theta}_1^{pMLE}$ with $M = 1$ under true $\rho = 0.5$. IF = 0.5 and Pocock type alpha spending function was used. Each panel represents ρ for $\widehat{\theta}_{21}^{MCMAE\left(\rho,\tau^*_{\max,\rho},w^*_\rho\right)}$ specified from $\{0, 0.1, 0.3, 0.5, 0.7, 0.9\}$. Each cell represents the difference and blue cell represents that the absolute CB of $\widehat{\theta}_{21}^{MCMAE\left(\rho,\tau^*_{\max,\rho},w^*_\rho\right)}$ is smaller than that of $\widehat{\theta}_1^{pMLE}$.

of CB than pMLE. The degree of difference tended to increase with true $\exp(-\theta_1)$ and these were because of overcorrection. Nevertheless, in the other scenarios including the case when true $\rho = 0.9$ (Figure S11), the absolute value of CB of $\widehat{\theta}_{21}^{MCMAE\left(\rho,\tau^*_{\max,\rho},w^*_\rho\right)}$ was nearly equal to or smaller than that of pMLE.

EmpSE of $\widehat{\theta}_{21}^{MCMAE\left(\rho,\tau^*_{\max,\rho},w^*_\rho\right)}$ tended to be smaller than that of $\widehat{\theta}_1^{pMLE}$ when true value of ρ was zero or 0.5 (Figures S8 and S10), and RMSE of $\widehat{\theta}_{21}^{MCMAE\left(\rho,\tau^*_{\max,\rho},w^*_\rho\right)}$ was nearly equal to or smaller than that of pMLE across all scenarios (Figure 2, and Figures S9 and S13).

5.2.2.2 | Scenario With Continuation to the Final Analysis. The results of simulation study 2 for $M = 2$ are shown in Figures S14–S22.

Figure S17 shows the differences in absolute CB between $\widehat{\theta}_{22}^{MCMAE,simple(\rho)}$ and $\widehat{\theta}_2^{CMAE}$ for $M = 2$ under true $\rho = 0.5$. Blue cell represents the absolute CB of $\widehat{\theta}_{22}^{MCMAE,simple(\rho)}$ being smaller

