Statistica Sinica Preprint No: SS-2025-0427	
Title	Model-robust Inference for Seamless Ii/iii Trials with Covariate Adaptive Randomization
Manuscript ID	SS-2025-0427
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0427
Complete List of Authors	Kun Yi and Lucy Xia
Corresponding Authors	Lucy Xia
E-mails	lucyxia.2010@gmail.com
Notice: Accepted author version.	

Model-robust Inference for Seamless II/III Trials with Covariate Adaptive Randomization

Kun Yi and Lucy Xia

Department of ISOM, HKUST

Abstract: Seamless phase II/III trials have become a cornerstone of modern drug development, offering a means to accelerate evaluation while maintaining statistical rigor. However, most existing inference procedures are model-based, designed primarily for continuous outcomes, and often neglect the stratification used in covariate-adaptive randomization (CAR), limiting their practical relevance. In this paper, we propose a unified, model-robust framework for seamless phase II/III trials grounded in generalized linear models (GLMs), enabling valid inference across diverse outcome types, estimands, and CAR schemes. Using Z-estimation, we derive the asymptotic properties of treatment effect estimators and explicitly characterize how their variance depends on the underlying randomization procedure. Based on these results, we develop adjusted Wald tests that, together with Dunnett's multiple-comparison procedure and the inverse–χ^2 combination method, ensure valid overall Type I error. Extensive simulation studies and a trial example demonstrate that the proposed model-robust tests achieve superior power and reliable inference compared to conventional approaches.

Key words and phrases: Seamless phase II/III trial, covariate-adaptive randomization, generalized linear model, model-robust inference, bootstrap adjustment.

1. Introduction

The integration of seamless phase II/III clinical trials into modern drug development represents a transformative shift toward accelerating therapeutic evaluation while maintaining statistical rigor (Jennison and Turnbull, 2007; Hampson and Jennison, 2015). Driven by

Corresponding author: lucyxia@ust.hk; ORCID:0000-0002-8805-6660.

regulatory mandates to streamline clinical research (FDA, 2022), seamless designs are introduced to combine dose selection (phase II) and confirmatory analysis (phase III) under a single protocol. By eliminating the traditional 6–12 month gap between phases, these designs reduce operational delays, lower costs, and enable continuous patient monitoring, thereby enhancing both pharmaceutical profitability and patient access to novel therapies (Prowell et al., 2016). However, their adoption introduces formidable statistical challenges, including inflated Type I error rates due to treatment selection and multiplicity (Bauer et al., 2010), and the need to reconcile adaptive designs with covariate-adaptive randomization (CAR), an essential mechanism for maintaining covariate balance in modern trials (Shao et al., 2010).

Most existing methodologies for seamless trials focus on continuous outcomes, relying on linear models and conventional Wald tests for inference (Liu et al., 2002; Schmidli et al., 2006). More recently, several studies have extended these approaches to accommodate binary or mixed endpoints (Jenkins et al., 2011; Chen et al., 2018; Li et al., 2024). However, the vast majority of these methods assume complete randomization, and research on seamless trials conducted under CAR remains limited. This creates an important gap between existing methodology and modern clinical-trial practice. Clinical endpoints are often discrete, such as binary tumor response rates or ordinal disease severity scores, or mixed, such as survival data with censoring. Even when continuous measures are collected, practitioners frequently dichotomize them for interpretability, a practice that can invalidate assumptions underpinning traditional analyses (Sverdlov, 2015). While generalized linear models (GLMs) provide a natural framework for non-Gaussian outcomes (McCullagh, 2019), their use in seamless trials under CAR remains underdeveloped. In addition, model misspecification, such as omitting prognostic covariates or mis-specifying link functions, can affect power and Type I

error control. Diverse estimands, including the average treatment effect, log relative risk, and log odds ratio, further call for a flexible inferential framework. This paper aims to develop a model-robust framework that accommodates diverse outcome types and estimands in seamless phase II/III trials under CAR.

We next review the relevant randomization and inference literature. In the existing literature, discussions of seamless phase II/III designs have largely focused on complete randomization. For example, following the approach of Bauer and Kieser (1999), researchers have employed the closure principle (Bretz et al., 2006; Marcus et al., 1976), combination tests (Bauer and Kohne, 1994), and multiple testing procedures (Simes, 1986; Dunnett, 1955) to control the family-wise Type I error rate. Additionally, Liu et al. (2002) provided a theoretical foundation for general two-stage adaptive designs, while Koenig et al. (2008) proposed the adaptive Dunnett test based on the conditional error rate (Müller and Schäfer, 2001).

However, considering only complete randomization is overly simplified. In practice, CAR procedures are widely implemented to mitigate covariate imbalances and improve statistical power, especially in trials with important prognostic factors (Baldi Antognini and Zagoraiou, 2011; Hu and Hu, 2012). Specifically, CAR is often implemented through stratified permuted block (STRPB) designs (Zelen, 1974) or minimization-based methods (Pocock and Simon, 1975; Taves, 1974), which enhance trial efficiency by dynamically balancing baseline covariates. Despite its widespread use in practice, CAR has only recently been formally considered in the analysis of seamless trials (Ma et al., 2022). Ma et al. (2022) pioneered this integration by developing a CAR-adjusted linear model and demonstrated that adjusted test statistics, such as replacing pooled variance with model-robust estimators, can restore Type I error

control and improve power under CAR. Nevertheless, their reliance on linear models limits applicability beyond continuous outcomes.

While CAR improves covariate balance, it complicates subsequent statistical inference and may lead to reduced Type I error when conventional variance estimators are used (Shao et al., 2010; Shao and Yu, 2013; Bugni et al., 2018, 2019; Liu and Hu, 2023). This has motivated the development of valid standard error estimators that adjust for the covariates used in CAR to ensure valid inference (Shao et al., 2010; Ma et al., 2015, 2020; Liu et al., 2024). Moreover, incorporating additional auxiliary covariates in the analysis can further improve estimation efficiency and increase test power (Ma et al., 2022; Ye et al., 2022, 2023; Gu et al., 2023). However, most existing methods are confined to linear models. Although recent work extends these results to generalized linear models under CAR (Wang et al., 2023; Zhao et al., 2025), it remains limited to two-arm trials. In contrast, our work generalizes robust inference to multi-arm CAR trials and integrates it into seamless Phase II/III designs, thereby providing a more comprehensive framework.

This paper bridges these gaps by proposing a unified framework for seamless phase II/III trials that harmonizes CAR with GLMs. Our contributions are threefold. First, we develop a model-robust Z-estimation approach under GLMs, yielding consistent and asymptotically normal estimators for treatment effects across outcome types. This generality accommodates diverse estimands, from relative risk to log odds ratios, without requiring correct specification of a full parametric likelihood. Second, we integrate CAR into seamless designs, extending the theoretical foundations established by Ma et al. (2022). By deriving the asymptotic properties of test statistics under CAR, including covariance structures that account for the underlying randomization procedure, we enable valid hypothesis testing and error control.

Third, we demonstrate practical impact through extensive simulations and a hypothetical trial example, illustrating how our framework improves power while maintaining Type I error control compared with conventional methods.

The remainder of this paper is organized as follows. Section 2 presents the general framework for seamless phase II/III trials with CAR Section 3 develops a model-robust Z-estimation framework under the GLM setting and establishes the consistency and asymptotic normality of the resulting estimators, supported by a precise decomposition of the standard error. Section 4 specializes these general results to three commonly used estimands. Section 5 evaluates the finite-sample performance of the proposed methods through extensive simulations, comparing Type I error control and power across diverse scenarios. Section 6 illustrates the practical utility of our framework in a hypothetical trial for treating alopecia areata. Finally, Section 7 concludes with a discussion of the broader implications and potential extensions. By unifying CAR with GLMs in seamless designs, this work advances statistical theory and offers a regulatory-compliant toolkit for the next generation of clinical research.

1.1 Notation

Let $[K] := \{1, 2, \ldots, K\}$ denote an index set. For a matrix $\boldsymbol{M} \in \mathbb{R}^{K\times K}$, let $[\boldsymbol{M}]_{(k,k)}$ denote its (k,k)-th entry, and for a vector $\boldsymbol{x} \in \mathbb{R}^q$, let $\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}^\top \boldsymbol{x}}$ denote the Euclidean norm; $\mathbf{1}_K$ and $\mathbf{0}_K$ denote the K-dimensional vectors of ones and zeros, respectively. For a square matrix $\boldsymbol{M} \in \mathbb{R}^{q\times q}$, the spectral norm is $\|\boldsymbol{M}\| = \sup\{\|\boldsymbol{M}\boldsymbol{x}\| : \|\boldsymbol{x}\| = 1, \boldsymbol{x} \in \mathbb{R}^q\}$. A sequence of random vectors $\boldsymbol{X}_n$ satisfies $\boldsymbol{X}_n = o_p(\mathbf{1})$ if $\boldsymbol{X}_n \xrightarrow{p} \mathbf{0}$ and $\boldsymbol{X}_n = O_p(\mathbf{1})$ if $\|\boldsymbol{X}_n\|$ is bounded in probability. For a vector-valued function $\boldsymbol{f}(\boldsymbol{\theta}) = (f_1(\boldsymbol{\theta}), \ldots, f_m(\boldsymbol{\theta}))^\top : \mathbb{R}^p \to \mathbb{R}^m$, the

Jacobian is $\dot{\boldsymbol{f}}(\boldsymbol{\theta}) = \partial \boldsymbol{f}(\boldsymbol{\theta})/\partial \boldsymbol{\theta} \in \mathbb{R}^{m\times p}$, where the j-th row equals the gradient $\dot{\boldsymbol{f}}_j(\boldsymbol{\theta}) = \partial f_j(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$, and the Hessian of component f_j is $\ddot{\boldsymbol{f}}_j(\boldsymbol{\theta}) = \partial^2 f_j(\boldsymbol{\theta})/\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top \in \mathbb{R}^{p\times p}$. Finally, for a vector $\boldsymbol{a} = (a_1, \ldots, a_K)^\top$, we write $\mathrm{diag}\{a_1, \ldots, a_K\}$ for the diagonal matrix with entries a_k on the diagonal and zeros elsewhere. $\boldsymbol{e}_k$ denotes the k-th standard basis vector in the Euclidean space. $\xrightarrow{p}$ and $\xrightarrow{d}$ denote convergence in probability and distribution, respectively.

