

---

# CALIBRATED BAYES ANALYSIS OF CLUSTER-RANDOMIZED TRIALS

---

A PREPRINT

**Ruyi Liu**

Department of Biostatistics, Yale School of Public Health,  
New Haven, CT, USA

**Joshua L. Warren**

Department of Biostatistics, Yale School of Public Health,  
New Haven, CT, USA

**Yuki Ohnishi**

Department of Biostatistics, Yale School of Public Health,  
New Haven, CT, USA

**Donna Spiegelman**

Department of Biostatistics, Yale School of Public Health,  
New Haven, CT, USA

**Liangyuan Hu**

Department of Biostatistics and Epidemiology, Rutgers University,  
New Brunswick, NJ, USA

**Fan Li\***

Department of Biostatistics, Yale School of Public Health,  
New Haven, CT, USA

November 27, 2025

## ABSTRACT

In cluster-randomized trials (CRTs), entire clusters of individuals are randomized to treatment, and outcomes within a cluster are typically correlated. While frequentist approaches are standard practice for CRT analysis, Bayesian methods have emerged as a strong alternative. Previous work has investigated the use of Bayesian hierarchical models for continuous, binary, and count outcomes in

---

\*Corresponding author: Fan Li, Department of Biostatistics, Yale School of Public Health. Email: [fan.f.li@yale.edu](mailto:fan.f.li@yale.edu).

CRTs, but these approaches focus on model-based treatment effect coefficients as the target estimands, which may have ambiguous interpretation under model misspecification and informative cluster size. In this article, we introduce a calibrated Bayesian procedure for estimand-aligned analysis of CRTs even in the presence of potentially misspecified models. We propose estimators targeting both the cluster-average treatment effect (cluster-ATE) and individual-average treatment effect (individual-ATE), particularly in scenarios with informative cluster sizes. We additionally explore strategies for summarizing the posterior samples that can achieve the frequentist coverage guarantee even under working model misspecification. We provide simulation evidence to demonstrate the model-robustness property of the proposed Bayesian estimators in CRTs, and further investigate the impact of covariate adjustment as well as the use of more flexible Bayesian nonparametric working models.

**Keywords** Causal inference · informative cluster size · estimands · model-robustness · covariate adjustment · Bayesian additive regression trees

## 1 Introduction

In a cluster-randomized trial (CRT), entire social units or groups of individuals, rather than individuals themselves, are randomized to different treatment conditions (Donner et al., 2000). CRTs are often the default when randomizing individuals is infeasible, when studying a cluster-level intervention, or when there are concerns about treatment contamination under individual randomization (Murray, 1998). There is a growing number of CRTs conducted to evaluate the effect of new healthcare innovations into the routine practice settings (Weinfurt et al., 2017). However, cluster randomization is typically less efficient than individual randomization because outcomes from individuals within the same cluster tend to be more similar than those from different clusters, requiring that this within-cluster correlation be accounted for when estimating the treatment effect. For this purpose, linear and generalized linear mixed models are a popular class of regression-based methods for analyzing CRTs that account for within-cluster correlations and adjust for baseline covariates (Wang et al., 2025, 2024b).

Although analyses of CRTs have traditionally relied on frequentist methods, a growing literature demonstrates the utility of Bayesian approaches. For example, Spiegelhalter (2001) considered a linear mixed model for analyzing continuous outcomes in CRTs, and assessed the sensitivity under different priors for variance components and intracluster correlation coefficient (ICC). Tong et al. (2024) extended this approach to further address heterogeneous variances across clusters. Turner et al. (2001) and Thompson et al. (2004) provided further extensions to binary outcomes using generalized linear mixed models and possibly an arm-specific random-effects structure. Clark and Bachmann (2010) focused on Bayesian hierarchical models to handle Poisson outcomes. In particular, Bayesian methods offer several appealing features in analyzing CRTs (Jones et al., 2021). First, the analyst has the flexibility in integrating informative priors on variance components and ICCs to inform the data analysis, especially when a limited number of clusters dictates fewer replication at the higher hierarchical level (Grantham et al., 2022). Second, Bayesian inference automatically quantifies uncertainty through the joint posterior distribution of the model parameters; in practice, even under complex models, we can approximate this posterior via Markov chain Monte Carlo (MCMC) to obtain samples for interval estimation and prediction (Spiegelhalter, 2001). In the context of CRTs, Bayesian methods naturally facilitate posterior inference on variance components and ICCs, which may not be directly available through standard model fitting procedures in

existing software (Turner et al., 2001; Clark and Bachmann, 2010). Finally, Bayesian nonparametric priors provide a promising avenue for relaxing the restrictive distributional assumptions on fixed effects and random effects inherent in hierarchical models, which can be valuable when cluster-specific heterogeneity departs from standard parametric forms. For instance, Bayesian Additive Regression Trees (BART) (Chipman et al., 2010), an ensemble approach that models outcomes as a sum of regression trees, allows treatment effects to be estimated without relying on strong parametric specifications of the outcome model. Thus, while the relaxation of distributional assumptions for random effects and the flexible estimation of fixed effects are distinct contributions, they are complementary in broadening the applicability of Bayesian nonparametric methods for CRT analysis.

With a few recent exceptions (e.g., Ohnishi and Li (2025) who focused on causal mediation analysis), previous efforts on Bayesian analysis of CRTs have primarily focused on model-based treatment effect estimands. In the context of CRTs, there has been little investigation, to date, on applying Bayesian methods to target nonparametric treatment effect estimands defined under the potential outcomes framework. Thus, it remains unclear, for example, whether a misspecified Bayesian hierarchical model can still target a clear treatment effect estimand under the potential outcomes framework in CRTs. In a parallel-arm CRT, Kahan et al. (2023) exemplified two different average treatment effect (ATE) estimands. The cluster-average treatment effect (cluster-ATE) assigns equal weight to each cluster regardless of the cluster size, and the individual-average treatment effect (individual-ATE) assigns equal weight to each individual participant regardless of cluster membership. These two estimands differ in magnitude in the presence of informative cluster size, that is, when the outcomes of individuals or the within-cluster treatment effect are marginally associated with cluster size. In the latter case, the cluster size is a marginal effect modifier that requires special considerations in the analysis. As Greenland (1982) highlighted, summary estimators may exhibit inconsistent behavior depending on the distribution of cluster sizes, underscoring the importance of accounting for possible interactions between treatment effects and cluster size distributions when estimating marginal treatment effects. The concept of informative cluster size can be viewed as a specific manifestation of this broader principle. When the cluster size is non-informative, Wang et al. (2025) have proved that the treatment coefficient estimator under an arbitrarily misspecified linear mixed model converges to the ATE (this is the case when cluster-ATE and individual-ATE coincide). However, in the presence of informative cluster size, Kahan et al. (2024) demonstrated that the treatment coefficient estimator under a linear mixed model (without covariates) targets neither cluster-ATE nor individual-ATE. These findings suggest that model specification in CRTs can inadvertently dictate the estimands being estimated. To that end, *model-robust* methods for inferring cluster-ATE and individual-ATE are of growing interest. Several frequentist methods have been proposed (Benitez et al., 2023; Balzer et al., 2019, 2023; Nugent et al., 2024; Wang et al., 2024a), for example, based on linear regression (Su and Ding, 2021) and the vehicle of efficient influence function (Wang et al., 2024b).