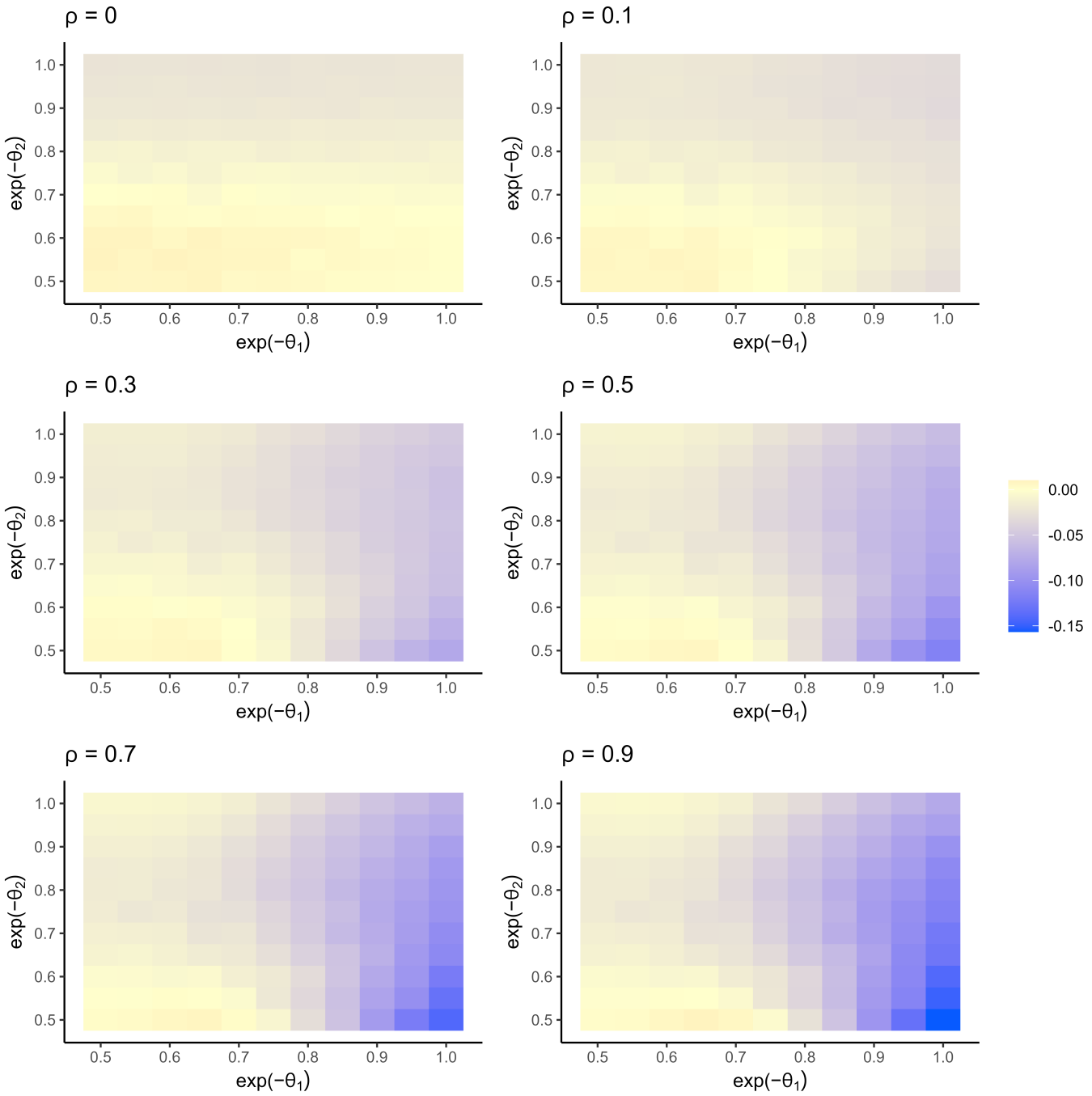

FIGURE 2 | Results of simulation study 2 for the differences in RMSE between $\widehat{\theta}_{21}^{MCMAE\left(\rho,\tau^*_{\max,\rho},w^*_\rho\right)}$ and $\widehat{\theta}_1^{pMLE}$ with $M = 1$ under true $\rho = 0.5$. IF = 0.5 and Pocock type alpha spending function was used. Each panel represents ρ for $\widehat{\theta}_{21}^{MCMAE\left(\rho,\tau^*_{\max,\rho},w^*_\rho\right)}$ specified from $\{0, 0.1, 0.3, 0.5, 0.7, 0.9\}$. Each cell represents the difference, and blue cell shows that the RMSE of $\widehat{\theta}_{21}^{MCMAE\left(\rho,\tau^*_{\max,\rho},w^*_\rho\right)}$ is smaller than that of $\widehat{\theta}_1^{pMLE}$.

than that of CMAE. Figure S17 showed that $\widehat{\theta}_{22}^{MCMAE,simple(\rho)}$ under $\exp(-\theta_2) \leq 0.7$ could have smaller absolute CB under $\exp(-\theta_1) \geq 0.8$ and larger absolute CB under $\exp(-\theta_1) \leq 0.65$ across all ρ specified for $\widehat{\theta}_{22}^{MCMAE,simple(\rho)}$. $\widehat{\theta}_{22}^{MCMAE,simple(\rho)}$ could have larger absolute CB under true $\rho = 0$ and $\exp(-\theta_2) \leq 0.7$ (Figure S14) and could have smaller absolute CB under true $\rho = 0.9$ and $\exp(-\theta_2) \leq 0.7$ (Figure S20).

EmpSE of $\widehat{\theta}_{22}^{MCMAE,simple(\rho)}$ was equal to or smaller than that of $\widehat{\theta}_2^{CMAE}$ across all scenarios and specified ρ (Figures S15, S18, and S21). Across all scenarios and specified ρ, the RMSE of $\widehat{\theta}_{22}^{MCMAE,simple(\rho)}$ was nearly equal to or smaller than that of CMAE (Figures S16, S19, and S22).

5.2.3 | High Level Summary of the Results of Simulation 2B

For $M = 1$, the absolute CB of the proposed method ($\widehat{\theta}_{21}^{MCMAE\left(\rho,\tau^*_{max,\rho},w^*_\rho\right)}$) was nearly equal to or smaller than that

of the existing method ($\widehat{\theta}_1^{pMLE}$) across almost all combinations of θ_1, θ_2, true ρ, and specified ρ. In cases where true θ_2 was large, true θ_1 was small, and ρ was truly zero, but the value of ρ was wrongly specified as positive, the absolute CB of the proposed method was larger than that of the existing method. Nevertheless, such a trend was distinct only when the specified ρ was quite larger than the true value of zero (e.g., $\rho \geq 0.7$). Roughly similar tendency was observed for $M = 2$, although the difference was slightly larger absolute CB under true $\rho = 0.5$ and both true θ_1 and θ_2 were large. This would be because of insufficient bias correction of $\widehat{\theta}_{22}^{MCMAE,simple(\rho)}$ similar to CMAE simple compared to $\widehat{\theta}_2^{CMAE}$. RMSE of proposed methods were nearly equal to or smaller than existing methods across all scenarios for both $M = 1$ and $M = 2$.

6 | Illustration

The FLAURA study [4, 5] is a confirmatory RCT for first-line *EGFR* mutated advanced non-small cell lung cancer patients comparing osimertinib (a third-generation EGFRi) with a first-generation EGFRi. This trial used an overall hierarchical strategy. No interim analyses were planned for PFS. Interim analysis of OS was planned using the O'Brien-Fleming type alpha spending function. An interim OS analysis was planned to be conducted at the time of the final PFS analysis. Interim OS analysis was performed if the PFS was statistically significant. Similarly, a final OS analysis was performed if the interim OS analysis recommended a continuation. The planned number of events for PFS was 359, and that for OS in the final analysis was 318. From the results, the test of PFS was significant, with 342 observed events and an estimated HR $\exp\left(-\widehat{\theta}_{11}\right)$ of 0.46. The OS interim analysis was not significant, with 141 observed events and an estimated HR $\exp\left(-\widehat{\theta}_{21}\right)$ of 0.63. The OS final analysis was significant, with 321 observed events and an estimated HR $\exp\left(-\widehat{\theta}_{22}\right)$ of 0.80. Again, on the label of U.S. approval, only the latter value was reported as the HR for OS [6].