2. Seamless Phase II/III Trial with CAR

2.1 An overview of the procedure

We work under a general framework for seamless phase II/III trials that accommodates various treatment effect metrics beyond the traditional average treatment effect. Let $\boldsymbol{Y} = (Y(0), \ldots, Y(K))^\top$, $\boldsymbol{\mu} = (\mu_0, \mu_1, \ldots, \mu_K) := \mathbb{E}[\boldsymbol{Y}] \in \mathbb{R}^{K+1}$ with μ_0 and μ_k represent the average potential outcome in the control arm and the k-th treatment arm for $k \in [K]$, respectively. Instead of focusing solely on the difference $\mu_k - \mu_0$, our framework allows treatment effects to be expressed as functions of these parameters. Specifically, for a chosen function g, we define the treatment effect metric as $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_K)$, where each component is given by $\delta_k = g(\mu_k) - g(\mu_0)$. This flexible formulation includes common metrics as special cases: *the log relative risk* (logRR) when g is the logarithm, *the log odds ratio* (LOR) when g is the logit function, and *the average treatment effect* (ATE) when g is the identity. Such flexibility enables efficient design and analysis, accommodating different targets. Building on this framework, the seamless phase II/III trial proceeds as follows:

- Phase II (e.g. Stage 1): Identification of promising treatments by testing

$$H_0^1 : \boldsymbol{\delta} = \mathbf{0}_K \quad \text{versus} \quad H_1^1 : \delta_k > 0 \quad \text{for some } k \in [K],$$

based on estimator $\widehat{\boldsymbol{\delta}} = (\widehat{\delta}_1, \ldots, \widehat{\delta}_K)$ constructed using data from N_1 patients. Dunnett's test (Dunnett, 1964) is used to control family-wise Type I error while accounting for multiplicity.

- Phase III (e.g. Stage 2): Selection of the final treatment

$$k^* = \arg \max_{1 \le k \le K} \widehat{\delta}_k,$$

and testing based on an additional N_2 patients:

$$H_0^2 : \delta_{k^*} = 0 \quad \text{versus} \quad H_1^2 : \delta_{k^*} > 0.$$

Conclusions are drawn by combining evidence from both stages using the inverse chi-square method (Bauer and Kohne, 1994).

2.2 Covariate-adaptive randomization

Recall that the numbers of patients enrolled in Stage 1 and Stage 2 are denoted by N_1 and N_2, respectively. For subject $i \in [N_j]$ and $j \in \{1, 2\}$, let $\boldsymbol{X}_i = (X_{i,1}, \ldots, X_{i,q})^\top \in \mathbb{R}^q$ collect baseline covariates, and define the treatment-assignment vector

$$\boldsymbol{T}_i = (T_i^0, T_i^1, \ldots, T_i^K)^\top \in \{0, 1\}^{K+1}, \qquad \sum_{k=0}^{K} T_i^k = 1,$$

where $T_i^k = 1$ indicates assignment to arm $k \in \{0, \ldots, K\}$. Following Hu and Rosenberger (2006), let $\mathcal{T}_i$, $\mathcal{X}_{i,\text{ex}}$, $\mathcal{X}_i$, and $\mathcal{Y}_i$ denote the σ-algebras generated by the first i treatment assignments, the discretized covariates used in randomization, the baseline covariates, and the observed outcomes, respectively. Define

$$\mathcal{F}_i = \mathcal{X}_{i+1,\text{ex}} \otimes \mathcal{T}_i \otimes \mathcal{X}_i \otimes \mathcal{Y}_i, \qquad \boldsymbol{\phi}_i = \mathbb{E}[\boldsymbol{T}_i \mid \mathcal{F}_{i-1}],$$

so $\boldsymbol{\phi}_i$ is the (vector) assignment probability for subject i given past assignments and observed information, allowing the randomization covariates $\boldsymbol{X}_{i,\text{ex}}$ to be a subset of the outcome-relevant baseline covariates $\boldsymbol{X}_i$. Forming discrete strata $S_i = S(\boldsymbol{X}_{i,\text{ex}}) = s \in \{1, \ldots, s_{\max}\}$ from $\boldsymbol{X}_{i,\text{ex}}$, CAR procedures sequentially assign patients to minimize a weighted imbalance score that combines overall imbalance, marginal imbalances of the j-th discrete (or discretized) covariate (for $j \in [q]$), and within-stratum imbalances, thereby improving covariate balance beyond what is achievable with complete randomization (Hu et al., 2023).

Our theory and numerical studies cover three lines of CAR schemes. Classical stratified permuted-block randomization (e.g. STRPB) is effective when each stratum has ample sample size (Zelen, 1974; Hallstrom and Davis, 1988), but its performance can degrade as the number of covariates (or levels) increases and strata fragment. To balance a relatively large number of covariates covariates, minimization-based methods such as the Pocock and Simon's procedure (PS) (Taves, 1974; Pocock and Simon, 1975) targeting marginal imbalance directly, though its within-stratum imbalance may remain insufficiently controlled (Toorawa et al., 2009). Building on these ideas, Hu and Hu (2012) proposed extensions (HH) that simultaneously temper overall, marginal, and within-stratum imbalance. Comprehensive reviews of CAR methods and their asymptotic properties can be found in Rosenberger et al. (2008), Hu and Hu (2012), Hu et al. (2014), and Rosenberger and Lachin (2015).

2.3 Z-estimation, Wald statistic and Dunnett's test

Having established the design framework, we describe in detail the estimation and inference procedures that utilize the collected data. For each treatment arm k with $0 \le k \le K$, let $Y_i(k)$ denote the potential outcome for the i-th patient if assigned to the k-th treatment.

After treatment assignment $\boldsymbol{T}_i$, the observed outcome is

$$Y_i = \sum_{k=0}^{K} T_i^k Y_i(k).$$

To estimate the treatment effect vector $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_K)^\top$, we follow the Z-estimation framework by Van der Vaart (2000). Let $\boldsymbol{\theta} \in \mathbb{R}^{K+1+q}$ denote the full parameter vector, comprising a $(K+1)$-dimensional parameter of interest, on which $\boldsymbol{\delta}$ depends, and a q-dimensional nuisance parameter vector. Then the Z-estimator $\widehat{\boldsymbol{\theta}}$ is defined as the solution to the following $(K+1+q)$-dimensional estimating equation:

$$\sum_{k=0}^{K} \sum_{i=1}^{n} T_i^k \boldsymbol{\psi}^k(Y_i(k), \boldsymbol{X}_i; \boldsymbol{\theta}) = \mathbf{0}_{K+1+q}, \tag{2.1}$$

where for $k \in 0 \cup [K]$

$$\boldsymbol{\psi}^0(Y_i(0), \boldsymbol{X}_i; \boldsymbol{\theta}) := \begin{pmatrix} h^0(\boldsymbol{X}_i; \boldsymbol{\theta}) - Y_i(0) \\ \mathbf{0}_K \\ \boldsymbol{\xi}^0(Y_i(0), \boldsymbol{X}_i; \boldsymbol{\theta}) \end{pmatrix}, \quad \boldsymbol{\psi}^k(Y_i(k), \boldsymbol{X}_i; \boldsymbol{\theta}) := \begin{pmatrix} \mathbf{0}_k \\ h^k(\boldsymbol{X}_i; \boldsymbol{\theta}) - Y_i(k) \\ \mathbf{0}_{K-k} \\ \boldsymbol{\xi}^k(Y_i(k), \boldsymbol{X}_i; \boldsymbol{\theta}) \end{pmatrix}.$$

Here, $h^k(\boldsymbol{X}_i; \boldsymbol{\theta})$ denotes the working model for the estimation of the parameters and $\boldsymbol{\xi}^k(Y_i(k), \boldsymbol{X}_i; \boldsymbol{\theta})$ corresponds to additional estimation equations for the nuisance parameters. We define the targeted $\boldsymbol{\theta}^*$ as the solution to

$$\frac{1}{K+1} \sum_{k=0}^{K} \mathbb{E}[\boldsymbol{\psi}^k(Y(k), \boldsymbol{X}; \boldsymbol{\theta}^*)] = \mathbf{0}_{K+1+q}.$$

We further obtain $\widehat{\mu}_k$ and $\widehat{\boldsymbol{\delta}}$ by

$$\widehat{\mu}_k = \frac{1}{n} \sum_{i=1}^{n} h^k(\boldsymbol{X}_i; \widehat{\boldsymbol{\theta}}), \ k \in 0 \cup [K]; \quad \text{and} \quad \widehat{\delta}_k = g(\widehat{\mu}_k) - g(\widehat{\mu}_0), \quad k \in [K]. \tag{2.2}$$

This formulation offers a flexible framework for estimation and subsequent analysis in various settings. Let us take the generalized linear model (GLM) (McCullagh, 2019) with canonical

link $\gamma(\mu)$ as an example. Define $\boldsymbol{\theta} = (\iota_0, \cdots, \iota_K, \boldsymbol{\beta}_X^\top)^\top$ and we obtain $\widehat{\boldsymbol{\theta}}$ by taking $h^k(\boldsymbol{X}_i; \boldsymbol{\theta}) = \gamma^{-1}(\iota_k + \boldsymbol{X}_i^\top \boldsymbol{\beta}_X)$ and $\boldsymbol{\xi}^k(Y_i(k), \boldsymbol{X}_i; \boldsymbol{\theta}) = \{Y_i(k) - h^k(\boldsymbol{X}_i; \boldsymbol{\theta})\}\boldsymbol{X}_i$ in the estimating equation, and obtain $\widehat{\mu}_k$ and $\widehat{\delta}_k$ subsequently by (2.2). In a special case with $q = 0$ where we do not include any information from $\boldsymbol{X}$, the $\boldsymbol{\xi}^k(\cdot)$ no longer exists, $\boldsymbol{\theta} = (\iota_0, \cdots, \iota_K)$ and $h^k(\boldsymbol{\theta}) = \gamma^{-1}(\iota_k)$. Solving the estimating equations, we obtain the unadjusted estimators $\widehat{\boldsymbol{\delta}}_{\mathsf{unadj}} = (\widehat{\delta}_{1,\mathsf{unadj}}, \ldots, \widehat{\delta}_{K,\mathsf{unadj}})^\top$, where $\widehat{\delta}_{k,\mathsf{unadj}} = g(\bar{Y}_k) - g(\bar{Y}_0)$ with $\bar{Y}_k = n_k^{-1} \sum_{i=1}^n T_i^k Y_i$ denoting the average of outcome in the k-th treatment arm, for $k \in 0 \cup [K]$. To carry out the test for individual $H_{0,k}^1 : \delta_k = 0$ in Stage 1, we define