The primary objective of this article is to develop *model-robust* Bayesian analysis methods for CRTs, delivering estimand-aligned inference even when analyses are based on potentially misspecified Bayesian hierarchical models. First, starting with a fitted Bayesian hierarchical model, we consider a model-based g-computation estimator (a direct standardization approach), with and without covariate adjustment, for estimating both the cluster-ATE and the individual-ATE. We then apply the model-robust standardization estimator in Li et al. (2025) to the fitted Bayesian hierarchical model to mitigate potential bias due to model misspecification. We establish that, under a Bayesian formulation and via a posterior contraction argument, this model-robust point estimator is root- $M$  consistent under CRT randomization, even when the outcome model is misspecified. Second, we examine the frequentist property of Bayesian interval

estimators derived from the posterior distributions, especially under violations of model assumptions. Previous research ([Hong and Martin, 2020](#)) has shown that posterior distributions under model misspecification may exhibit bias, resulting in misleading inference. Building upon [Antonelli et al. \(2022\)](#) (whose focus was on independent rather than correlated data), we propose a calibrated variance estimator that integrates posterior samples with cluster-level bootstrap. Finally, we explore how Bayesian nonparametric priors (specifically the BART priors) may contribute to the quality of inference by relaxing parametric assumptions commonly made in linear mixed models for CRT analyses. Collectively, our study seeks to address the following three interconnected questions:

- Q1 How can we construct point estimators within Bayesian hierarchical models that remain robust to potential model misspecification when analyzing CRTs?
- Q2 What posterior inferential strategy can offer reliable frequentist coverage properties for interval estimation under conditions of model misspecification in CRTs?
- Q3 To what extent can Bayesian nonparametric methods, such as BART ([Chipman et al., 2010](#)), improve the estimation of cluster- and individual-ATE in CRTs?

The remainder of this article is organized as follows. In Section 2.1, we introduce the notation, assumptions and the two estimands of interest: the cluster-ATE and individual-ATE. In Sections 2.2 and 2.3, we review two types of estimand formulation method that can be used to target cluster-ATE and individual-ATE, both with and without covariate adjustment. In Section 3, we then detail how to combine each point estimation strategy with the posterior samples from a fitted Bayesian hierarchical model, and introduce a calibrated variance estimation approach that can restore the frequentist property. In Section 4, we report on a series of simulations that examine the finite-sample performance of point and variance estimation methods under data-generating processes of varying complexity. We provide an illustrative data example in Section 5, and offer concluding remarks in Section 6. Example code to implement all methods introduced in this article is provided in <https://github.com/Ruyi-Liu/Model-robust-Bayesian-inference-in-cluster-randomized-trials>.

## 2 Estimands and model-assisted estimation

### 2.1 Notation and setup

We consider a parallel-arm CRT, where data are collected among  $M$  clusters involving a total of  $N$  individuals. Each cluster  $i$  has  $N_i$  individuals ( $i = 1, \dots, M$ ), and the total number of individuals is  $N = \sum_{i=1}^M N_i$ . We consider the observed cluster size as the true cluster size  $N_i$  and do not consider within-cluster subsampling. We let  $A_i \in \{0, 1\}$  be cluster-level treatment indicator for cluster  $i$ , with  $A_i = 1$  indicating the assignment to the treatment arm and  $A_i = 0$  indicating the assignment to the control arm. In a CRT, individuals within a cluster receive the same treatment condition, and we assume they comply to the assigned treatment. We let subscript  $i$  represent the cluster index ( $i = 1, \dots, M$ ) and  $j$  represent the individual index within cluster  $i$  ( $j = 1, \dots, N_i$ ). Define  $\mathbf{X}_{ij}$  as the  $q$ -dimensional vector of baseline covariates for individual  $j$  in cluster  $i$ , including both individual-level and cluster-level characteristics. We further let  $\mathbf{X}_i = (\mathbf{X}_{i1}, \dots, \mathbf{X}_{iN_i})^\top \in \mathbb{R}^{N_i \times q}$  be the collection of all baseline covariates in cluster  $i$ .

For individual  $j$  in cluster  $i$ , we define  $Y_{ij}$  as the observed outcome. Let  $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iN_i})^\top$  be the vector of all observed outcomes in cluster  $i$ . Under the potential outcomes framework, the individual potential outcome is defined as

$Y_{ij}(a)$  under treatment condition  $a \in \{0, 1\}$ . Under the cluster-level Stable Unit Treatment Value Assumption (SUTVA), we have  $Y_{ij} = A_i Y_{ij}(1) + (1 - A_i) Y_{ij}(0)$ . Cluster-level SUTVA requires a well-defined intervention, and implies that individual potential outcomes in cluster  $i$  depend only on their own cluster's treatment assignment, but not the assignments to any other clusters. Let  $\mathbf{Y}_i(a) = \{Y_{i1}(a), \dots, Y_{iN_i}(a)\} \in \mathbb{R}^{N_i}$  be the collection of potential outcomes for cluster  $i$  under condition  $a$ . Then, the collection of random variables in cluster  $i$  is  $\mathbf{O}_i^{\text{full}} = \{\mathbf{Y}_i(1), \mathbf{Y}_i(0), N_i, \mathbf{X}_i\}$ . The complete data  $\{(O_1^{\text{full}}, A_1), \dots, (O_M^{\text{full}}, A_M)\}$  is partially observed. We assume that the random variable vectors from different clusters are mutually independent, but correlation cannot be ignored within the same cluster.

We consider two typical types of marginal causal estimands in CRTs described in [Kahan et al. \(2023\)](#). These estimands are cluster-ATE and individual-ATE, considering the unit of inference at the cluster level and the individual level, respectively. The cluster-ATE addresses the expected change in outcomes due to treatment across the population of clusters. This estimand weights each cluster equally, irrespective of its size  $N_i$ . On the other hand, the individual-ATE addresses the expected change in outcomes due to treatment for the population of individuals, and mimics the conventional estimand in an individually-randomized trial. The individual-ATE weights each individual equally, regardless of the cluster to which they belong. We pursue a super population framework in defining estimands (a detailed comparison between finite-population and super-population estimands is given in [Kahan et al. \(2024\)](#) in CRTs), such that the observed data are a random sample from an infinite hypothetical population of clusters. The cluster-ATE ( $\Delta_C$ ) and individual-ATE ( $\Delta_I$ ) are therefore defined by invoking expectations over this super-population. Specifically, let function  $f$  be a predetermined smooth function that defines the scale of the effect measure, the two types of estimands are defined as  $\Delta_C \equiv f(\mu_C(1), \mu_C(0))$  and  $\Delta_I \equiv f(\mu_I(1), \mu_I(0))$ , where for  $a \in \{0, 1\}$ ,  $\mu_C(a) = \mathbb{E}\left(\frac{1}{N_i} \sum_{j=1}^{N_i} Y_{ij}(a)\right)$ ,  $\mu_I(a) = \mathbb{E}\left(\sum_{j=1}^{N_i} Y_{ij}(a)\right) / \mathbb{E}(N_i)$ . Here, setting  $f(x, y) = x - y$ ,  $f(x, y) = x/y$  and  $f(x, y) = x(1 - y)/\{y(1 - x)\}$  lead to the difference, risk ratio, and odds ratio estimands, respectively.