MCMAE (τ_{max}) was applied to the MLE $\widehat{\theta}_{22}$ with τ_{max} varied in $\{1, 5, 100\}$ and ρ varied in $\{0, 0.5, 0.9\}$. This resulted in almost identical HR estimates. That is, $\exp\left(-\widehat{\theta}_{22}^{MCMAE(\tau_{max})}\right)$ was 0.80 with almost no dependence on τ_{max} and ρ. This implies that the uncorrected MLE might not have a large CB. The reported HR of 0.8 for OS in the label of U.S. approval may not need to be corrected from the viewpoint of CB.

To adjust for CB due to the significance of the final OS analysis in labeling, CB correction mentioned in Section S.5 (Supporting Information File) was adopted. $\exp\left(-\widehat{\theta}_{22}^{MCMAE(1)}\right)$ was 0.87 and $\exp\left(-\widehat{\theta}_{22}^{MCMAE(5)}\right)$ was 1.05, almost independence of the value of ρ.

For the illustration of early stopping ($M = 1$), we considered a hypothetical situation where the OS interim analysis was significant with the estimated value of HR $\exp\left(-\widehat{\theta}_{21}\right)$ of 0.5 for OS. Application of the MCMAE $\left(\tau^*_{\max}, w^*\right)$ produced the estimates $\exp\left(-\widehat{\theta}_{21}^{MCMAE\left(\tau^*_{\max}, w^*\right)}\right)$ of 0.55 across the values of ρ in $\{0, 0.5, 0.9\}$. This observed independence of ρ would be because the observed HR $\exp\left(-\widehat{\theta}_{11}\right)$ of 0.46 for PFS was much greater than the critical value $\exp(-b_{11})$ of 0.81 for PFS. When $\exp\left(-\widehat{\theta}_{11}\right)$ was hypothetically set to 0.8, $\exp\left(-\widehat{\theta}_{21}^{MCMAE\left(\tau^*_{\max}, w^*\right)}\right)$ takes its value of 0.55, 0.62, and 0.66 for $\rho = 0, 0.5,$ and 0.9, respectively. Here, the $\exp\left(-\widehat{\theta}_{11}\right)$ value of 0.8 is the value just barely crossing the critical value $\exp(-b_{11})$ of 0.81. Even in this hypothetical scenario where OS interim analysis was significant in the FLAURA study, adjustment of CB may not be required. However, other trials may suffer from large CB. One should be always cautious about the reporting of point HR estimates when a trial applies overall hierarchical strategies.

7 | Discussion

In RCTs testing PFS and OS using an overall hierarchical strategy, a simple version of a CMAE ($\widehat{\theta}_1^{CMAE,simple}$) and pMLE can leave CB regarding MLE of HR for OS when a trial stopped early for efficacy ($M = 1$). This is partly because the existing methods disregard the amount of selection driven by an early stop owing to PFS testing. This selection bias becomes stronger when the correlation between endpoints is high. Simulation study 1 found that a simple extension of $\widehat{\theta}_1^{CMAE,simple}$ for the multiple primary endpoint situations, termed as $\widehat{\theta}_{21}^{MCMAE,simple}$, left CB. Compared to this, our proposed multi-step version of $\widehat{\theta}_{21}^{MCMAE,simple}$, denoted as MCMAE $\left(\tau^*_{\max}, w^*\right)$, considering the idea of existing work (pMLE), had a smaller CB compared to the existing methods with comparable EmpSE to pMLE. The magnitude of CB correction was highlighted, especially when the true PFS effect was small and the trial was stopped early for efficacy ($M = 1$).

When the trial continued to the final analysis of OS ($M = 2$), simulation study 1 found that the existing methods tended to have CB toward overestimation because they did not consider PFS analysis. From the viewpoint of the RMSE, the MLE tended to perform better than the CB-adjusted estimators. This would be partly because the direction of the CB due to the PFS test and that of the group sequential test were opposite; thus, they would mutually cancel each other out. Nevertheless, when the CB due to either OS or PFS is much larger than the other CB, as in the case of small or large information fraction, $\widehat{\theta}_{22}^{MCMAE,simple}$ had smaller CB and RMSE than MLE and the existing methods.

To select the method to be adopted, we summarized the comparison between our proposed methods ($\widehat{\theta}_{21}^{MCMAE,simple}$ and $\widehat{\theta}_{21}^{MCMAE\left(\tau^*_{max}, w^*\right)}$ for $M = 1$, and $\widehat{\theta}_{22}^{MCMAE,simple}$ for $M = 2$) and existing methods as described below. Theoretically, $\widehat{\theta}_{21}^{MCMAE,simple}$ has a smaller CB than MLE for $M = 1$ and $\rho \geq 0$ unless it overcorrects CB according to proposition 7. $\widehat{\theta}_{21}^{MCMAE\left(\tau^*_{max}, w^*\right)}$ with $\tau^*_{max} \geq 1$ corrects CB stronger than $\widehat{\theta}_{21}^{MCMAE,simple}$, although larger τ^*_{max} may lead to overcorrection. For $M = 2$, no theoretical finding was obtained. According to simulation study 2, unless large ρ was specified under true ρ was zero, $\widehat{\theta}_{21}^{MCMAE\left(\tau^*_{max}, w^*\right)}$ and $\widehat{\theta}_{22}^{MCMAE,simple}$, even with mis-specified ρ, tended to have smaller absolute CB than existing methods, $\widehat{\theta}_1^{pMLE}$ for $M = 1$ and $\widehat{\theta}_2^{CMAE}$ for $M = 2$, respectively. $\widehat{\theta}_{21}^{MCMAE\left(\tau^*_{max}, w^*\right)}$ for $M = 1$ and $\widehat{\theta}_{22}^{MCMAE,simple}$ for

$M = 2$ had nearly equal to or smaller RMSE than $\widehat{\theta}_1^{pMLE}$ and $\widehat{\theta}_2^{CMAE}$, respectively.