$$g'(x) = \frac{d}{dx} g(x), \quad \boldsymbol{G} = \left(-g'(\mu_0)\mathbf{1}_K, \ \ \mathrm{diag}\{g'(\mu_1), \ldots, g'(\mu_K)\}\right) \in \mathbb{R}^{K \times (K+1)},$$

and let $\boldsymbol{\Gamma}$ be the sandwich variance covariance based on (2.1). Then the model-robust Wald statistic is defined as

$$W_k^{\mathsf{robust}} = \frac{\widehat{\delta}_k}{\mathrm{se}_{\mathsf{robust}}(\widehat{\delta}_k)}, \ \text{where} \ \ \mathrm{se}^2_{\mathsf{robust}}(\widehat{\delta}_k) = \frac{1}{n}\left[\widehat{\boldsymbol{G}}\widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{G}}^\top\right]_{(k,k)},$$

with $\widehat{\boldsymbol{\Gamma}}$ being the empirical plug-in estimate of $\boldsymbol{\Gamma}$. To control the Type I error under the global null hypothesis $H_0^1 : \boldsymbol{\delta} = \mathbf{0}_K$, we compute Stage 1 p-value evaluated using Dunnett's test as

$$P_1 = 1 - \mathbb{P}(Z_1 \le \max_k W_k, \ldots, Z_K \le \max_k W_k), \ \text{with} \ (Z_1, \ldots, Z_K)^\top \sim N\left(\mathbf{0}, \widehat{\boldsymbol{R}}\right),$$

where $\widehat{\boldsymbol{R}}$ denotes the estimated correlation matrix derived from Theorem 6 and Remark 2, based on $\widehat{\boldsymbol{\Gamma}}$ and $\widehat{\boldsymbol{G}}$. Let P_2 denote the Stage 2 p-value corresponding to $H_0^2 : \delta_{k^*} = 0$. The overall test combines evidence across the two stages using the inverse–χ^2 method, which rejects H_0 if $-\log(P_1 P_2) > \frac{1}{2}\chi_4^2(1-\alpha)$, where $\chi_4^2(1-\alpha)$ is the $(1-\alpha)$-quantile of the chi-squared distribution with four degrees of freedom. In our design, the Stage 1 data are used exclusively for dose selection and for computing the Stage 1 p-value, while the Stage 2 data

are used solely for the confirmatory analysis. Because the two stages involve disjoint and independent data, both stage-wise p-values are valid and independent. Consequently, their combination via the inverse-χ^2 method controls the overall Type I error. Similar approaches have been adopted and validated in the seamless trial literature (Jenkins et al., 2011; Ma et al., 2022). Consider the logistic model for illustration, with $\text{logit}^{-1}(x) = [1+\exp(-x)]^{-1}$,

$$\mathbb{E}[Y_i \mid \boldsymbol{X}_i, \boldsymbol{T}_i] = \text{logit}^{-1}\left(\boldsymbol{\iota}^\top \boldsymbol{T}_i + \boldsymbol{X}_i^\top \boldsymbol{\beta}_X\right), \quad \boldsymbol{\iota} = (\iota_0, \ldots, \iota_K).$$

Here we take $\gamma(x) = \text{logit}(x)$, which yields

$$h^k(\boldsymbol{X}_i; \boldsymbol{\theta}) = \text{logit}^{-1}\left(\iota_k + \boldsymbol{X}_i^\top \boldsymbol{\beta}_X\right),$$

$$\boldsymbol{\xi}^k(Y_i(k), \boldsymbol{X}_i; \boldsymbol{\theta}) = \left[Y_i(k) - \text{logit}^{-1}\left(\iota_k + \boldsymbol{X}_i^\top \boldsymbol{\beta}_X\right)\right]\boldsymbol{X}_i.$$

For hypothesis testing, we set $g(x) = \log(x)$ when assessing log relative risks, and $g(x) = \log(x/(1-x))$ when evaluating log odds ratios.

3. Theoretical property

In this section, we first state the key assumptions and discuss their implications. We then present the main theoretical results, including the consistency of $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\mu}}$, the asymptotic normality of $\hat{\boldsymbol{\mu}}$ and $g(\hat{\boldsymbol{\mu}})$, and, finally, the asymptotic joint distribution of the Stage 1 test statistics, which establishes the validity of our model-robust tests.

Assumption 1 (Data-generating process)**.**

1. $\{(\boldsymbol{X}_i, Y_i(0), \ldots, Y_i(K)\}_{i=1}^n$ *are independently identically distributed as* $(\boldsymbol{X}, Y(0), \ldots, Y(K))$*, with* $\mathbb{E}\left[\|\boldsymbol{X}\|^2\right] < \infty$*,* $\max_{k \in 0 \cup [K]} \text{Var}[Y(k)] < \infty$*.*

2. *Let* π_s *denote the proportion of observations in stratum* s*. For any* $s \in [s_{\max}]$*, we have* $0 < \pi_s < 1$*, and* $\sum_{s \in [s_{\max}]} \pi_s = 1$*.*

3. *Recall that* $\mathcal{F}_{i-1} = \mathcal{X}_{i,\mathrm{ex}} \otimes \mathcal{T}_{i-1} \otimes \mathcal{X}_{i-1} \otimes \mathcal{Y}_{i-1}$ *and the assignment rule satisfies* $\boldsymbol{\phi}_i = \mathbb{E}[\boldsymbol{T}_i|\mathcal{F}_{i-1}]$ *for* $i \in [n]$. *Let us define the imbalance within stratum* s *and treatment arm* k *as* $D_n^k(s)$ *with the following decomposition*

$$\begin{aligned} D_n^k(s) &:= \sum_{i=1}^{n} \left(T_i^k - \frac{1}{K+1} \right) \mathbb{I}\{S_i = s\} \\ &= \sum_{i=1}^{n} M_i^k(s) + d_n^k(s), \quad s \in [s_{\max}], \quad k \in 0 \cup [K]. \end{aligned}$$

We assume that the CAR procedure satisfies

(a) $\mathbb{E}[\{d_n^k(s)\}^2] = o(n)$; *and*

(b) $\{M_i^k(s)\}_{i=1}^n$ *is a sequence of bounded zero-mean martingale differences with respect to* $\mathcal{F}_{i-1}$. *Further with vectorized*

$$\boldsymbol{M}_i := \mathrm{vec}\Big(\big(M_i^k(s)\big)_{s\in[s_{\max}],\, k\in[K]} \Big) \in \mathbb{R}^{s_{\max}(K+1)},$$

the averaged conditional second moment converges in probability to a block covariance matrix:

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big[\boldsymbol{M}_i \boldsymbol{M}_i^\top \,\big|\, \mathcal{F}_{i-1}\big] \xrightarrow{p} \boldsymbol{\Sigma}^{\mathsf{CAR}} \in \mathbb{R}^{s_{\max}(K+1)\times s_{\max}(K+1)}.$$

If we index $\boldsymbol{\Sigma}^{\mathsf{CAR}}$ *by* $(s, s') \in [s_{\max}] \times [s_{\max}]$ *as a block matrix:*

$$\boldsymbol{\Sigma}^{\mathsf{CAR}} = \big[\, \boldsymbol{\Sigma}^{\mathsf{CAR}}(s,s') \,\big]_{s,s'\in[s_{\max}]}, \qquad \boldsymbol{\Sigma}^{\mathsf{CAR}}(s,s') \in \mathbb{R}^{(K+1)\times(K+1)},$$

and defined the (k, k')*-th element in* $\boldsymbol{\Sigma}^{\mathsf{CAR}}(s, s')$ *as* $\boldsymbol{\Sigma}^{\mathsf{CAR}}_{kk'}(s, s')$, *then*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[M_i^k(s) M_i^{k'}(s') \mid \mathcal{F}_{i-1}] \xrightarrow{p} \boldsymbol{\Sigma}^{\mathsf{CAR}}_{kk'}(s, s').$$

Assumption 1.3 is designed to cover the covariate balance properties induced by several commonly used CAR procedures; see Hu et al. (2023) for their formal definitions. Under

STRPB with a fixed block size compatible with the target allocation ratio, treatment counts are exactly balanced within each completed block in each stratum. Hence, the within-stratum imbalance can only arise from the current incomplete block and is uniformly bounded in n. Consequently, Assumption 1.3 holds by taking the martingale component to be degenerate, so that the remainder satisfies $\mathbb{E}\{d_n^k(s)^2\} = O(1) = o(n)$. For HH, if the allocation rule favors assignments that reduce the imbalance criterion and the criterion includes a positive within-stratum weight, then within-stratum imbalances remain bounded in probability. Since marginal and overall imbalances are linear combinations of within-stratum imbalances, they are also bounded. Thus Assumption 1.3 is satisfied with no nontrivial $\sqrt{n}$-scale martingale fluctuation. PS corresponds to the marginal-balancing case: it yields bounded marginal and overall imbalances, but its within-stratum imbalances are generally $O_p(n^{1/2})$. Assumption 1.3 accommodates this latter case through the martingale decomposition of the within-stratum imbalance, with Σ^{CAR} representing the corresponding design-induced limiting covariance. Assumption **??** in the Supplementary Material outlines standard conditions for Z-estimation, analogous to those in Wang et al. (2023) and Van der Vaart (2000), which ensure the consistency and asymptotic convergence of the estimator $\widehat{\boldsymbol{\theta}}$ to its population counterpart $\boldsymbol{\theta}^*$.

Proposition 1 (Consistency of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\mu}}$)**.** *Under Assumptions 1 and* **??**,

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* = O_p(n^{-1/2}), \qquad \widehat{\boldsymbol{\mu}} - \boldsymbol{\mu} = O_p(n^{-1/2}).$$

Let $\boldsymbol{\Sigma}^{\mathsf{CR}}$ be the counterpart of $\boldsymbol{\Sigma}^{\mathsf{CAR}}$ under complete randomization. Beyond the consistency provided in Proposition 5, in the following theorem, we derive the asymptotic normality of $\widehat{\boldsymbol{\mu}}$ and provide a clear decomposition of the limiting covariance structure to reflect the influence of covariate adjustment in the Z-estimation and the CAR procedure.