To identify the cluster-ATE and individual-ATE, the core challenge lies in the fact that for each individual, we only observe the potential outcome for the actual treatment received, denoted as  $Y_{ij}^{\text{obs}} \equiv Y_{ij} = Y_{ij}(A_i)$ , which is factual. The other potential outcome, represented by  $Y_{ij}^{\text{mis}} = Y_{ij}(1 - A_i)$ , remains counterfactual ([Li et al., 2023](#)). In CRTs, the treatment indicator  $A_i$  is randomized such that  $A_i \perp \{\mathbf{Y}_i(0), \mathbf{Y}_i(1), \mathbf{X}_i, N_i\}$ . For every cluster in the study, the probability of receiving the treatment,  $\pi = \mathbb{P}(A_i = 1) \in (0, 1)$ , is also often a constant. Due to cluster randomization, the marginal distribution of the potential outcomes can be identified from the observed data. To see this, using the Law of Iterated Expectation, the estimands of interest can be further expressed as functions of the observed data as  $\mu_C(a) = \mathbb{E}\{\mathbb{E}(\bar{Y}_i | A_i = a, \mathbf{X}_i, N_i)\}$  and  $\mu_I(a) = \mathbb{E}\{\mathbb{E}(Y_{i+} | A_i = a, \mathbf{X}_i, N_i)\} / \mathbb{E}(N_i)$ .

In practice, estimation of  $\mu_C(a)$  and  $\mu_I(a)$  reduces to estimating the conditional mean functions  $m_C(a, \mathbf{x}, n) = \mathbb{E}(\bar{Y}_i | A_i = a, \mathbf{X}_i = \mathbf{x}, N_i = n)$  and  $m_I(a, \mathbf{x}, n) = \mathbb{E}(Y_{i+} | A_i = a, \mathbf{X}_i = \mathbf{x}, N_i = n)$ , where  $\bar{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} Y_{ij}$  and  $Y_{i+} = \sum_{j=1}^{N_i} Y_{ij}$ . These functions can be modeled using linear mixed models or flexible nonparametric regressions. Given estimators  $\hat{m}_C$  and  $\hat{m}_I$ , plug-in (g-computation) estimators are obtained by averaging over the empirical distribution of  $(\mathbf{X}_i, N_i)$  across observed clusters. Equivalently, one may integrate  $\hat{m}_C$  and  $\hat{m}_I$  over a fitted distribution for  $(\mathbf{X}_i, N_i)$ . Under standard identification conditions (consistency, exchangeability, and positivity), these plug-in estimators target the desired estimands; see Section 2.2 for an implementation using posited linear mixed models.

## 2.2 Model-based g-computation formula

Next, we review two distinct approaches to formulating the estimand. We begin with the model-based approach. In what follows, we focus on continuous outcomes and structure the discussion around the commonly used linear mixed model. In the context of parallel-arm CRTs, a linear mixed model that adjust for covariates with main effects and treatment-covariate interactions (referred to as the analysis of covariance II model (Wang et al., 2025; Yang et al., 2020)) is given by, for  $i = 1, \dots, M$  and  $j = 1, \dots, N_i$ ,

$$Y_{ij} = \beta_0 + \beta_1 A_i + \boldsymbol{\beta}_2^\top \mathbf{X}_{ij} + \beta_3 N_i + \boldsymbol{\beta}_4^\top A_i \mathbf{X}_{ij} + \beta_5 A_i N_i + \phi_i + \epsilon_{ij}, \quad (1)$$

where  $\phi_i \sim \mathcal{N}(0, \sigma_\phi^2)$  is the random effect for cluster  $i$  that measures the cluster-specific departure from the overall mean,  $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$  is the residual error for individual  $j$  in cluster  $i$ , and  $\beta_0, \beta_1, (\boldsymbol{\beta}_2^\top, \beta_3)$ , and  $(\boldsymbol{\beta}_4^\top, \beta_5)$  stand for the intercept, main effect of treatment, main effect of potential effect modifier, and the treatment-covariate interaction effects. Under this working model, the cluster random effects satisfy  $\mathbb{E}(\phi_i) = 0$  and are independent of  $(A_i, \mathbf{X}_{ij}, N_i)$ ; hence they integrate out of the conditional means used for standardization. In CRTs, it is commonly assumed within mixed model frameworks that random effects are mutually independent and are further independent of the treatment assignment and all covariates (Wang et al., 2025). Within this framework, the fraction of the entire variance in  $Y_{ij}$  attributable to the between-group variation, represented by  $\sigma_\phi^2 / (\sigma_\phi^2 + \sigma_\epsilon^2)$ , is known as the ICC (Murray, 1998). We can then define the covariate adjusted model-based g-computation formulas as  $\Delta_C^{\text{mb}} = f(\mu_C^{\text{mb}}(1), \mu_C^{\text{mb}}(0))$ , and  $\Delta_I^{\text{mb}} = f(\mu_I^{\text{mb}}(1), \mu_I^{\text{mb}}(0))$ , where

$$\mu_C^{\text{mb}}(a) = \beta_0 + \beta_1 a + \mathbb{E} \left\{ N_i^{-1} \sum_{j=1}^{N_i} (\boldsymbol{\beta}_2^\top \mathbf{X}_{ij} + \beta_3 N_i + \boldsymbol{\beta}_4^\top a \mathbf{X}_{ij} + \beta_5 a N_i) \right\}, \quad (2)$$

$$\mu_I^{\text{mb}}(a) = \beta_0 + \beta_1 a + \frac{\mathbb{E} \left\{ \sum_{j=1}^{N_i} (\boldsymbol{\beta}_2^\top \mathbf{X}_{ij} + \beta_3 N_i + \boldsymbol{\beta}_4^\top a \mathbf{X}_{ij} + \beta_5 a N_i) \right\}}{\mathbb{E}(N_i)}. \quad (3)$$

In Equation (2)–(3), expectations are taken over the joint distribution of  $(\mathbf{X}_i, N_i)$  across clusters. Intuitively, both formulas involve calculating the mean of model predictions for each cluster when the treatment is set to  $a$  (Wang et al., 2024b). This approach is standard in the causal inference literature for independent data, and is sometimes referred to as the population standardization approach to estimate marginal estimands in observational studies (Rosenbaum, 1987).