The proposed method may be extended to more complex designs where more than two endpoints are tested hierarchically and at multiple timepoints. In some situations, an application of MCMAE $\left(\tau_{max}^*, w^*\right)$ may be easy. A typical situation includes a case where all three primary endpoints are reported as statistically significant following an overall hierarchical strategy. However, if any interim analysis is not significant, correcting CB would be challenging because the third endpoint may be influenced by the negative direction of CB due to the insignificance of interim analysis and positive direction of CB due to the significance of the preceding endpoint.

One limitation is that the proposed method is calculated under the given ρ. However, this may not be the case in actual clinical trials. One approach is to calculate the estimator by changing ρ within an appropriate range. Then, one can evaluate the variability of CB with respect to ρ. Another approach is estimating ρ from the data at hand, although the estimation of ρ may have CB. The second limitation is that interval estimation was not addressed. The third limitation is no clear guidance to determine the maximum number of iterations (τ_{max}) for $M = 2$. For $M = 1$, we prioritized a criterion that an excess bias correction should be avoided, which contradicts the study's main result of rejecting the null hypothesis. This criterion is the same as that used for pMLE. We now consider that this guidance seems appropriate in actual situations; however, the other criteria can be options. For $M = 2$, we do not have clear guidance to determine a maximum number of iterations, and these are future directions. The fourth limitation is that the setting of multiple interim analyses was not addressed. If the trial stops early for efficacy at the second or later interim analysis, the direction of CB will not be known. Thus, although MCMAE simple would be expanded to the situation, MCMAE $\left(\tau_{max}^*, w^*\right)$ would not be expanded because tau could not be determined as τ_{max}^*.

Acknowledgments

The authors acknowledge Dr. Mitsunori Ogawa for his helpful advice on early versions of our proposed method. This research was partly supported by AMED under Grant Number 701100.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

References

1. S. Green, A. Smith, B. Jacqueline, and J. Crowley, *Clinical Trials in Oncology*, 3rd ed. (CRC Press, 2012).

2. EMA, "Guideline on Clinical Evaluation of Anticancer Medicinal Products" (2019).

3. E. Glimm, W. Maurer, and F. Bretz, "Hierarchical Testing of Multiple Endpoints in Group-Sequential Trials," *Statistics in Medicine* 29, no. 2 (2010): 219–228, https://doi.org/10.1002/sim.3748.

4. J. C. Soria, Y. Ohe, J. Vansteenkiste, et al., "Osimertinib in Untreated EGFR-Mutated Advanced Non–Small-Cell Lung Cancer," *New England Journal of Medicine* 378, no. 2 (2018): 113–125, https://doi.org/10.1056/NEJMoa1713137.

5. S. S. Ramalingam, J. Vansteenkiste, D. Planchard, et al., "Overall Survival With Osimertinib in Untreated, EGFR-Mutated Advanced NSCLC," *New England Journal of Medicine* 382, no. 1 (2020): 41–50, https://doi.org/10.1056/NEJMoa1913662.

6. AstraZeneca, "Tagrisso (Osimertinib) [Package Insert]", https://www.accessdata.fda.gov/drugsatfda_docs/label/2024/208065s033lbl.pdf.

7. J. J. Zhang, G. M. Blumenthal, K. He, S. Tang, P. Cortazar, and R. Sridhara, "Overestimation of the Effect Size in Group Sequential Trials," *Clinical Cancer Research* 18, no. 18 (2012): 4872–4876, https://doi.org/10.1158/1078-0432.CCR-11-3118.

8. D. J. Slamon, P. Neven, S. Chia, et al., "Overall Survival With Ribociclib Plus Fulvestrant in Advanced Breast Cancer," *New England Journal of Medicine* 382, no. 6 (2020): 514–524, https://doi.org/10.1056/NEJMoa1911149.

9. C. Zhou, K. J. Tang, B. C. Cho, et al., "Amivantamab Plus Chemotherapy in NSCLC With EGFR Exon 20 Insertions," *New England Journal of Medicine* 389, no. 22 (2023): 2039–2051, https://doi.org/10.1056/NEJMoa2306441.

10. A. T. Shaw, T. M. Bauer, F. de Marinis, et al., "First-Line Lorlatinib or Crizotinib in Advanced ALK-Positive Lung Cancer," *New England Journal of Medicine* 383, no. 21 (2020): 2018–2029, https://doi.org/10.1056/NEJMoa2027187.

11. I. Ray-Coquard, P. Pautier, S. Pignata, et al., "Olaparib Plus Bevacizumab as First-Line Maintenance in Ovarian Cancer," *New England Journal of Medicine* 381, no. 25 (2019): 2416–2428, https://doi.org/10.1056/NEJMoa1911361.

12. D. S. Robertson, B. Choodari-Oskooei, M. Dimairo, L. Flight, P. Pallmann, and T. Jaki, "Point Estimation for Adaptive Trial Designs I: A Methodological Review," *Statistics in Medicine* 42, no. 2 (2023): 122–145, https://doi.org/10.1002/sim.9605.