Theorem 2 (Asymptotic normality of $\widehat{\boldsymbol{\mu}}$)**.** *Under Assumptions 1 and* **??**, $\widehat{\boldsymbol{\mu}}$ *in (2.2) satisfies:*

$$\sqrt{n}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Gamma}), \tag{3.3}$$

with $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}_{\mathsf{base}} - \boldsymbol{\Gamma}_{\mathsf{adj}} - \boldsymbol{\Gamma}_{\mathsf{CAR,adj}}$*. Here,*

- $\boldsymbol{\Gamma}_{\mathsf{base}} = (K+1)\mathrm{diag}\left\{\mathrm{Var}\left[Y(0)\right], \ldots, \mathrm{Var}\left[Y(K)\right]\right\}$ *represents the baseline covariance matrix that does not account for covariate adjustment or the CAR procedures.*

- *With* $r^k(\boldsymbol{\theta}^*) = Y(k) - h^k(\boldsymbol{X}, \boldsymbol{\theta}^*)$ *and* $\boldsymbol{r} = (r^0(\boldsymbol{\theta}^*), \ldots, r^K(\boldsymbol{\theta}^*))^\top$,

$$\begin{aligned}\boldsymbol{\Gamma}_{\mathsf{adj}} = (K+1)\mathrm{diag}\left\{\mathrm{Var}\left[Y(0)\right] - \mathrm{Var}\left[r^0(\boldsymbol{\theta}^*)\right], \ldots, \mathrm{Var}\left[Y(K)\right] - \mathrm{Var}\left[r^K(\boldsymbol{\theta}^*)\right]\right\} \\ + \mathrm{Var}\left[\boldsymbol{r}\right] - \mathrm{Var}\left[\boldsymbol{Y}\right]\end{aligned}$$

  captures the contribution of covariate adjustment in Z-estimation.

- *With* $\boldsymbol{L}(s,k) = \mathbb{E}\left[r^k(\boldsymbol{\theta}^*) \mid S = s\right]\boldsymbol{e}_k$*, recall the definition of* $\boldsymbol{\Sigma}^{\mathsf{CAR}}_{kk'}(s,s')$ *from Assumption 1,*

$$\boldsymbol{\Gamma}_{\mathsf{CAR,adj}} = (K+1)^2 \sum_{k,k' \in 0 \cup [K]} \sum_{s,s' \in [s_{\max}]} \{\boldsymbol{\Sigma}^{\mathsf{CR}}_{kk'}(s,s') - \boldsymbol{\Sigma}^{\mathsf{CAR}}_{kk'}(s,s')\}\boldsymbol{L}(s,k)\boldsymbol{L}(s',k')^\top$$

  depicts the interplay between covariate adjustment in Z-estimation and the CAR procedures.

Corollary 3 (Special cases)**.**

1. *Under CR, Assumption 1 holds with* $\boldsymbol{\Sigma}^{\mathsf{CR}}_{kk'}(s,s') = \boldsymbol{\Sigma}^{\mathsf{CAR}}_{kk'}(s,s')$ *for* $s, s' \in [s_{\max}]$ *and* $k, k' \in 0 \cup [K]$*, and* $\boldsymbol{\Gamma}_{\mathsf{CAR,adj}} = 0$*.*

2. *When* $q = 0$ *in the Z-estimation,* $\boldsymbol{\Gamma}_{\mathsf{adj}} = 0$*.*

3. *Under HH or STRPB,* $\boldsymbol{\Gamma}_{\mathsf{CAR,adj}}$ *is positive definite as* $\boldsymbol{\Sigma}^{\mathsf{CAR}}_{kk'}(s,s') = 0$ *for all* $k, k' \in 0 \cup [K]$ *and* $s, s' \in [s_{\max}]$*.*

In contrast, under PS , $\boldsymbol{\Sigma}^{\mathsf{CAR}}_{kk'}(s,s')$ does not have a closed-form and thus may not guarantee the positive definiteness of $\boldsymbol{\Gamma}_{\mathsf{CAR,adj}}$.

Remark 1. It is worth emphasizing that the simplest and widely used estimation procedure neither adjusts for covariates nor the impact of CAR. This corresponds to the unadjusted estimator $\widehat{\boldsymbol{\mu}}_{\mathsf{unadj}}$ and its asymptotic covariance matrix $\boldsymbol{\Gamma}_{\mathsf{unadj}} = \boldsymbol{\Gamma}_{\mathsf{base}}$. By contrast, for model-based methods, while they incorporate covariate adjustment in estimation, the conventional calculation of the asymptotic covariance still ignores the influence of CAR, effectively treating $\boldsymbol{\Gamma}_{\mathsf{CAR,adj}} = 0$. In this case, the estimator is $\widehat{\boldsymbol{\mu}}$, with $\boldsymbol{\Gamma}_{\mathsf{conv}} = \boldsymbol{\Gamma}_{\mathsf{base}} - \boldsymbol{\Gamma}_{\mathsf{adj}}$. Our proposed model-robust method fully accounts for both covariate adjustment and CAR, yielding the same estimator $\widehat{\boldsymbol{\mu}}$ but with asymptotic covariance matrix $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}_{\mathsf{base}} - \boldsymbol{\Gamma}_{\mathsf{adj}} - \boldsymbol{\Gamma}_{\mathsf{CAR,adj}}$. More specifically, CAR procedures, such as HH or STRPB designs, improve covariate balance and guarantee that $\boldsymbol{\Sigma}^{\mathsf{CAR}}_{kk'}(s,s') = 0$ for all $k,k' \in \{0\} \cup [K]$ and $s,s' \in [s_{\max}]$, thereby further ensuring the positive definiteness of $\boldsymbol{\Gamma}_{\mathsf{CAR,adj}}$. A systematic comparison of the three approaches will be presented in Section 5 and 6. Lastly, when the working model used for estimation and inference is correct, $\boldsymbol{\Gamma}_{\mathsf{CAR,adj}} = 0$ and thus the results from CR and CAR procedures are asymptotically the same.

Theorem 4 (Asymptotic normality of the general form $\widehat{\boldsymbol{\delta}}$)**.** *Consider the parameter of interest* $\boldsymbol{\delta} = (g(\mu_1) - g(\mu_0), \ldots, g(\mu_K) - g(\mu_0))^\top$*, and denote by* $\widehat{\boldsymbol{\delta}}$ *its estimator obtained by replacing* $\boldsymbol{\mu}$ *with* $\widehat{\boldsymbol{\mu}}$*. Then under Assumptions 1 and* **??***,*

$$\sqrt{n}(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{G}\boldsymbol{\Gamma}\boldsymbol{G}^\top), \tag{3.4}$$

where $\boldsymbol{G} = \frac{\partial \boldsymbol{\delta}}{\partial \boldsymbol{\mu}}$ *is the* $K \times (K+1)$ *Jacobian matrix of* $\boldsymbol{\delta}$ *evaluated at* $\boldsymbol{\mu}$*, and* $\boldsymbol{\Gamma}$ *is defined in Theorem 2.*

To construct both the conventional and model-robust test statistics for inference on $\boldsymbol{\delta}$, we require an estimate of $\boldsymbol{G}\boldsymbol{\Gamma}\boldsymbol{G}^\top$. We first evaluate the Jacobian of $\boldsymbol{\delta}$ at $\widehat{\boldsymbol{\mu}}$ to obtain $\widehat{\boldsymbol{G}}$. Recall that

$$\boldsymbol{\Gamma} = \boldsymbol{\Gamma}_{\mathsf{base}} - \boldsymbol{\Gamma}_{\mathsf{adj}} - \boldsymbol{\Gamma}_{\mathsf{CAR,adj}} = \boldsymbol{\Gamma}_{\mathsf{conv}} - \boldsymbol{\Gamma}_{\mathsf{CAR,adj}},$$

where $\widehat{\boldsymbol{\Gamma}}_{\mathsf{conv}} := \widehat{\boldsymbol{\Gamma}}_{\mathsf{base}} - \widehat{\boldsymbol{\Gamma}}_{\mathsf{adj}}$ can be estimated directly using plug-in estimators. By contrast, estimating $\boldsymbol{\Gamma}_{\mathsf{CAR,adj}}$ is more involved, as $\boldsymbol{\Sigma}^{\mathsf{CAR}}_{kk'}(s, s')$ does not admit a closed form under PS. Following Liu and Hu (2023), we therefore employ a bootstrap procedure to estimate $\boldsymbol{\Sigma}^{\mathsf{CAR}}$ and subsequently construct $\widehat{\boldsymbol{\Gamma}}_{\mathsf{CAR,adj}}$. The resulting estimator is $\widehat{\boldsymbol{\Gamma}} = \widehat{\boldsymbol{\Gamma}}_{\mathsf{conv}} - \widehat{\boldsymbol{\Gamma}}_{\mathsf{CAR,adj}}$. Additional details on the construction and asymptotic properties are provided in Section **??** of the Supplementary Material. Finally, the standard error estimate of $\widehat{\delta}_k$ is obtained as the square root of the kth diagonal element of the estimated variance–covariance matrix of $\widehat{\boldsymbol{\delta}}$:

$$\mathrm{se}^2_{\mathsf{conv}}(\widehat{\delta}_k) = n^{-1}[\widehat{\boldsymbol{G}}\,\widehat{\boldsymbol{\Gamma}}_{\mathsf{conv}}\,\widehat{\boldsymbol{G}}^\top]_{(k,k)}, \qquad \mathrm{se}^2_{\mathsf{robust}}(\widehat{\delta}_k) = n^{-1}[\widehat{\boldsymbol{G}}\,\widehat{\boldsymbol{\Gamma}}\,\widehat{\boldsymbol{G}}^\top]_{(k,k)}.$$

Proposition 5 (The consistency of $\widehat{\boldsymbol{\Gamma}}_{\mathsf{conv}}$ and $\widehat{\boldsymbol{\Gamma}}$)**.** *Under Assumptions 1 and* **??**, $\widehat{\boldsymbol{\Gamma}}_{\mathsf{conv}} \xrightarrow{p} \boldsymbol{\Gamma}_{\mathsf{conv}}$ *and* $\widehat{\boldsymbol{\Gamma}} \xrightarrow{p} \boldsymbol{\Gamma}$.