A special case of the linear mixed model in Equation (1) is the unadjusted analysis, in which neither baseline covariates nor cluster-size adjustments are included. Under this unadjusted outcome model, the parameter vectors  $(\boldsymbol{\beta}_2^\top, \beta_3)$  and  $(\boldsymbol{\beta}_4^\top, \beta_5)$  in the above analysis of covariance model are set to zero, and the resulting model-based g-computation (Rosenbaum, 1987) formulas reduce to  $\Delta_C^{\text{mb,unadj}} = f(\mu_C^{\text{mb,unadj}}(1), \mu_C^{\text{mb,unadj}}(0))$ , and  $\Delta_I^{\text{mb,unadj}} = f(\mu_I^{\text{mb,unadj}}(1), \mu_I^{\text{mb,unadj}}(0))$ , where  $\mu_C^{\text{mb,unadj}}(a) = \beta_0 + \beta_1 a$  and  $\mu_I^{\text{mb,unadj}}(a) = \beta_0 + \beta_1 a$ . Therefore, in their mathematical forms, the g-computation formulas without baseline covariates lead to identical expressions, and consequently the same estimated values, for both the cluster-ATE and the individual-ATE (even though the true cluster-ATE and individual-ATE can differ due to informative cluster size). This reflects an implicit limitation of the unadjusted model-based g-computation approach. For example, the difference estimands are equal under this approach,  $\Delta_C^{\text{mb,unadj}} = \beta_0 + \beta_1 \times 1 - (\beta_0 + \beta_1 \times 0) = \beta_1 = \Delta_I^{\text{mb,unadj}}$ . However, in the presence of informative cluster size, the

true values of the cluster-ATE and individual-ATE are not necessarily equal (Kahan et al., 2023). Consequently, this unadjusted model-based g-computation approach is expected to be biased under informative cluster size.

### 2.3 Model-robust standardization formula

For model-based covariate adjustment to provide model-robust inference, Wang et al. (2024b) demonstrated that it generally requires independence between the cluster size and cluster characteristics, an assumption often violated in the presence of informative cluster size. Cluster size can be influenced by geographical features or other contextual factors. Thus, aside from the g-computation formulation for cluster-ATE and individual-ATE, which may subject to outcome model misspecification bias, we further consider the model-robust standardization formula introduced in Li et al. (2025) to leverage the known randomization mechanism in CRTs and avoid reliance on correct outcome model specification. The estimators in Li et al. (2025) are frequentist in nature and serve as the foundation for our proposed Bayesian analogs, and they also demonstrate superior performance over conventional treatment effect coefficient estimators in the presence of informative cluster size. The model-robust standardization formula under covariate adjustment are given by  $\Delta_C^{\text{mr}} = f(\mu_C^{\text{mr}}(1), \mu_C^{\text{mr}}(0))$  and  $\Delta_I^{\text{mr}} = f(\mu_I^{\text{mr}}(1), \mu_I^{\text{mr}}(0))$ , where

$$\mu_C^{\text{mr}}(a) = \mathbb{E} \left\{ \mathbb{E}(\bar{Y}_i | A_i = a, \mathbf{X}_i, N_i) + \frac{\mathbb{I}(A_i = a)(\bar{Y}_i - \mathbb{E}(\bar{Y}_i | A_i = a, \mathbf{X}_i, N_i))}{\pi^a (1 - \pi)^{1-a}} \right\}, \quad (4)$$

$$\mu_I^{\text{mr}}(a) = \frac{1}{\mathbb{E}(N_i)} \cdot \mathbb{E} \left\{ \mathbb{E}(Y_{i+} | A_i = a, \mathbf{X}_i, N_i) + \frac{\mathbb{I}(A_i = a)(Y_{i+} - \mathbb{E}(Y_{i+} | A_i = a, \mathbf{X}_i, N_i))}{\pi^a (1 - \pi)^{1-a}} \right\}, \quad (5)$$

where  $\mathbb{E}(\bar{Y}_i | A_i = a, \mathbf{X}_i, N_i)$  and  $\mathbb{E}(Y_{i+} | A_i = a, \mathbf{X}_i, N_i)$  denote the conditional expectations of the average outcome and the sum outcome, respectively, within cluster  $i$ , given cluster-specific baseline covariates and cluster size, which can be estimated using any appropriate outcome regression model.

Typically, the conditional mean of the average outcome and the sum outcome within cluster are induced from an individual-level outcome model, for which there are numerous options. For example, when the mean function is equal to zero,  $\mu_C^{\text{mr}}(a)$  and  $\mu_I^{\text{mr}}(a)$  become nonparametric two-sample mean estimators with different weights. If the mean function is induced by the unadjusted linear mixed model, the conditional expectation of the average outcome becomes  $\mathbb{E}(\bar{Y}_i | A_i = a, \mathbf{X}_i, N_i) = \mathbb{E}(\bar{Y}_i | A_i = a) = \beta_0 + \beta_1 a$ , and the conditional expectation of the sum outcome becomes  $\mathbb{E}(Y_{i+} | A_i = a, \mathbf{X}_i, N_i) = N_i \times (\beta_0 + \beta_1 a)$ . Under this setting, the first term in both  $\mu_C^{\text{mr}}(a)$  and  $\mu_I^{\text{mr}}(a)$  are equal, reflecting the unadjusted model's mean structure. However, the second term in each expression incorporates residuals that differ due to the use of cluster-level versus individual-level aggregation, thereby enabling adjustment for informative cluster size. Under model (1), which is the mixed model with linear adjustment for covariates and treatment-covariate interactions, we obtain the following expression of the cluster-level mean outcome  $\mathbb{E}(\bar{Y}_i | A_i = a, \mathbf{X}_i, N_i) = \beta_0 + \beta_1 a + \left( \frac{1}{N_i} \sum_{j=1}^{N_i} \boldsymbol{\beta}_2^\top \mathbf{X}_{ij} \right) + \beta_3 N_i + \left( \frac{1}{N_i} \sum_{j=1}^{N_i} \boldsymbol{\beta}_4^\top a \mathbf{X}_{ij} \right) + \beta_5 a N_i$ , and the cluster-level sum outcome  $\mathbb{E}(Y_{i+} | A_i = a, \mathbf{X}_i, N_i) = \beta_0 N_i + \beta_1 a N_i + \left( \sum_{j=1}^{N_i} \boldsymbol{\beta}_2^\top \mathbf{X}_{ij} \right) + \beta_3 N_i^2 + \left( \sum_{j=1}^{N_i} \boldsymbol{\beta}_4^\top a \mathbf{X}_{ij} \right) + \beta_5 a N_i^2$ . Although this standardization procedure relies on a working outcome model for  $\bar{Y}_i$  or  $Y_{i+}$ , these efficient influence function-based model-robust estimators can achieve consistency of the target estimands, namely the cluster-ATE and individual-ATE, even under outcome model misspecification, since the randomization probability  $\pi_i$  in CRTs is often predetermined by the investigator. A proof of model robustness is provided in Li et al. (2025).