13. D. S. Robertson, B. Choodari-Oskooei, M. Dimairo, L. Flight, P. Pallmann, and T. Jaki, "Point Estimation for Adaptive Trial Designs II: Practical Considerations and Guidance," *Statistics in Medicine* 42, no. 14 (2023): 2496–2520, https://doi.org/10.1002/sim.9734.

14. X. Fan, D. L. DeMets, and K. K. G. Lan, "Conditional Bias of Point Estimates Following a Group Sequential Test," *Journal of Biopharmaceutical Statistics* 14, no. 2 (2004): 505–530, https://doi.org/10.1081/BIP-120037195.

15. J. F. Troendle and K. F. Yu, "Conditional Estimation Following a Group Sequential Clinical Trial," *Communications in Statistics - Theory and Methods* 28, no. 7 (1999): 1617–1634, https://doi.org/10.1080/03610929908832376.

16. H. Zhong and R. L. Prentice, "Bias-Reduced Estimators and Confidence Intervals for Odds Ratios in Genome-Wide Association Studies," *Biostatistics* 9, no. 4 (2008): 621–634, https://doi.org/10.1093/biostatistics/kxn001.

17. M. S. Pepe, Z. Feng, G. Longton, and J. Koopmeiners, "Conditional Estimation of Sensitivity and Specificity From a Phase 2 Biomarker Study Allowing Early Termination for Futility," *Statistics in Medicine* 28, no. 5 (2009): 762–779, https://doi.org/10.1002/sim.3506.

18. J. S. Koopmeiners, Z. Feng, and M. S. Pepe, "Conditional Estimation After a Two-Stage Diagnostic Biomarker Study That Allows Early Termination for Futility," *Statistics in Medicine* 31, no. 5 (2012): 420–435, https://doi.org/10.1002/sim.4430.

19. M. Shimura, K. Maruo, and M. Gosho, "Conditional Estimation Using Prior Information in 2-Stage Group Sequential Designs Assuming Asymptotic Normality When the Trial Terminated Early," *Pharmaceutical Statistics* 17, no. 5 (2018): 400–413, https://doi.org/10.1002/pst.1859.

20. I. C. Marschner, M. Schou, and A. J. Martin, "Estimation of the Treatment Effect Following a Clinical Trial That Stopped Early for Benefit," *Statistical Methods in Medical Research* 31, no. 12 (2022): 2456–2469, https://doi.org/10.1177/09622802221122445.

21. C. Jennison and B. W. Turnbull, *Group Sequential Methods With Applications to Clinical Trials* (Chapman and Hall/CRC, 1999).

22. M. Shimura, M. Gosho, and A. Hirakawa, "Comparison of Conditional Bias-Adjusted Estimators for Interim Analysis in Clinical Trials With Survival Data," *Statistics in Medicine* 36, no. 13 (2017): 2067–2080, https://doi.org/10.1002/sim.7258.

23. J. Whitehead, "On the Bias of Maximum Likelihood Estimation Following a Sequential Test," *Biometrika* 73, no. 3 (1986): 573–581, https://doi.org/10.2307/2336521.

24. I. C. Marschner and I. M. Schou, "Underestimation of Treatment Effects in Sequentially Monitored Clinical Trials That Did Not Stop Early for Benefit," *Statistical Methods in Medical Research* 28, no. 10–11 (2019): 3027–3041, https://doi.org/10.1177/0962280218795320.

25. M. J. Grayling and J. M. Wason, "Point Estimation Following a Two-Stage Group Sequential Trial," *Statistics in Medicine* 32, no. 2 (2023): 287–304, https://doi.org/10.1177/09622802221137745.

26. H. Y. Guo and A. Liu, "A Simple and Efficient Bias-Reduced Estimator of Response Probability Following a Group Sequential Phase II Trial," *Journal of Biopharmaceutical Statistics* 15, no. 5 (2005): 773–781, https://doi.org/10.1081/BIP-200067771.

27. A. Liu, J. F. Troendle, K. F. Yu, and V. W. Yuan, "Conditional Maximum Likelihood Estimation Following a Group Sequential Test," *Biometrical Journal* 46, no. 6 (2004): 760–768, https://doi.org/10.1002/bimj.200410076.

28. J. Gou and D. Xi, "Hierarchical Testing of a Primary and a Secondary Endpoint in a Group Sequential Design With Different Information Times," *Statistics in Biopharmaceutical Research* 11, no. 4 (2019): 398–406, https://doi.org/10.1080/19466315.2018.1546613.

29. R. Kan and C. Robotti, "On Moments of Folded and Truncated Multivariate Normal Distributions," *Journal of Computational and Graphical Statistics* 26, no. 4 (2017): 930–934, https://doi.org/10.1080/10618600.2017.1322092.

30. A. A. Tsiatis, "The Asymptotic Joint Distribution of the Efficient Scores Test for the Proportional Hazards Model Calculated Over Time," *Biometrika* 68, no. 1 (1981): 311–315, https://doi.org/10.2307/2335832.