Theorem 6 (Comparison between conventional and model-robust inference)**.** *Under Assumptions 1 and* **??**, *we consider the individual hypothesis test* $H_0 : \delta_k = 0$ *versus* $H_1 : \delta_k > 0$. *Define the individual conventional and model-robust Wald test statistics for* $k \in [K]$ *as*

$$W_k^{\mathsf{conv}} = \frac{\sqrt{n}\,\widehat{\delta}_k}{\sqrt{[\widehat{\boldsymbol{G}}\,\widehat{\boldsymbol{\Gamma}}_{\mathsf{conv}}\,\widehat{\boldsymbol{G}}^\top]_{(k,k)}}}, \quad \textit{and} \quad W_k^{\mathsf{robust}} = \frac{\sqrt{n}\,\widehat{\delta}_k}{\sqrt{[\widehat{\boldsymbol{G}}\,\widehat{\boldsymbol{\Gamma}}\,\widehat{\boldsymbol{G}}^\top]_{(k,k)}}}.$$

W_k^{robust} *always produces a valid type I error by Theorem 4 and Proposition 5. Because* $\boldsymbol{\Gamma}_{\mathsf{conv}} - \boldsymbol{\Gamma}$ *is positive definite under HH and STRPB, the conventional statistic* W_k^{conv} *exhibits a reduced type I error under these procedures.*

Theorem 7 (Asymptotic joint distribution of the $\{W_k^{\mathsf{robust}}\}_{k\in[K]}$)**.** *Under Assumptions 1 and* **??***, we aim to conduct the following hypothesis test*

$$H_0^1 : \delta_1 = \delta_2 = \cdots = \delta_K = 0,$$

$$H_1^1 : \exists k \in [K], \delta_k > 0.$$

In particular, $(W_1^{\mathsf{robust}}, \ldots, W_K^{\mathsf{robust}})^\top \xrightarrow{d} N(\mathbf{0}, \boldsymbol{R})$ *with*

$$\boldsymbol{R}_{(k,k')} = \frac{[\boldsymbol{G}\boldsymbol{\Gamma}\boldsymbol{G}^\top]_{(k,k')}}{\sqrt{[\boldsymbol{G}\boldsymbol{\Gamma}\boldsymbol{G}^\top]_{(k,k)}[\boldsymbol{G}\boldsymbol{\Gamma}\boldsymbol{G}^\top]_{(k',k')}}}, \quad k, k' \in [K].$$

Remark 2. The plug-in estimator $\widehat{\boldsymbol{R}}$ is constructed by replacing $\boldsymbol{G}$ and $\boldsymbol{\Gamma}$ with their empirical counterparts $\widehat{\boldsymbol{G}}$ and $\widehat{\boldsymbol{\Gamma}}$. The Stage 1 p-value is then computed using Dunnett's test and combined with the Stage 2 p-value through the inverse–χ^2 method to obtain the overall test decision, as detailed in Section 2.3. By Theorem 7, each individual statistic W_k^{robust} maintains valid Type I error control; consequently, the combined procedures using Dunnett's test and the inverse–χ^2 method also preserve overall Type I error validity. However, since W_k^{conv} tends to be conservative, the joint procedure may no longer guarantee exact Type I error control.

4. Application to important estimands

In clinical and epidemiological studies with binary outcomes and multiple treatment arms, it is often necessary to report treatment effects on scales that are both scientifically interpretable and compatible with standard modeling. Our framework is designed to be flexible: by specifying an appropriate $g(\cdot)$, we link $\boldsymbol{\mu} = (\mu_0, \mu_1, \ldots, \mu_K)^\top$ to $\boldsymbol{\delta}$ through the relation $\delta_k = g(\mu_k) - g(\mu_0)$. It encompasses three special cases that are particularly important in applications, log relative risk (logRR), log odds ratio (LOR) and average treatment effect

(ATE). Each arises from a distinct choice of g yet is treated within a unified asymptotic framework.

We proceed by introducing each treatment effect measure $\boldsymbol{\delta}$ together with their associated g and the Jacobian required by Theorem 4. Throughout, $\widehat{\boldsymbol{\mu}}$ denotes plug-in estimators obtained from a working model (e.g., logistic regression), and $\boldsymbol{\Gamma}$ denotes the asymptotic covariance of $\sqrt{n}\widehat{\boldsymbol{\mu}}$.

4.1 Log relative risk (log**RR**)

Log relative risk provides a multiplicative comparison standard in epidemiology and risk communication. For treatment arm k,

$$\mathrm{RR}_k = \frac{\mu_k}{\mu_0}, \qquad \text{and} \quad \boldsymbol{\delta} = \big(\log\mu_1 - \log\mu_0, \ldots, \log\mu_K - \log\mu_0\big)^\top.$$

The log transformation stabilizes variability and ensures additivity in regression. A convenient working model is the logistic regression

$$\mathbb{E}[Y_i \mid \boldsymbol{X}_i, \boldsymbol{T}_i] = \mathrm{logit}^{-1}\big(\boldsymbol{\tau}^\top \boldsymbol{T}_i + \boldsymbol{\beta}_X^\top \boldsymbol{X}_i\big)\,,$$

and valid inference for $\log(\mathrm{RR}_k)$ follows from delta method and Theorem 4.

Proposition 8. *When* log*RR is the target,* $g(x) = \log x$ *and*

$$\boldsymbol{G}_{\mathsf{RR}} = \big(-\mu_0^{-1}\mathbf{1}_K,\ \ \mathrm{diag}\{\mu_1^{-1}, \ldots, \mu_K^{-1}\}\big), \qquad \widehat{\boldsymbol{G}}_{\mathsf{RR}} = \boldsymbol{G}_{\mathsf{RR}}\big|_{\mu=\widehat{\mu}}.$$

Thus Theorem 4 implies

$$\sqrt{n}\left(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}\right) \xrightarrow{d} N\big(0,\ \boldsymbol{G}_{\mathsf{RR}}\,\boldsymbol{\Gamma}\,\boldsymbol{G}_{\mathsf{RR}}^\top\big)\,.$$

4.2 Log Odds Ratio (LOR)

The LOR is a natural scale for expressing effects in logistic regression, where model coefficients correspond to log-odds differences. For arm k,

$$\mathrm{LOR}_k = \log\left(\frac{\mu_k}{1-\mu_k}\right) - \log\left(\frac{\mu_0}{1-\mu_0}\right),$$

$$\boldsymbol{\delta} = \left(\log(\tfrac{\mu_1}{1-\mu_1}) - \log(\tfrac{\mu_0}{1-\mu_0}), \ldots, \log(\tfrac{\mu_K}{1-\mu_K}) - \log(\tfrac{\mu_0}{1-\mu_0})\right)^\top.$$

Estimation utilizes the fitted logistic model, followed by the transformation of $\widehat{\mu}_k$ and calculation of variance using delta method.

Proposition 9. *For LOR,* $g(x) = \log\big(x/(1-x)\big)$,

$$\boldsymbol{G}_{\mathsf{LOR}} = \left([(\mu_0-1)^{-1} - \mu_0^{-1}]\mathbf{1}_K,\ \ \mathrm{diag}\{\mu_1^{-1} - (\mu_1-1)^{-1}, \ldots, \mu_K^{-1} - (\mu_K-1)^{-1}\}\right)$$

and $\widehat{\boldsymbol{G}}_{\mathsf{LOR}} = \boldsymbol{G}_{\mathsf{LOR}}\big|_{\mu=\widehat{\mu}}$. *Consequently Theorem 4 implies*

$$\sqrt{n}\left(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}\right) \xrightarrow{d} N\big(0,\ \boldsymbol{G}_{\mathsf{LOR}}\,\boldsymbol{\Gamma}\,\boldsymbol{G}_{\mathsf{LOR}}^\top\big).$$

Details on ATE are provided in Section **??** of the Supplementary Material. Taken together, the logRR, LOR and ATE specifications illustrate how our unified inferential mechanism yields effect estimates and valid large-sample inference across the three scales most commonly reported in practice, while maintaining a common theoretical and computational pipeline.

5. Numerical studies

We conducted extensive simulation studies to examine the finite-sample performance of the proposed testing procedures. The simulations are structured as follows: we first outline the general setup common to all experiments, and then present two examples based on

distinct data-generating models: M1 (logistic) and M2 (probit). Within each example, we consider three working models ($\mathcal{A}_0$, $\mathcal{A}_1$ and $\mathcal{A}_2$), ranging from unadjusted to fully adjusted specifications, and compare the conventional Wald test with the new model-robust Wald test. We investigate four randomization procedures: complete randomization (CR), stratified permuted block design (STRPB), Pocock–Simon procedure (PS), and Hu and Hu's procedure (HH). Randomization is performed using two baseline covariates, X_1 and X_2, which will be specified later. For continuous covariates, values are discretized as 0 for negative values and 1 for positive values.

We consider a two-stage seamless trial design. In Stage 1, a total of 420 patients are sequentially enrolled and randomized among $K = 2$ experimental arms and a control arm. For each experimental arm $k = 1, \ldots, K$, we compute a test statistic of the following form (both the conventional and model-robust versions, which differ in the specification of se, using $\mathbf{\Gamma}_{\mathsf{conv}}$ and $\mathbf{\Gamma}$ respectively):

$$W_k = \frac{\widehat{\delta}_k}{\mathrm{se}(\widehat{\delta}_k)}.$$

The treatment arm with the largest statistic W_k (denoted $k^\star$) is selected as the most promising and advanced to Stage 2. In Stage 2, an additional 500 patients are enrolled and randomized between control and treatment $k^\star$. At the end of the study, we test the global null hypothesis $H_0 : \boldsymbol{\delta} = \mathbf{0}_K$ by combining the stage-wise p-values from Stage 1 (obtained via Dunnett's test) and Stage 2 using the inverse χ^2 combination method (Bauer and Kohne, 1994). We evaluate treatment effects using (logRR), LOR, and ATE. All simulation summaries are based on 10,000 replicates.

5.1 Example 1: logistic model (M1)

The data are generated from the logistic regression model, where $Y_i \sim \text{Bernoulli}(p_i)$ with $p_i = 1/(1+\exp(-\eta_i)) = \text{logit}^{-1}(\eta_i)$, and

$$\eta_i = \beta_0 + \iota_1 T_i^1 + \iota_2 T_i^2 + \beta_1 X_{i,1} + \beta_2 X_{i,2}.$$

Covariates are generated as $X_{i,1} \sim \text{Bernoulli}(1/2)$ and $X_{i,2} \sim \mathcal{N}(0,1)$. Parameters are set to $(\beta_0, \beta_1, \beta_2) = (-1, 1, 2)$. We evaluate three working models for estimation and inference with $p_i \sim \text{logit}^{-1}(\eta_i)$, i.e., $\gamma(\cdot) = \text{logit}(\cdot)$:

$\mathcal{A}_0$ No covariates adjustment: $\eta_i = \beta_0 + \iota_1 T_i^1 + \iota_2 T_i^2$.

$\mathcal{A}_1$ Adjust for $X_{i,1}$ only: $\eta_i = \beta_0 + \iota_1 T_i^1 + \iota_2 T_i^2 + +\beta_1 X_{i,1}$.

$\mathcal{A}_2$ Adjust for both $X_{i,1}$ and $X_{i,2}$: $\eta_i = \beta_0 + \iota_1 T_i^1 + \iota_2 T_i^2 + \beta_1 X_{i,1} + \beta_2 X_{i,2}$.