### 3 Bayesian hierarchical model for cluster-average and individual-average treatment effects

#### 3.1 Point estimation

There are a range of literature on Bayesian methods (Carlin and Louis, 1997; Gelman et al., 1995), as well as review papers specifically addressing the application of Bayesian methods to randomized clinical trials (Spiegelhalter et al., 1994; Kadane, 1995). Assuming we have unknown quantities  $\theta$ , standard frequentist methods make inferences based on a presumed likelihood  $p(\mathbf{Y} \mid \theta)$ . Bayesian approaches enhance this by incorporating a prior distribution  $p(\theta)$ , which represents our initial uncertainty regarding  $\theta$  prior to considering the data. In Bayesian analysis, the selection of an appropriate prior distribution remains an important consideration, aiming either to incorporate relevant external information or to provide a non-informative formulation in certain scenarios (Spiegelhalter, 2001). In what follows, we anchor our discussion in the linear mixed model presented in Section 2.2. We assume the following weakly-informative, conjugate priors for  $\beta$ ,  $\sigma_\phi^2$  and  $\sigma_\epsilon^2$ :  $\beta \sim \mathcal{N}_p(\mathbf{0}, \mathbf{L}_0) = \mathcal{N}_p(\mathbf{0}, 100\mathbf{I}_p)$ , where  $\beta = (\beta_0, \beta_1, \beta_2^\top, \beta_3, \beta_4^\top, \beta_5)^\top$ ,  $\sigma_\epsilon^2 \sim \text{IG}(a, b) = \text{IG}(0.001, 0.001)$ , and  $\sigma_\phi^2 \sim \text{IG}(\alpha, c) = \text{IG}(0.001, 0.001)$ . The derived full conditional distributions are presented as follows:

$$\begin{aligned}\beta \mid \mathbf{Y}, \phi, \sigma_\phi^2, \sigma_\epsilon^2 &\sim \mathcal{N}\left(\left(\frac{\mathbf{H}^\top \mathbf{H}}{\sigma_\epsilon^2} + \mathbf{L}_0^{-1}\right)^{-1} \cdot \frac{\mathbf{H}^\top (\mathbf{Y} - \mathbf{Z}\phi)}{\sigma_\epsilon^2}, \left(\frac{\mathbf{H}^\top \mathbf{H}}{\sigma_\epsilon^2} + \mathbf{L}_0^{-1}\right)^{-1}\right), \\ \phi_i \mid \mathbf{Y}, \phi_{-i}, \beta, \sigma_\phi^2, \sigma_\epsilon^2 &\sim \mathcal{N}\left(\frac{\sigma_\phi^2 \cdot \sum I(\mathbf{Z}_{ki} = 1) (\mathbf{Y}_k - \mathbf{H}_k^\top \beta)}{\sigma_\epsilon^2 + \sigma_\phi^2 \cdot N_i}, \frac{\sigma_\phi^2 \cdot \sigma_\epsilon^2}{\sigma_\epsilon^2 + \sigma_\phi^2 \cdot N_i}\right), \\ \sigma_\phi^2 \mid \mathbf{Y}, \beta, \phi, \sigma_\epsilon^2 &\sim \text{IG}\left(\alpha + \frac{M}{2}, c + \frac{1}{2}\phi^\top \phi\right), \\ \sigma_\epsilon^2 \mid \mathbf{Y}, \beta, \phi, \sigma_\phi^2 &\sim \text{IG}\left(a + \frac{N}{2}, b + \frac{1}{2}(\mathbf{Y} - \mathbf{H}\beta - \mathbf{Z}\phi)^\top (\mathbf{Y} - \mathbf{H}\beta - \mathbf{Z}\phi)\right),\end{aligned}$$

where  $\mathbf{Y}$  is the  $N \times 1$  vector of all observed outcomes,  $\mathbf{H}$  is the  $N \times p$  design matrix of all fixed effect covariates with corresponding  $p \times 1$  vector of fixed effect regression coefficients,  $\mathbf{Z}$  is the  $N \times M$  indicator matrix for cluster membership,  $\phi = (\phi_1, \dots, \phi_M)^\top$ ,  $\phi_{-i} = (\phi_1, \phi_{i-1}, \dots, \phi_{i+1}, \phi_M)^\top$ ,  $I(\cdot)$  is the indicator function, which equals 1 if the condition inside is true and 0 otherwise. With the explicit forms of these full conditional distributions established, we can construct a Gibbs sampling algorithm to generate posterior samples for all model parameters. This allows us to estimate the g-computation estimands defined in Section 2.2 and model-robust estimands defined in Section 2.3, either with or without adjusting for covariates. Consider  $B$  samples,  $\{\beta^{(b)}\}_{b=1}^B$ , drawn from the posterior distribution. Let  $\hat{\Delta}$  serve as a general notation for the point estimator of either cluster-ATE or individual-ATE, derived using a model-based g-computation formula (e.g., Equations (2) and (3)) or a model-robust standardization formula (e.g., Equations (4) and (5)). We define the point estimator  $\hat{\Delta}$  as  $\hat{\Delta} = E_{\beta \mid \mathbf{D}}[\Delta(\mathbf{D}, \beta)] \approx \frac{1}{B} \sum_{b=1}^B \Delta(\mathbf{D}, \beta^{(b)})$ , where  $\Delta(\mathbf{D}, \beta^{(b)})$  represents the estimated value computed using the observed dataset  $\mathbf{D}$  and the parameter values from the  $b$ -th draw,  $\beta^{(b)}$ . Note that while we specify conjugate priors for the linear mixed model in this example, the framework is flexible and can accommodate a variety of prior choices. Previous work on Bayesian methods for CRTs has explored diverse prior specifications, for instance, log-uniform priors for cluster-level random effect variances and residual error variances (Spiegelhalter, 2001), fitted half-normal priors or informative beta priors for the ICC (Turner et al., 2001). These alternative priors can also be incorporated within our framework. For the point estimator derived from the model-robust standardization, under CRT randomization, we prove in Web Appendix 2, via empirical-process arguments (Van der

Vaart, 2000), that this estimator is model-robust and root- $M$  consistent, even under outcome model misspecification. This result is formalized in Proposition 1. Let  $\hat{\mu}_\omega^{\text{mr}}(a)$  denote the model-robust estimator of the arm-specific mean for treatment arm  $a$ , where  $\omega \in \{C, I\}$  indexes the cluster- and individual-level estimands, respectively. This estimator,  $\hat{\mu}_\omega^{\text{mr}}(a)$ , uses the working models to estimate the conditional means  $\mathbb{E}(\bar{Y}_i \mid A_i = a, \mathbf{X}_i, N_i)$  when  $\omega = C$  and  $\mathbb{E}(Y_{i+} \mid A_i = a, \mathbf{X}_i, N_i)$  when  $\omega = I$ , which enter the corresponding augmentation terms in Equations (4) and (5).

**Proposition 1** *Under the true law  $P_0$ , we assume there is a pointwise measurable class  $\mathcal{F}$  is  $P_0$ -Donsker with a square-integrable envelope. Then, under mild regularity conditions on the data-generating process stated in Web Appendix 1 and under cluster randomization, the model-robust point estimator  $\hat{\mu}_\omega^{\text{mr}}(a)$  achieves root- $M$  consistency:*

$$\hat{\mu}_\omega^{\text{mr}}(a) - \mu_\omega(a) = O_p(M^{-1/2}),$$

even under outcome model misspecification.