31. K. K. G. Lan and D. L. DeMets, "Discrete Sequential Boundaries for Clinical Trials," *Biometrika* 70, no. 3 (1983): 659–663, https://doi.org/10.2307/2336502.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Figure S1.** Relation between $\widehat{\theta}_{21}^{MCMAE(4)}$ and $\widehat{\theta}_{21}$ (Panel A), between $\widehat{\theta}_{21}^{MCMAE(4)}$ and $\widehat{\theta}_{11}$ (Panel B), and between $\widehat{\theta}_{21}^{MCMAE(\tau_{max})}$ and τ_{max} (Panel C). **Figure S2.** Contour lines of $\widehat{\theta}_{21}^{MCMAE(\tau_{max})}$ as function of τ_{max} and $\widehat{\theta}_{21}$ (Panel A), of τ_{max} and $\widehat{\theta}_{11}$ (Panel B), and of $\widehat{\theta}_{11}$ and $\widehat{\theta}_{21}$ (Panel C). **Figure S3.** Contour lines of $\widehat{\theta}_{21}^{MCMAE(\tau_{max})}$ as function of $\widehat{\theta}_{11}$ and $\widehat{\theta}_{21}$ with τ_{max} varied from 1 to 25. **Figure S4.** Contour lines of $\widehat{\theta}_{21}^{MCMAE(\tau_{max})}$ as function of $\widehat{\theta}_{11}$ and $\widehat{\theta}_{21}$ with τ_{max} varied from 26 to 50. **Figure S5.** Contour lines of $\widehat{\theta}_{21}^{MCMAE(\tau_{max})}$ as function of $\widehat{\theta}_{11}$ and $\widehat{\theta}_{21}$ with τ_{max} varied from 51 to 75. **Figure S6.** Contour lines of $\widehat{\theta}_{21}^{MCMAE(\tau_{max})}$ as function of $\widehat{\theta}_{11}$ and $\widehat{\theta}_{21}$ with τ_{max} varied from 76 to 100. **Table S1.** Results of simulation study 1 and 2A in $M = 1$, IF = 0.5, $\exp(-\theta_1) = 0.6$, $\exp(-\theta_2) = 0.65$. **Table S2.** Results of simulation study 1 and 2A in $M = 1$, IF = 0.5, $\exp(-\theta_1) = 0.6$, $\exp(-\theta_2) = 0.9$. **Table S3.** Results of simulation study 1 and 2A in $M = 1$, IF = 0.5, $\exp(-\theta_1) = 0.85$, $\exp(-\theta_2) = 0.65$. **Table S4.** Results of simulation study 1 and 2A in $M = 1$, IF = 0.5, $\exp(-\theta_1) = 0.85$, $\exp(-\theta_2) = 0.9$. **Table S5.** Results of simulation study 1 and 2A in $M = 2$ IF = 0.25, $\exp(-\theta_1) = 0.6$, $\exp(-\theta_2) = 0.65$. **Table S6.** Results of simulation study 1 and 2A in $M = 2$, IF = 0.25, $\exp(-\theta_1) = 0.6$, $\exp(-\theta_2) = 0.9$. **Table S7.** Results of simulation study 1 and 2A in $M = 2$, IF = 0.25, $\exp(-\theta_1) = 0.85$, $\exp(-\theta_2) = 0.65$. **Table S8.** Results of simulation study 1 and 2A in $M = 2$, IF = 0.25, $\exp(-\theta_1) = 0.85$, $\exp(-\theta_2) = 0.9$. **Table S9.** Results of simulation study 1 and 2A in $M = 2$, IF = 0.5, $\exp(-\theta_1) = 0.6$, $\exp(-\theta_2) = 0.65$. **Table S10.** Results of simulation study 1 and 2A in $M = 2$, IF = 0.5, $\exp(-\theta_1) = 0.6$, $\exp(-\theta_2) = 0.9$. **Table S11.** Results of simulation study 1 and 2A in $M = 2$, IF = 0.5, $\exp(-\theta_1) = 0.85$, $\exp(-\theta_2) = 0.65$. **Table S12.** Results of simulation study 1 and 2A in $M = 2$, IF = 0.5, $\exp(-\theta_1) = 0.85$, $\exp(-\theta_2) = 0.9$. **Table S13.** Results of simulation study 1 and 2A in $M = 2$, IF = 0.75, $\exp(-\theta_1) = 0.6$, $\exp(-\theta_2) = 0.65$. **Table S14.** Results of simulation study 1 and 2A in $M = 2$, IF = 0.75, $\exp(-\theta_1) = 0.6$, $\exp(-\theta_2) = 0.9$. **Table S15.** Results of simulation study 1 and 2A in $M = 2$, IF = 0.75, $\exp(-\theta_1) = 0.85$, $\exp(-\theta_2) = 0.65$. **Table S16.** Results of simulation study 1 and 2A in $M = 2$, IF = 0.75, $\exp(-\theta_1) = 0.85$, $\exp(-\theta_2) = 0.9$. **Figure S7.** Results of simulation study 2 for the differences in the absolute CB between $\widehat{\theta}_{21}^{MCMAE\left(\rho,\tau^*_{\max,\rho},w^*_\rho\right)}$ and $\widehat{\theta}_1^{pMLE}$ with $M = 1$ under true $\rho = 0$. **Figure S8.** Results of simulation study 2 for the differences in EmpSE between $\widehat{\theta}_{21}^{MCMAE\left(\rho,\tau^*_{\max,\rho},w^*_\rho\right)}$ and $\widehat{\theta}_1^{pMLE}$ with $M = 1$ under true $\rho = 0$. **Figure S9.** Results of simulation study 2 for the differences in RMSE between $\widehat{\theta}_{21}^{MCMAE\left(\rho,\tau^*_{\max,\rho},w^*_\rho\right)}$ and $\widehat{\theta}_1^{pMLE}$ with $M = 1$ under true $\rho = 0$. **Figure S10.** Results of simulation study 2 for the differences in EmpSE between $\widehat{\theta}_{21}^{MCMAE\left(\rho,\tau^*_{\max,\rho},w^*_\rho\right)}$ and $\widehat{\theta}_1^{pMLE}$ with $M = 1$ under true $\rho = 0.5$. **Figure S11.** Results of simulation study 2 for the differences in absolute CB between $\widehat{\theta}_{21}^{MCMAE\left(\rho,\tau^*_{\max,\rho},w^*_\rho\right)}$ and $\widehat{\theta}_1^{pMLE}$ with $M = 1$ under true $\rho = 0.9$. **Figure S12.** Results of simulation study 2 for the differences in EmpSE between $\widehat{\theta}_{21}^{MCMAE\left(\rho,\tau^*_{\max,\rho},w^*_\rho\right)}$ and $\widehat{\theta}_1^{pMLE}$ with $M = 1$ under true $\rho = 0.9$. **Figure S13.** Results of simulation study 2 for the differences in RMSE between $\widehat{\theta}_{21}^{MCMAE\left(\rho,\tau^*_{\max,\rho},w^*_\rho\right)}$ and $\widehat{\theta}_1^{pMLE}$ with $M = 1$ under true $\rho = 0.9$. **Figure S14.** Results of simulation study 2 for the differences in absolute CB between $\widehat{\theta}_{22}^{MCMAE,simple(\rho)}$ and $\widehat{\theta}_2^{CMAE}$ with $M = 2$ under true $\rho = 0$. **Figure S15.** Results of simulation study 2 for the differences in EmpSE between $\widehat{\theta}_{22}^{MCMAE,simple(\rho)}$ and $\widehat{\theta}_2^{CMAE}$ with, $M = 2$ under true $\rho = 0$. **Figure S16.** Results of simulation study 2 for the differences in RMSE between $\widehat{\theta}_{22}^{MCMAE,simple(\rho)}$ and $\widehat{\theta}_2^{CMAE}$ with $M = 2$ under true $\rho = 0$. **Figure S17.** Results of simulation study 2 for the differences in absolute CB between $\widehat{\theta}_{22}^{MCMAE,simple(\rho)}$ and $\widehat{\theta}_2^{CMAE}$ with $M = 2$ under true $\rho = 0.5$. **Figure S18.** Results of simulation study 2 for the differences in EmpSE between $\widehat{\theta}_{22}^{MCMAE,simple(\rho)}$ and $\widehat{\theta}_2^{CMAE}$ with, $M = 2$ under true $\rho = 0.5$. **Figure S19.** Results of simulation study 2 for the differences in RMSE between $\widehat{\theta}_{22}^{MCMAE,simple(\rho)}$ and $\widehat{\theta}_2^{CMAE}$ with $M = 2$ under true $\rho = 0.5$. **Figure S20.** Results of simulation study 2 for the differences in absolute CB between $\widehat{\theta}_{22}^{MCMAE,simple(\rho)}$ and $\widehat{\theta}_2^{CMAE}$ with $M = 2$ under true $\rho = 0.9$. **Figure S21.** Results of simulation study 2 for the differences in EmpSE between $\widehat{\theta}_{22}^{MCMAE,simple(\rho)}$ and $\widehat{\theta}_2^{CMAE}$ with, $M = 2$ under true $\rho = 0.9$. **Figure S22.** Results of simulation study 2 for the differences in RMSE between $\widehat{\theta}_{22}^{MCMAE,simple(\rho)}$ and $\widehat{\theta}_2^{CMAE}$ with $M = 2$ under true $\rho = 0.9$.