Thus $\mathcal{A}_0$ and $\mathcal{A}_1$ are misspecified working models, while $\mathcal{A}_2$ is correct.

Table **??** in the Supplementary Material reports the conventional Wald test statistics computed without adjustment for CAR, across complete randomization and the three CAR procedures under different working models. As the impact of CAR is not accounted for, the standard errors are systematically inflated for all CAR schemes in the misspecified models $\mathcal{A}_0$ and $\mathcal{A}_1$; as a result, the nominal 5% Wald tests become conservative (Type I error < 0.05). By contrast, CR is unaffected and has valid Type I errors because it does not use covariates in assignment. Moreover, consistent with expectations, the correctly specified working model $\mathcal{A}_2$ leads to valid Type I errors and yields smaller sampling variability for the estimators than the misspecified models $\mathcal{A}_0$ and $\mathcal{A}_1$. Finally, combining the asymptotically independent p-values from Stage 1 and 2 via the inverse–χ^2 method preserves overall type I error control provided each phase maintains valid type I error.

Using the logRR as a representative estimand, Table 1 contrasts the conventional Wald tests (ignoring the impact of CAR) with our model-robust Wald tests. The findings are as follows: (i) The model-robust standard error estimates properly account for the CAR schemes, yielding valid 5% type I error across all working models, while the conventional tests are conservative in misspecified models $\mathcal{A}_0$, $\mathcal{A}_1$. Accordingly, under the alternative hypothesis, the model-robust tests achieve substantially higher power, showing gains of at least 50% under $\mathcal{A}_0$ and 35% under $\mathcal{A}_1$ across all CAR procedures. (ii) Under the correctly specified working model $\mathcal{A}_2$, we have $\mathbf{\Gamma}_{\mathsf{CAR,adj}} = 0$ asymptotically (cf. Remark 1). Consequently, CR and the CAR procedures exhibit comparable performance. For the same reason, the conventional Wald test, which ignores $\mathbf{\Gamma}_{\mathsf{CAR,adj}}$ when constructing $\mathbf{\Gamma}$, performs similarly to the model-robust Wald test. (iii) Under misspecification (e.g., $\mathcal{A}_0$ and $\mathcal{A}_1$), residuals retain covariate signal; better covariate balance improves covariance estimation by Theorem 2, so CAR, by achieving a superior strata balance, outperforms CR in power. (iv) The model-robust standard errors are computed using the sandwich estimator, which does not rely on a particular generative model. This yields more accurate estimation and greater robustness, even when the working model is misspecified. Similar observations are found with LOR and ATE, with results summarized in Table **??** and **??** in the Supplementary Material.

To further illustrate our advantage, we vary (ι_1, ι_2) systematically. Starting from $(0, 0)$, we sequentially increase them to $(0, 0.1)$, $(0.1, 0.2)$, and so on, up to $(0.9, 1)$. Let $\iota_{\max} = \max(\iota_1, \iota_2)$, and Figure 1 displays the statistical power of the conventional and model-robust tests for logRR under different randomization schemes as $\iota_{\max}$ increases. When the working models are misspecified (e.g., under $\mathcal{A}_0$ and $\mathcal{A}_1$), the model-robust tests achieve substantial power gains over their conventional counterparts. Although CR maintains valid Type I error

control when $\iota_{\max} = 0$, its power remains lower than that of the model-robust tests. Results on LOR and ATE are shown in Figure **??** and **??**, which convey similar findings.

Table 1: logRR results for Example 1, the logistic model: Type I error rates and power (in 10^{-2}) for Stage 1, 2, and the combined analysis, across different working models and randomization schemes.

	Procedure	Test	Type I			Power		
			Stage 1	Stage 2	All	Stage 1	Stage 2	All
			$(\iota_1, \iota_2) = (0, 0)$			$(\iota_1, \iota_2) = (0.3, 0.4)$		
	CR		5.15	5.02	5.11	21.72	29.62	38.61
	STRPB	conv	1.76	2.41	1.53	14.69	26.29	31.73
		robust	4.82	4.88	4.89	28.29	39.10	50.82
$\mathcal{A}_0$	HH	conv	1.62	2.67	1.60	15.80	26.30	32.12
		robust	4.99	5.42	5.15	29.53	39.34	51.67
	PS	conv	1.76	2.23	1.61	15.39	26.42	32.34
		robust	4.76	4.86	4.71	28.29	39.16	50.72
	CR		5.13	5.05	4.85	23.68	31.71	41.68
	STRPB	conv	2.23	2.84	2.01	17.64	29.11	35.92
		robust	4.82	4.93	4.91	28.23	39.16	50.91
$\mathcal{A}_1$	HH	conv	2.17	3.21	2.15	18.46	29.33	36.58
		robust	5.05	5.38	5.08	29.57	39.41	51.56
	PS	conv	2.37	2.88	2.07	18.16	29.63	36.86
		robust	4.86	4.84	4.77	28.37	39.15	50.72
	CR		5.49	5.22	5.14	33.06	43.64	57.79
	STRPB	conv	4.84	4.70	4.62	31.38	44.30	56.64
		robust	4.96	4.71	4.66	31.72	44.41	56.91
$\mathcal{A}_2$	HH	conv	5.00	5.23	4.96	32.27	43.93	57.72
		robust	5.13	5.24	5.05	32.56	43.95	58.03
	PS	conv	4.87	4.97	4.95	31.80	43.71	57.87
		robust	4.92	4.99	4.97	31.96	43.75	57.99

5.2 Example 2: probit model (M2)

The data are generated from the probit regression model, where $Y_i \sim \text{Bernoulli}(p_i)$ with $p_i = \Phi(\eta_i)$, $\Phi(\cdot)$ being the standard normal CDF and

$$\eta_i = \beta_0 + \iota_1 T_i^1 + \iota_2 T_i^2 + \beta_1 X_{i,1} X_{i,2} + \beta_2 \exp(X_{i,1} + X_{i,2}) + \beta_3 X_{i,3}.$$

Covariates are generated as $X_{i,1}, X_{i,2}, X_{i,3} \sim \mathcal{N}(0, 1)$ and parameters are set to $(\beta_0, \beta_1, \beta_2, \beta_3) = (-1, 1, 1, 0.5)$. We consider three working models for analysis:

$\mathcal{A}_0$ No covariates adjustment: $\eta_i = \beta_0 + \iota_1 T_i^1 + \iota_2 T_i^2$.

$\mathcal{A}_1$ Adjust for $X_{i,1}$ and $X_{i,2}$: $\eta_i = \beta_0 + \iota_1 T_i^1 + \iota_2 T_i^2 + \beta_1 X_{i,1} + \beta_2 X_{i,2}$.

$\mathcal{A}_2$ Adjust for $X_{i,1}$, $X_{i,2}$, and $X_{i,3}$: $\eta_i = \beta_0 + \iota_1 T_i^1 + \iota_2 T_i^2 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3}$.

The key distinction between Examples 1 and 2 is that, in Example 2, all three working models $\mathcal{A}_0$, $\mathcal{A}_1$, and $\mathcal{A}_2$ are misspecified. As a result, $\mathbf{\Gamma}_{\mathsf{CAR,adj}}$ is nonzero, and ignoring it, as

in the conventional Wald test, produces inflated standard error estimates and, consequently, conservative type I error rates across all working models. By contrast, our model-robust procedure does not depend on the correctness of the working model. Its performance is therefore far less affected by misspecification: type I errors remain close to the nominal level, and power is consistently higher than that of the conventional Wald test. These findings are confirmed by the empirical results reported in Tables **??** and 2; for instance, under $\mathcal{A}_0$, power increases by at least 10% across different CAR procedures. The phenomenon is further illustrated by varying (ι_1, ι_2) in a manner similar to that in Example 1, with results summarized in Figure **??**. Notably, in the working model $\mathcal{A}_2$, although all covariates are included, the model is still misspecified; as a result, the model-robust tests achieve substantial power gains over their conventional counterparts. Similar observations are found with LOR and ATE, with results summarized in Figure **??** and **??**.

Table 2: logRR results for Example 2, the probit model: Type I error rates and power (in 10^{-2}) for Stage 1, Stage 2, and the combined analysis, across different working models and randomization schemes.

	Procedure	**Test**	**Type I**			**Power**		
			Stage 1	**Stage 2**	**All**	**Stage 1**	**Stage 2**	**All**
			$(\iota_1, \iota_2) = (0, 0)$			$(\iota_1, \iota_2) = (0.2, 0.3)$		
$\mathcal{A}_0$	CR		4.92	5.16	5.13	26.68	37.38	48.64
	STRPB	conv	2.59	3.16	2.47	23.92	36.03	46.90
		robust	4.82	4.80	5.07	31.90	43.50	56.95
	HH	conv	2.68	3.18	2.58	22.99	35.74	46.81
		robust	4.66	4.89	4.85	31.13	42.55	56.06
	PS	conv	3.63	3.78	3.72	25.29	36.89	48.24
		robust	4.86	5.21	4.95	29.08	41.35	53.33
$\mathcal{A}_1$	CR		5.30	5.16	5.54	30.28	41.43	54.43
	STRPB	conv	3.75	4.15	3.85	29.17	41.73	53.76
		robust	5.11	4.98	5.14	33.58	45.44	58.95
	HH	conv	3.67	3.81	3.64	27.83	41.32	53.88
		robust	4.96	4.83	4.88	32.44	44.91	58.72
	PS	conv	5.21	4.97	4.84	30.71	43.01	55.71
		robust	5.10	4.99	4.84	30.39	42.95	55.49
$\mathcal{A}_2$	CR		5.29	5.31	5.55	32.33	43.47	56.41
	STRPB	conv	3.51	4.10	3.62	30.21	44.20	56.41
		robust	4.88	5.03	4.93	35.49	48.26	61.75
	HH	conv	3.79	3.90	3.76	29.77	43.75	56.62
		robust	5.02	4.88	4.97	34.78	47.58	61.68
	PS	conv	5.15	4.93	4.96	32.11	45.11	57.89
		robust	5.06	4.96	4.89	31.85	45.14	57.64