### 3.2 Calibrated variance estimation

When the model is misspecified, standard Bayesian posterior inference does not guarantee nominal frequentist coverage, because it accounts only for uncertainty in the posterior distribution of the misspecified outcome model parameters and not for additional sampling variability in the data (Antonelli et al., 2022). The model-robust formulation in Equations (4) and (5) incorporates both the parameter vector  $\beta$  from the conditional expectation and aggregated observed outcomes in  $\mathbf{D}$ , motivating a correction to the posterior variance to reflect sampling variability and recover nominal coverage.

Following Antonelli et al. (2022), we further quantify uncertainty in inferring causal estimands by combining the posterior distribution of parameters, conditional on the observed data, with a computationally efficient resampling procedure. Given the design of CRTs, we adopt a *cluster-level* resampling approach, extending the individual-level resampling proposed by Antonelli et al. (2022). First, we generate  $K$  new datasets,  $\mathbf{D}^{(1)}, \dots, \mathbf{D}^{(K)}$ , by sampling clusters with replacement from the empirical distribution of the observed clusters. For each bootstrapped datasets and posterior parameter samples, we compute  $\Delta(\mathbf{D}^{(k)}, \beta^{(b)})$  for  $k = 1, \dots, K$  and  $b = 1, \dots, B$ . Here,  $\beta^{(b)}$  represents a sample from the posterior distribution of  $\beta$  given the original observed data  $\mathbf{D}$ , and is fixed across different bootstrap samples. By calculating the posterior distribution of the estimator over multiple resampled datasets, we can estimate the potential variability from the data. Define  $G_\Delta = [\Delta(\mathbf{D}^{(k)}, \beta^{(b)})]_{k=1, \dots, K; b=1, \dots, B}$ , which is a matrix with  $K$  rows and  $B$  columns, where each entry represents the estimated value computed using the  $k$ -th resampled dataset and the  $b$ -th posterior draw of model parameters. For each  $k = 1, \dots, K$ , we approximate  $E_{\beta|D}[\Delta(\mathbf{D}^{(k)}, \beta)]$  by averaging over the  $B$  columns in row  $k$ . This results in a mean vector of length  $K$ , given by  $\frac{1}{B}G_\Delta\mathbf{1}$ , where  $\mathbf{1}$  denotes a vector of all ones of length  $B$ . We then estimate the variability using the sample variance of the  $K$  means, denoted by  $\text{Var}_{\mathbf{D}^{(k)}}\{E_{\beta|D}[\Delta(\mathbf{D}^{(k)}, \beta)]\}$ .

However, this variance has not yet accounted for the uncertainty that arises from potential variability in the posterior distribution of model parameters. Thus, the calibrated variance formula is  $\text{Var}(\widehat{\Delta}) = \text{Var}_{\mathbf{D}^{(k)}}\{E_{\beta|D}[\Delta(\mathbf{D}^{(k)}, \beta)]\} + \text{Var}_{\beta|D}[\Delta(\mathbf{D}, \beta)]$ . The first term quantifies variability arising from the data by evaluating the posterior distribution of the estimator across resampled datasets. The second term addresses the uncertainty arising from parameter variability, which is identical to the posterior variance in the uncalibrated posterior inference. For example, if we use this calibrated

posterior inference for the model-robust estimator for cluster-ATE using linear mixed model with covariate adjustment, we would execute the following steps.

1. Draw  $B$  posterior samples  $\{\boldsymbol{\beta}^{(b)}\}_{b=1}^B \sim \Pi(\boldsymbol{\beta} | \mathbf{D})$  from the fitted linear mixed model (Bayesian setting), where  $\mathbf{D} = \{\mathbf{Y}_i, A_i, N_i, \mathbf{X}_i\}_{i=1}^M$ .

2. Compute the point estimate.  $\widehat{\Delta}_C^{\text{mr}} \approx \frac{1}{B} \sum_{b=1}^B \widehat{\Delta}_C^{\text{mr}}(\mathbf{D}, \boldsymbol{\beta}^{(b)})$ , where  $\frac{1}{B} \sum_{b=1}^B f\left(\widehat{\mu}_C^{\text{mr}}\left(1, \mathbf{D}, \boldsymbol{\beta}^{(b)}\right), \widehat{\mu}_C^{\text{mr}}\left(0, \mathbf{D}, \boldsymbol{\beta}^{(b)}\right)\right)$ , where

$$\widehat{\mu}_C^{\text{mr}}\left(a, \mathbf{D}, \boldsymbol{\beta}^{(b)}\right) = \frac{1}{M} \sum_{i=1}^M \left\{ \frac{I(A_i = a) \left( \bar{Y}_i - \widehat{\mathbb{E}}^{(b)}(\bar{Y}_i | A_i = a, \mathbf{X}_i, N_i) \right)}{\pi^a (1 - \pi)^{1-a}} + \widehat{\mathbb{E}}^{(b)}(\bar{Y}_i | A_i = a, \mathbf{X}_i, N_i) \right\},$$

$$\widehat{\mathbb{E}}^{(b)}(\bar{Y}_i | A_i = a, \mathbf{X}_i, N_i) = \beta_0^{(b)} + \beta_1^{(b)} a + \left( \frac{1}{N_i} \sum_{j=1}^{N_i} \boldsymbol{\beta}_2^{(b)\top} \mathbf{X}_{ij} \right) + \beta_3^{(b)} N_i + \left( \frac{1}{N_i} \sum_{j=1}^{N_i} \boldsymbol{\beta}_4^{(b)\top} a \mathbf{X}_{ij} \right) + \beta_5^{(b)} a N_i.$$

3. Generate  $K$  new datasets,  $\mathbf{D}^{(1)}, \dots, \mathbf{D}^{(K)}$ , by resampling clusters with replacement from the empirical distribution of the data, where  $\mathbf{D}^{(k)} = \left\{ \mathbf{Y}_i^{(k)}, A_i^{(k)}, N_i^{(k)}, \mathbf{X}_i^{(k)} \right\}_{i=1}^M$ .