Appendix A

Repeated Correction for Single Primary Endpoint

We explain the iterated correction method for a single primary endpoint following the notation in Section 2. Let $\tau \in \mathbb{N}_{\geq 1}$ be the number of the repetition and $\widehat{\theta}_1^{CMAE(\tau)}$ be the CB-adjusted estimator at the τ-th repetition. The repeated correction can be achieved by the following steps until τ reaches the upper limit, $\tau_{\max} \in \mathbb{N}_{\geq 1}$:

Step 1. Set $\tau = 1$.

Step 2. By substituting $\widehat{\theta}_1^{CMAE(\tau-1)}$ for θ in the CB term (1), calculate $\sigma_1 \frac{\phi\left(\frac{\widehat{\theta}_1^{CMAE(\tau-1)}-b_1}{\sigma_1}\right)}{\Phi\left(\frac{\widehat{\theta}_1^{CMAE(\tau-1)}-b_1}{\sigma_1}\right)}$, where the initial values are set as $\widehat{\theta}_1^{CMAE(0)} = \widehat{\theta}_1$.

Step 3. Update the CB-corrected estimators at the τ-th iteration:

$$\widehat{\theta}_1^{CMAE(\tau)} \coloneqq \widehat{\theta}_1 - \sigma_1 \frac{\phi\left(\frac{\widehat{\theta}_1^{CMAE(\tau-1)}-b_1}{\sigma_1}\right)}{\Phi\left(\frac{\widehat{\theta}_1^{CMAE(\tau-1)}-b_1}{\sigma_1}\right)}$$

Step 4. Increment τ by 1. If τ is greater than τ_{max}, stop the iteration steps; otherwise, repeat Step 2.