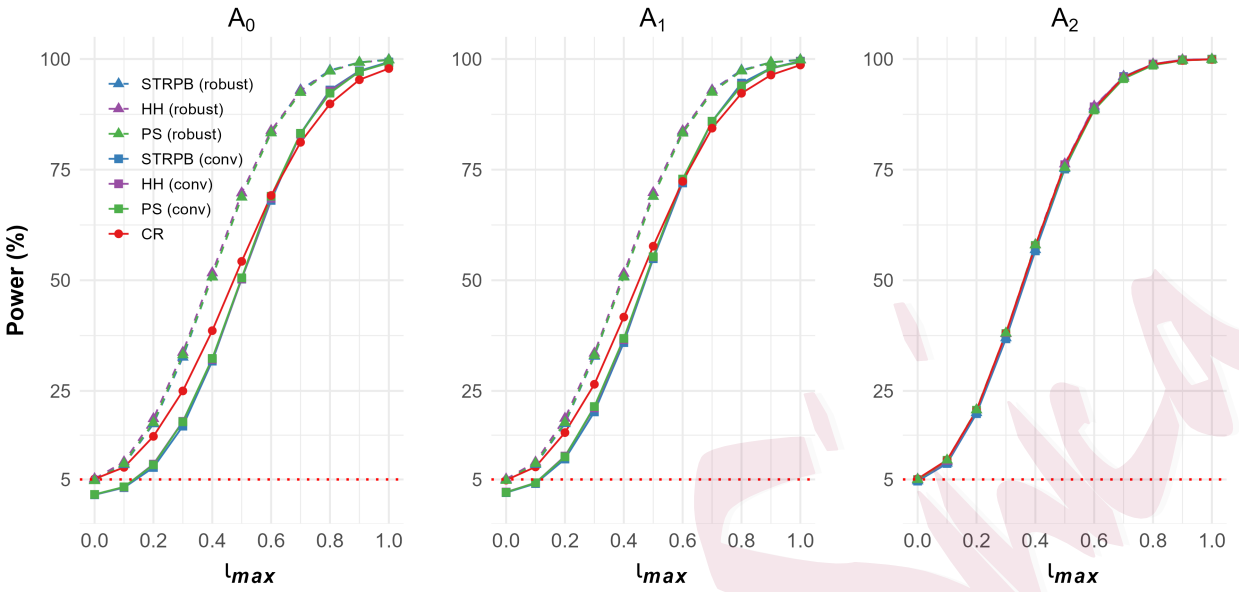

Figure 1: logRR for Example 1, power comparison of conventional and model-robust tests under $(\mathcal{A}_0, \mathcal{A}_1, \mathcal{A}_2)$ across $\iota_{\max}$; the red dashed line denotes $\alpha = 0.05$ under the null hypothesis.

Table 3: logRR results for the hypothetical trial: Type I error rates and power (in 10^{-2}) for Stage 1, Stage 2, and the combined analysis, across different working models and randomization schemes.

	Procedure	Test	Type I			Power		
			Stage 1	Stage 2	All	Stage 1	Stage 2	All
			$(\iota_1, \iota_2) = (0, 0)$			$(\iota_1, \iota_2) = (0.2, 0.4)$		
$\mathcal{A}_0$	CR		4.90	4.86	4.76	28.07	41.90	52.12
	STRPB	conv	3.99	3.81	3.33	26.08	41.90	51.67
		robust	5.00	5.00	4.93	29.83	44.89	56.04
	HH	conv	3.89	4.30	4.11	26.27	41.34	51.32
		robust	4.68	4.88	4.68	30.33	44.45	55.50
	PS	conv	4.23	4.10	3.98	26.99	41.99	51.58
		robust	5.21	4.86	4.93	30.24	44.73	55.55
$\mathcal{A}_1$	CR		4.73	5.12	4.87	28.70	43.97	53.75
	STRPB	conv	4.36	4.11	3.76	27.87	43.40	53.86
		robust	5.00	4.50	4.43	30.39	45.47	56.70
	HH	conv	4.07	4.74	4.54	28.07	42.99	53.59
		robust	4.74	5.30	5.19	30.69	45.14	56.24
	PS	conv	4.69	4.46	4.43	28.90	43.36	53.71
		robust	5.30	4.94	5.02	30.84	44.97	56.30
$\mathcal{A}_2$	CR		4.74	5.18	5.09	30.19	45.44	56.45
	STRPB	conv	5.05	4.79	4.81	30.15	45.62	56.72
		robust	5.09	4.57	4.44	30.50	45.66	56.88
	HH	conv	4.83	4.64	4.90	30.41	45.09	55.99
		robust	4.77	5.36	5.21	30.64	45.12	56.15
	PS	conv	4.67	4.96	4.68	30.94	45.18	56.62
		robust	5.25	4.93	5.01	31.13	45.18	56.78

6. Hypothetical trial example

We illustrate the implications of our findings through a hypothetical trial modeled on a recent multi-arm Phase II/III study (King et al., 2022, 2021) that compared multiple doses of baricitinib against placebo in treating alopecia areata (AA). AA is an autoimmune disorder that causes sudden, patchy hair loss; its most severe forms, alopecia totalis (AT) and alopecia universalis (AU), involve complete scalp or body hair loss and can severely diminish patients' quality of life. Conventional treatments such as topical corticosteroids or phototherapy may benefit patients with mild, patch-type AA, but are largely ineffective for chronic or refractory AT/AU. In contrast, oral Janus kinase (JAK) inhibitors have recently emerged as a promising therapeutic class. Baricitinib, a reversible and selective JAK1/JAK2 inhibitor already approved in over 70 countries for rheumatoid arthritis, has shown particular promise. In two Phase II clinical trials (King et al., 2022), 36 weeks of baricitinib 4 mg treatment led to nearly 40% of patients achieving a Scalp Alopecia Tool (SALT) score below 20, indicating over 85% hair regrowth from baseline—establishing baricitinib as an effective therapy for severe AA.

Motivated by this clinical evidence, we constructed a seamless Phase II/III design to demonstrate the performance of our model-robust Wald tests. The trial adopts an adaptive two-stage framework, where the transition from Stage 1 to Stage 2 is guided by interim efficacy results without interrupting patient recruitment. In Stage 1, 420 patients are randomized equally to one control arm (placebo) and two treatment arms (baricitinib 2 mg and 4 mg). Based on interim outcomes, the superior treatment arm is advanced to Stage 2, where 500 additional patients are enrolled for confirmatory comparison with placebo. This integrated design efficiently combines dose selection with confirmatory evaluation, highlighting

how our proposed methodology enhances decision-making in modern adaptive clinical trials.

The primary binary efficacy endpoint is defined as achieving a SALT score below 20, coded as 1 if achieved and 0 otherwise. Two clinically relevant baseline covariates are incorporated to capture patient heterogeneity. The first is the baseline SALT score, which quantifies the percentage of scalp hair loss at enrollment; values between 50 and 100 reflect the moderate-to-severe disease required for trial eligibility. The second covariate is disease duration, representing the length of the current episode. Consistent with observed distributions in chronic AA populations, we assume that 65% of patients have a disease duration shorter than four years, while 35% have a duration of four years or longer. Disease duration serves as an important prognostic factor, since shorter episodes are generally associated with more favorable treatment responses.

To ensure balance across these key prognostic factors, we implement CAR procedures. Patients are stratified jointly by baseline SALT score (<75 vs. ≥ 75) and disease duration (<4 years vs. ≥ 4 years), yielding four strata. Within each stratum, randomization is carried out using one of three widely used CAR approaches: STRPB, PS, and HH. For comparison, we also consider complete randomization. This design mirrors the stratification logic used in the baricitinib Phase II/III trials and ensures that treatment arms are comparable across clinically relevant subgroups.

The binary outcome is modeled using logistic regression:

$$\text{logit}(p_i) = \beta_0 + \iota_1 T_i^1 + \iota_2 T_i^2 + \beta_1 X_{i,1} + \beta_2 X_{i,2},$$

where T_i^1 and T_i^2 are indicators for the two treatment arms (with placebo as reference), $X_{i,1}$ denotes baseline SALT score, and $X_{i,2}$ denotes the binary indicator for disease duration (1 if $\geq$ 4 years, 0 otherwise). The parameter vector is set to $(\beta_0, \iota_1, \iota_2, \beta_1, \beta_2) =$

$(1.8, 0.2, 0.4, -0.04, -1.5)$. Here, $\beta_0 = 1.8$ represents the baseline intercept; $\iota_1 = 0.2$ and $\iota_2 = 0.4$ encode increasing treatment efficacy across the active arms; $\beta_1 = -0.04$ reflects the negative effect of more severe baseline SALT scores; and $\beta_2 = -1.5$ captures the reduced efficacy associated with longer disease duration. Each outcome Y_i is generated from a Bernoulli distribution with success probability p_i.

We assess the performance of competing statistical procedures under CR, STRPB, PS, and HH, using 10,000 replicates and three working models $\mathcal{A}_0$, $\mathcal{A}_1$, and $\mathcal{A}_2$ as defined in Example 1. Table **??** reports type I error rates from conventional Wald tests that ignore the contribution of CAR procedures. Consistent with our simulation findings, these tests are conservative under misspecified working models $\mathcal{A}_0$ and $\mathcal{A}_1$ across all three estimands: logRR, LOR, and ATE. Additionally, Tables 3, **??**, and **??** compare the conventional Wald test with the proposed model-robust Wald test for each metric. The results show a clear and consistent gain in power across all scenarios while maintaining valid type I error control, thereby demonstrating the practical advantage of the model-robust approach.

7. Discussion

The proposed framework offers a flexible inferential approach for seamless Phase II/III trials under CAR, but several aspects of its scope merit further discussion. The model robustness of our approach stems from the fact that it does not require the correct specification of a full parametric likelihood. Instead, the proposed method is based on an estimating equation, or Z-estimation, framework. Specifically, consistency and valid large-sample inference rely on the population estimating equation identifying the target parameter, together with the regularity conditions stated in our theoretical results and the covariate-balancing properties

induced by the CAR procedure, rather than on a fully correct outcome regression model. Under these conditions, the proposed approach remains valid under misspecification of the outcome model, misspecification of the link function, and omission of prognostic covariates. Our simulation studies support this robustness across all three misspecification settings. A limitation, however, is that finite-sample performance may deteriorate when the number of strata is relatively large, an issue also noted in Zhao et al. (2025).

Beyond the robustness considerations above, recent work has developed CAR procedures that can directly accommodate continuous covariates without prior discretization, such as the unified balancing framework of Ma et al. (2024). Extending our inferential framework to this setting would require replacing the current stratum-based imbalance quantities with suitable continuous-covariate imbalance metrics, and deriving the corresponding design-induced variance component, analogous to $\mathbf{\Gamma}_{\mathsf{CAR,adj}}$ in the present paper. One possible route is to combine our variance-decomposition argument with the asymptotic theory developed for continuous-covariate CAR procedures (Zhang, 2023; Ma et al., 2024). Bootstrap methods may also provide a practical alternative for estimating the design-induced variance when an explicit analytic form is difficult to obtain. We leave this extension to future work, as it would allow the proposed inference methods to cover trials with continuous or mixed-type covariates more flexibly.