4. For each  $\mathbf{D}^{(k)}$ , compute the estimated cluster-ATE based on the model-robust estimator,  $\widehat{\Delta}_C^{\text{mr}}(\mathbf{D}^{(k)}, \boldsymbol{\beta}^{(b)})$ , for  $b \in \{1, \dots, B\}$ .  $\widehat{\Delta}_C^{\text{mr}}(\mathbf{D}^{(k)}, \boldsymbol{\beta}^{(b)}) = f\left(\widehat{\mu}_C^{\text{mr}}\left(1, \mathbf{D}^{(k)}, \boldsymbol{\beta}^{(b)}\right), \widehat{\mu}_C^{\text{mr}}\left(0, \mathbf{D}^{(k)}, \boldsymbol{\beta}^{(b)}\right)\right)$ , where

$$\widehat{\mu}_C^{\text{mr}}\left(a, \mathbf{D}^{(k)}, \boldsymbol{\beta}^{(b)}\right) = \frac{1}{M} \sum_{i=1}^M \left\{ \frac{I(A_i = a) \left( \bar{Y}_i^{(k)} - \widehat{\mathbb{E}}^{(b)}(\bar{Y}_i^{(k)} | A_i = a, \mathbf{X}_i^{(k)}, N_i^{(k)}) \right)}{\pi^a (1 - \pi)^{1-a}} + \widehat{\mathbb{E}}^{(b)}(\bar{Y}_i^{(k)} | A_i = a, \mathbf{X}_i^{(k)}, N_i^{(k)}) \right\},$$

$$\begin{aligned} \widehat{\mathbb{E}}^{(b)}(\bar{Y}_i^{(k)} | A_i = a, \mathbf{X}_i^{(k)}, N_i^{(k)}) &= \beta_0^{(b)} + \beta_1^{(b)} a + \left( \frac{1}{N_i^{(k)}} \sum_{j=1}^{N_i^{(k)}} \boldsymbol{\beta}_2^{(b)\top} \mathbf{X}_{ij}^{(k)} \right) + \beta_3^{(b)} N_i^{(k)} \\ &\quad + \left( \frac{1}{N_i^{(k)}} \sum_{j=1}^{N_i^{(k)}} \boldsymbol{\beta}_4^{(b)\top} a \mathbf{X}_{ij}^{(k)} \right) + \beta_5^{(b)} a N_i^{(k)}. \end{aligned}$$

5. The potential variability from the data is estimated by the variance of  $(\widehat{\Delta}_C^{\text{mr}}(\mathbf{D}^{(1)}, \boldsymbol{\beta}), \dots, \widehat{\Delta}_C^{\text{mr}}(\mathbf{D}^{(K)}, \boldsymbol{\beta}))$ , where  $\widehat{\Delta}_C^{\text{mr}}(\mathbf{D}^{(k)}, \boldsymbol{\beta}) \approx \frac{1}{B} \sum_{b=1}^B \widehat{\Delta}_C^{\text{mr}}(\mathbf{D}^{(k)}, \boldsymbol{\beta}^{(b)})$ .

6. The uncertainty arising from parameter variability is then estimated by computing the variance of posterior samples of the estimated cluster-ATE given the original observed dataset, which is the variance of  $(\widehat{\Delta}_C^{\text{mr}}(\mathbf{D}, \boldsymbol{\beta}^{(1)}), \dots, \widehat{\Delta}_C^{\text{mr}}(\mathbf{D}, \boldsymbol{\beta}^{(B)}))$ . This estimated variance is identical to the posterior variance in uncalibrated posterior inference.

7. The calibrated variance estimation is obtained by summing the estimated data variability and parameter variability from steps 6 to 7. With this calibrated variance estimate, the calibrated 95% credible interval can be derived using an asymptotic normal approximation.

This example illustrates the variance correction procedure for posterior inference of cluster-ATE using a covariate adjusted, model-robust estimator. This procedure can be easily adapted to both model-based and model-robust estimators, with or without leveraging covariates, for estimating cluster-ATE or individual-ATE. The following Proposition 2

establishes the consistency of the calibrated variance estimator under correct specification of the outcome regression model, whose posterior contracts at a sufficiently fast rate. The proof is provided in Web Appendix 1.

**Proposition 2** *Assuming cluster randomization, and that the outcome model contracts at a rate faster than  $M^{-1/4}$ . Then, under mild regularity conditions on the data-generating process and the posterior distribution, the variance estimator  $\widehat{V} = \text{Var}_{\mathbf{D}^{(k)}} \left\{ E_{\beta|\mathbf{D}} [\Delta(\mathbf{D}^{(k)}, \beta)] \right\} + \text{Var}_{\beta|\mathbf{D}} [\Delta(\mathbf{D}, \beta)]$  is consistent, i.e.,  $\widehat{V} - V = o_p(M^{-1})$ , where  $V = \text{Var}_{\mathbf{D}} \{ E_{\beta|\mathbf{D}} [\Delta(\mathbf{D}, \beta)] \}$ .*

Under randomized treatment assignment and a well-behaved outcome model, the calibrated variance estimator converges to the true variance as the number of clusters increases. More specifically, our proof in Web Appendix 1 shows that, when the outcome model is correctly specified, the contribution of posterior parameter uncertainty to the variability of the model-robust estimator becomes asymptotically negligible. Conversely, under model misspecification, the proposed variance estimator tends to overestimate the true variance, which can, at worst, lead to conservative inference when the point estimator's bias is small. To further evaluate the practical utility and robustness of this correction, we will empirically investigate the finite-sample performance of the calibrated variance estimator under several Bayesian working models in Section 4. Specifically, we first consider the case where the working outcome model is specified as a linear mixed model, which, while computationally convenient, may be prone to misspecification in complex data-generating processes. Under such misspecification, for instance when the model omits key nonlinearities, the calibrated variance for the model-robust estimator overestimates the variance, yielding valid but wider confidence intervals. By contrast, when BART's flexible model space successfully approximates the true data-generating mechanism, the calibrated variance estimator achieves consistency as the number of clusters increases. Accordingly, we examine mixed-effects BART, which accommodates nonlinearities and higher-order interactions and can mitigate misspecification bias by exploring a richer model space. Together with the linear mixed model analysis, these results provide empirical evidence on the reliability and generality of the calibrated variance estimator.

## 4 Simulation Studies

### 4.1 Simulation design

We conduct three simulation studies to evaluate the empirical performance of model-based g-computation and model-robust standardization estimators under combinations of covariate adjustment (with or without), outcome models (linear mixed model or BART), and posterior variance estimation methods (uncalibrated or calibrated), as summarized in Table 1. The first simulation study examines the performance of the model-based g-computation estimator combined with the uncalibrated posterior variance estimation for constructing credible intervals. We evaluate the impact of covariate adjustment on the efficiency and accuracy of estimating the cluster-ATE and the individual-ATE, and investigate the impact of outcome model misspecification on posterior inference. The second simulation study compares the model-based g-computation estimator and the model-robust standardization approach, both incorporating covariate adjustment and using a linear mixed model as the outcome model, under potential model misspecification. The third simulation study investigates the performance of Bayesian nonparametric methods under both model-based and model-robust frameworks, with and without posterior variance correction, across scenarios with model misspecification. We specifically focus on BART-based estimators to their linear mixed model counterparts, focusing on frequentist

coverage, relative bias, and robustness to model misspecification in small-sample settings. Methods corresponding to empty cells in Table 1 are not the focus of our simulation studies, as they are not expected to yield additional insights beyond those already covered by the selected combinations. All twelve methods (full combination of all factors) are compared in the application section.

Table 1: Categorization of methods. The table presents combinations of estimators (model-based and model-robust), posterior variance estimation methods (uncalibrated and calibrated), covariate adjustment (with or without), and outcome models (linear mixed model and BART). Cells marked with (1), (2), and (3) indicate the focus of the first, second, and third simulation studies, respectively.