Here, $\widehat{\theta}_1^{CMAE(1)}$ is identical to $\widehat{\theta}_1^{CMAE,simple}$. Additionally, $\widehat{\theta}_1^{CMAE(\tau_{\max})}$ decreases with respect to $\tau_{\max}$ (Supporting Information File S.3). Thus, the following inequality holds for τ_{max} and τ'_{max} that satisfy $\tau_{max} < \tau'_{max}$:

$$\widehat{\theta}_1^{CMAE(\tau'_{max})} < \widehat{\theta}_1^{CMAE(\tau_{max})} < \widehat{\theta}_1$$

This implies that more iteration results in a stronger correction of the CB.

Details of $\widehat{\theta}_{21}^{MCMAE(\tau^*_{\max},w^*)}$

In this subsection, we explain the details of $\widehat{\theta}_{21}^{MCMAE(\tau^*_{\max},w^*)}$, including how to determine $\tau^*_{\max}$ and w^*. Because $\widehat{\theta}_{21}^{MCMAE(\tau_{\max})}$ is the function of $\left(\widehat{\theta}_{11}, \widehat{\theta}_{21}, \tau_{\max}\right)$, let $g\left(\widehat{\theta}_{11}, \widehat{\theta}_{21}, \tau_{\max}\right)$ represent $\widehat{\theta}_{21}^{MCMAE(\tau_{\max})}$, where $\widehat{\theta}_{11} > b_{11}, \widehat{\theta}_{21} > b_{21}$. Then, $g\left(\widehat{\theta}_{11}, \widehat{\theta}_{21}, 0\right)$ represents MLE $\widehat{\theta}_{21}$, and $g\left(\widehat{\theta}_{11}, \widehat{\theta}_{21}, \tau_{\max}\right)$ with large $\tau_{\max}$ was observed to be close to $\hat{\theta}_{21}^{MCMAE}$. Additionally, the numerical examination illustrated in Figures S1–S6 in the Supporting Information File S.8 suggested that $g\left(\widehat{\theta}_{11}, \widehat{\theta}_{21}, \tau_{\max}\right)$ may monotonically increase with respect to $\widehat{\theta}_{11}, \widehat{\theta}_{21}$ and monotonically decrease with respect to $\tau_{\max}$. Note that the calculation error in integration may affect the estimates under a large $\tau_{\max}$. Thus, let $\tau^*_{max} = \max\left\{\tau \in \mathbb{N}_{\geq 0} | g\left(b_{11}, b_{21}, \tau\right) \geq 0\right\}$. The following inequality would hold:

$$g\left(\widehat{\theta}_{11}, \widehat{\theta}_{21}, \tau^*_{\max}\right) > g\left(b_{11}, b_{21}, \tau^*_{\max}\right) \geq 0$$

If this formula is true, it implies that $g\left(\widehat{\theta}_{11}, \widehat{\theta}_{21}, \tau^*_{\max}\right)$ would correct CB to the degree that it does not contradict the results of hypothesis testing.

The infimum of $g\left(\widehat{\theta}_{11}, \widehat{\theta}_{21}, \tau^*_{\max}\right)$ is not zero but $g\left(b_{11}, b_{21}, \tau^*_{\max}\right)$, because $\tau^*_{\max}$ is not a continuous value. Zero may be better for the infimum of the CB-adjusted estimator, because it is the infimum of the values that imply the superiority of the experimental group. Thus, let $h\left(\widehat{\theta}_{11}, \widehat{\theta}_{21}, \tau^*_{\max}, w\right)$ be the weighted mean between $g\left(\widehat{\theta}_{11}, \widehat{\theta}_{21}, \tau^*_{\max}+1\right)$ and $g\left(\widehat{\theta}_{11}, \widehat{\theta}_{21}, \tau^*_{\max}\right)$:

$$h\left(\widehat{\theta}_{11}, \widehat{\theta}_{21}, \tau^*_{\max}, w\right) \coloneqq wg\left(\widehat{\theta}_{11}, \widehat{\theta}_{21}, \tau^*_{\max}+1\right) + (1-w)g\left(\widehat{\theta}_{11}, \widehat{\theta}_{21}, \tau^*_{\max}\right)$$

where $w \in (0, 1]$. $h\left(\widehat{\theta}_{11}, \widehat{\theta}_{21}, \tau^*_{\max}, w\right)$ would monotonically decrease with respect to w; thus let w^* be the value satisfying

$$h\left(b_{11}, b_{21}, \tau^*_{\max}, w^*\right) = 0$$

Then, we define the CB-adjusted estimator as:

$$\widehat{\theta}_{21}^{MCMAE(\tau^*_{\max},w^*)} \coloneqq h\left(\widehat{\theta}_{11}, \widehat{\theta}_{21}, \tau^*_{\max}, w^*\right)$$

Here, the infimum of $\widehat{\theta}_{21}^{MCMAE(\tau^*_{\max},w^*)}$ would be zero (i.e., $\widehat{\theta}_{21}^{MCMAE(\tau^*_{\max},w^*)} > 0$); this was true in the simulation study (Section 4).