A further extension concerns the endpoint structure across stages. For simplicity, the proposed method is developed under the assumption that the same outcome and estimand are used for Stage 1 dose selection and Stage 2 confirmatory inference, as in Ma et al. (2022). In seamless trials where different endpoints are used across stages, such as a surrogate endpoint for interim dose selection and a long-term clinical endpoint for confirmation, the general

Z-estimation framework could potentially be adapted by defining endpoint-specific estimating equations and deriving the corresponding CAR-adjusted covariance structures. In such mixed-endpoint designs, the main additional complication arises when Stage 1 data are used both to select the dose based on a surrogate endpoint and to contribute primary-endpoint evidence for the selected dose. If the selected-dose Stage 1 p-value is computed without accounting for this data-dependent selection, post-selection bias may occur, particularly when the surrogate and primary endpoints are correlated. By contrast, the Stage 2 p-value remains conditionally valid given the Stage 1 data, since Stage 2 observations are independent of the selection event. A rigorous extension would therefore need to specify how Stage 1 evidence is constructed and to account for the endpoint-specific estimating equations, the selection rule, and the design-induced variation from CAR. We leave this extension to future work.

Other extensions that worth pursuing include survival outcomes with censoring or recurrent events, high-dimensional covariates, machine-learning-based treatment effect estimators, and computationally efficient algorithms for real-time implementation. Developing these directions would further broaden the applicability of the proposed framework while maintaining rigorous error control in complex seamless trial settings.

Supplementary Material

The Supplementary Material provides the proofs of the theoretical results and additional numerical results.

References

Baldi Antognini, A. and M. Zagoraiou (2011). The covariate-adaptive biased coin design for balancing clinical trials in the presence of prognostic factors. *Biometrika 98*(3), 519–535.

Bauer, P. and M. Kieser (1999). Combining different phases in the development of medical treatments within a single trial. *Statistics in medicine 18*(14), 1833–1848.

Bauer, P., F. Koenig, W. Brannath, and M. Posch (2010). Selection and bias—two hostile brothers. *Statistics in Medicine 29*(1), 1–13.

Bauer, P. and K. Kohne (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics*, 1029–1041.

Bretz, F. et al. (2006). Confirmatory seamless phase ii/iii clinical trials with hypotheses selection at interim: general concepts. *Biometrical Journal: Journal of Mathematical Methods in Biosciences 48*(4), 623–634.

Bugni, F. A., I. A. Canay, and A. M. Shaikh (2018). Inference under covariate-adaptive randomization. *Journal of the American Statistical Association 113*(524), 1784–1796.

Bugni, F. A., I. A. Canay, and A. M. Shaikh (2019). Inference under covariate-adaptive randomization with multiple treatments. *Quantitative Economics 10*(4), 1747–1785.

Chen, C., K. Anderson, D. V. Mehrotra, E. H. Rubin, and A. Tse (2018). A 2-in-1 adaptive phase 2/3 design for expedited oncology drug development. *Contemporary clinical trials 64*, 238–242.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association 50*(272), 1096–1121.

Dunnett, C. W. (1964). New tables for multiple comparisons with a control. *Biometrics 20*(3), 482–491.

FDA (2022). Critical path opportunities initiated during 2006. Technical report, U.S. Department of Health and Human Services, Food and Drug Administration (FDA).

Gu, Y., H. Liu, and W. Ma (2023). Regression-based multiple treatment effect estimation under covariate-adaptive

randomization. *Biometrics 79*(4), 2869–2880.

Hallstrom, A. and K. Davis (1988). Imbalance in treatment assignments in stratified blocked randomization. *Controlled Clinical Trials 9*(4), 375–382.

Hampson, L. V. and C. Jennison (2015). Optimizing the data combination rule for seamless phase ii/iii clinical trials. *Statistics in Medicine 34*(1), 39–58.

Hu, F., Y. Hu, Z. Ma, and W. F. Rosenberger (2014). Adaptive randomization for balancing over covariates. *WIREs Computational Statistics 6*(4), 288–303.

Hu, F. and W. F. Rosenberger (2006). *The Theory of Response-Adaptive Randomization in Clinical Trials*. Hoboken, New Jersey: John Wiley & Sons.

Hu, F., X. Ye, and L.-X. Zhang (2023). Multi-arm covariate-adaptive randomization. *Science China Mathematics 66*(1), 163–190.

Hu, Y. and F. Hu (2012). Asymptotic properties of covariate-adaptive randomization. *The Annals of Statistics 40*(3), 1794–1815.

Jenkins, M., A. Stone, and C. Jennison (2011). An adaptive seamless phase ii/iii design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical statistics 10*(4), 347–356.

Jennison, C. and B. W. Turnbull (2007). Adaptive seamless designs: selection and prospective testing of hypotheses. *Journal of biopharmaceutical statistics 17*(6), 1135–1161.

King, B. et al. (2021). Efficacy and safety of the oral janus kinase inhibitor baricitinib in the treatment of adults with alopecia areata: phase 2 results from a randomized controlled study. *Journal of the American Academy of Dermatology 85*(4), 847–853.

King, B. et al. (2022). Two phase 3 trials of baricitinib for alopecia areata. *New England Journal of Medicine 386*(18), 1687–1699.

Koenig, F., W. Brannath, F. Bretz, and M. Posch (2008). Adaptive dunnett tests for treatment selection. *Statistics*

in medicine 27(10), 1612–1625.

Li, R., L. Wu, R. Liu, and J. Lin (2024). Flexible seamless 2-in-1 design with sample size adaptation. *Journal of Biopharmaceutical Statistics 34*(6), 1007–1025.

Liu, Q., M. A. Proschan, and G. W. Pledger (2002). A unified theory of two-stage adaptive designs. *Journal of the American Statistical Association 97*(460), 1034–1041.

Liu, Y. and F. Hu (2023). The impacts of unobserved covariates on covariate-adaptive randomized experiments. *The Annals of Statistics 51*(5), 1895–1920.

Liu, Y., L. Xia, and F. Hu (2024). Testing heterogeneous treatment effect with quantile regression under covariate-adaptive randomization. *Journal of Econometrics*, 105808.

Ma, W., F. Hu, and L. Zhang (2015). Testing hypotheses of covariate-adaptive randomized clinical trials. *Journal of the American Statistical Association 110*(510), 669–680.

Ma, W., P. Li, L.-X. Zhang, and F. Hu (2024). A new and unified family of covariate adaptive randomization procedures and their properties. *Journal of the American Statistical Association 119*(545), 151–162.

Ma, W., Y. Qin, Y. Li, and F. Hu (2020). Statistical inference for covariate-adaptive randomization procedures. *Journal of the American Statistical Association 115*(531), 1488–1497.

Ma, W., F. Tu, and H. Liu (2022). Regression analysis for covariate-adaptive randomization: a robust and efficient inference perspective. *Statistics in Medicine 41*(29), 5645–5661.

Ma, W., M. Wang, and H. Zhu (2022). Seamless phase ii/iii clinical trials with covariate adaptive randomization. *Statistica Sinica 32*(2), 1079–1098.

Marcus, R., P. Eric, and K. R. Gabriel (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika 63*(3), 655–660.

McCullagh, P. (2019). *Generalized linear models*. Routledge.

Müller, H.-H. and H. Schäfer (2001). Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics 57*(3), 886–891.

Pocock, S. J. and R. Simon (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*, 103–115.

Prowell, T. M., M. R. Theoret, and R. Pazdur (2016). Seamless oncology-drug development. *New England Journal of Medicine 374*(21), 2001–2003.

Rosenberger, W. F. and J. M. Lachin (2015). *Randomization in Clinical Trials: Theory and Practice* (2nd ed.). Hoboken, New Jersey: John Wiley & Sons.

Rosenberger, W. F., O. Sverdlov, et al. (2008). Handling covariates in the design of clinical trials. *Statistical Science 23*(3), 404–419.

Schmidli, H., F. Bretz, A. Racine, and W. Maurer (2006). Confirmatory seamless phase ii/iii clinical trials with hypotheses selection at interim: applications and practical considerations. *Biometrical Journal 48*(4), 635–643.

Shao, J. and X. Yu (2013). Validity of tests under covariate-adaptive biased coin randomization and generalized linear models. *Biometrics 69*(4), 960–969.

Shao, J., X. Yu, and B. Zhong (2010). A theory for testing hypotheses under covariate-adaptive randomization. *Biometrika 97*(2), 347–360.

Simes, R. J. (1986). An improved bonferroni procedure for multiple tests of significance. *Biometrika 73*(3), 751–754.

Sverdlov, O. (2015). *Modern adaptive randomized clinical trials: statistical and practical aspects*. CRC Press.

Taves, D. R. (1974). Minimization: a new method of assigning patients to treatment and control groups. *Clinical Pharmacology & Therapeutics 15*(5), 443–453.

Toorawa, R. et al. (2009). Use of simulation to compare the performance of minimization with stratified blocked randomization. *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry 8*(4), 264–278.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, Volume 3. Cambridge university press.

Wang, B. et al. (2023). Model-robust inference for clinical trials that improve precision by stratified randomization and covariate adjustment. *Journal of the American Statistical Association 118*(542), 1152–1163.

Ye, T., J. Shao, Y. Yi, and Q. Zhao (2023). Toward better practice of covariate adjustment in analyzing randomized clinical trials. *Journal of the American Statistical Association 118*(544), 2370–2382.

Ye, T., Y. Yi, and J. Shao (2022). Inference on the average treatment effect under minimization and other covariate-adaptive randomization methods. *Biometrika 109*(1), 33–47.

Zelen, M. (1974). The randomization and stratification of patients to clinical trials. *Journal of Chronic Diseases 27*(7), 365–375.

Zhang, L.-X. (2023). Asymptotic properties of multi-treatment covariate adaptive randomization procedures for balancing observed and unobserved covariates. *arXiv preprint arXiv:2305.13842*.

Zhao, F., Y. Liu, and F. Hu (2025). Statistical inference on the relative risk following covariate-adaptive randomization. *Biometrics 81*(2), ujaf036.

Kun Yi

Department of ISOM, School of Business and Management, HKUST

E-mail: `kyiae@connect.ust.hk`

Lucy Xia

Department of ISOM, School of Business and Management, HKUST

E-mail: `lucyxia@ust.hk`