Covariate adjustment	Outcome model	Model-based g-computation		Model-robust standardization	
		Uncalibrated	Calibrated	Uncalibrated	Calibrated
Unadjusted	Linear mixed model	(1)			
Adjusted	Linear mixed model	(1)&(2)	(2)	(2)	(2)
	BART	(3)			(3)

We describe our data-generating process as follows. The covariate-generating models are adapted from the simulation experiment in [Wang et al. \(2024b\)](#). Let  $(N_i, C_{i1}, C_{i2})$ , for  $i = 1, \dots, M$ , be independent draws from distribution  $\mathcal{P}^N \times \mathcal{P}^{C_1|N} \times \mathcal{P}^{C_2|N, C_1}$ , where  $\mathcal{P}^N$  is uniform over support  $(10, 50)$ ,  $\mathcal{P}^{C_1|N} = \mathcal{N}(N/10, 4)$ ,  $\mathcal{P}^{C_2|N, C_1} = \mathcal{B}[\text{expit}\{\log(N/10)C_1\}]$  is a Bernoulli distribution with  $\text{expit}(x) = (1 + e^{-x})^{-1}$ . The coefficient of variation of cluster size is approximately 0.394. Next, for each individual, we generate the two individual-level covariates from  $X_{ij1} \sim \mathcal{B}(N_i/50)$ , and  $X_{ij2} \sim \mathcal{N}\left\{\frac{(\sum_{j=1}^{N_i} X_{ij1}) \times (2C_{i2}-1)}{N_i}, 9\right\}$ . Then, we draw a cluster random intercept  $\phi_i \sim \mathcal{N}(0, \sigma_\phi^2)$  with  $\sigma_\phi^2 = 0.25$  and generate potential outcomes  $Y_{ij}(A_i) \sim \mathcal{N}(\eta_{ij}(A_i) + \phi_i, 1)$ , so that  $\text{ICC} = \sigma_\phi^2 / (\sigma_\phi^2 + 1) = 0.2$ . Here,  $\eta_{ij}(A_i)$  denotes the fixed effect component of the outcome-generating model, which varies across the designed scenarios. We consider three data-generating processes of increasing complexity: (i)  $\eta_{ij}(A_i) = 0.3A_i + \frac{N_i}{200} + \frac{C_{i1}}{3} + \frac{2C_{i2}-1}{5} + \frac{A_i C_{i2}}{10} + \frac{1}{4}(X_{ij1} + X_{ij2}) + \frac{A_i N_i}{200} + \frac{A_i X_{ij1}}{10}$ , (ii)  $\eta_{ij}(A_i) = -0.1 - \frac{0.3}{1+\exp(-6(N_i+C_{i1}+C_{i2}+X_{ij1}+X_{ij2}))} + 0.2X_{ij1}X_{ij2} + 0.3(\frac{N_i}{100} + C_{i1} + C_{i2})A_i - \frac{1}{1+\exp(-4(X_{ij1}+X_{ij2}))}A_i$ , and (iii)  $\eta_{ij}(A_i) = \frac{N_i A_i}{60} + \frac{1}{2} N_i \sin(C_{i1}) \cdot (2C_{i2}-1) + \frac{1}{2} e^{X_{ij1}} \cdot |X_{ij2}| + \frac{1}{4} A_i \cdot X_{ij2} \cdot \log(|C_{i1}|)$ . The observed outcomes are generated by independently sampling  $A_i \sim \mathcal{B}(0.5)$  and defining  $Y_{ij} = A_i Y_{ij}(1) + (1 - A_i) Y_{ij}(0)$ . Scenario (iii) is adapted from the continuous potential outcomes model in [Wang et al. \(2024b\)](#).

In each simulation study scenario, we consider a relatively small ( $M = 30$ ) or moderate ( $M = 60$  or  $M = 90$ ) number of clusters. In the first simulation study, each scenario is independently replicated  $R = 500$  times. We generate 3000 posterior draws using MCMC, discarding the first 1000 as burn-in. In the second study, which involves a more complex data-generating process, each scenario is replicated  $R = 1000$  times, with 2000 MCMC draws and the first 1000 discarded as burn-in. For the third simulation study, BART results are based on  $R = 250$  replications. We generate 10000 posterior draws using MCMC, discarding the first 5000 as burn-in. BART requires longer chains than linear mixed models to ensure convergence, due to its nonparametric and data-adaptive structure. Convergence is assessed using traceplots of the cluster-ATE and individual-ATE and the Geweke diagnostic. All simulations were performed using R version 4.3.1.

We evaluate the overall accuracy in the estimation of cluster-ATE and individual-ATE for each method using relative bias. There are  $B$  posterior samples of the model parameter generated for each simulated dataset  $r$ , and  $(\hat{\Delta}^{(r,1)}, \dots, \hat{\Delta}^{(r,B)})$  is the sequence of estimated posterior cluster-ATE or individual-ATE based on the simulated dataset  $r$ ,  $r \in \{1, \dots, R\}$ . The estimated value of cluster-ATE or individual-ATE of dataset  $r$  is  $\hat{\Delta}^{(r)} = \frac{1}{B} (\hat{\Delta}^{(r,1)} + \dots + \hat{\Delta}^{(r,B)})$ . The final estimate is the average of the posterior mean across all simulations,  $\bar{\Delta} = \frac{1}{R} \sum_{r=1}^R \hat{\Delta}^{(r)}$ . Then, relative bias is  $(\bar{\Delta} - \Delta) / \Delta$ , where the true value of the estimand  $\Delta$ , which is calculated from 100,000 clusters (Zhu et al., 2024). The relative bias evaluates the accuracy of the estimation in comparison to the actual value and provides a sense of how large the error is in relation to the size of the true ATE. The Monte Carlo Standard Deviation (MCSD) is given by  $\text{MCSD} = \sqrt{\frac{1}{R-1} \sum_{r=1}^R (\hat{\Delta}^{(r)} - \bar{\Delta})^2}$ , which measures the spread of the estimated values across different simulation runs, i.e., different Monte Carlo replications. A smaller MCSD indicates that the method is more efficient. The average estimated standard error (AESE) is defined as the average, across simulation replicates, of the posterior standard error estimates of  $\hat{\Delta}^{(r)}$ :  $\text{AESE} = \frac{1}{R} \sum_{r=1}^R \sqrt{\frac{1}{B-1} \sum_{b=1}^B (\hat{\Delta}^{(r,b)} - \hat{\Delta}^{(r)})^2}$ . The match between the MCSD and AESE suggests that the posterior standard error estimates are well-calibrated and appropriately reflect the sampling variability across replicates. In addition, we calculate the 95% credible interval coverage, using both calibrated and uncalibrated variance estimates, for each method to assess whether it achieves nominal frequentist coverage for the cluster-ATE and individual-ATE, thereby supporting valid estimand-aligned inference. Empirical coverage is computed as the proportion of 95% credible intervals that contain the true value of the estimand across all  $R$  replicated datasets. To compare the efficiency of estimators and the impact of leveraging or not leveraging covariates, we calculate the relative efficiency, which is dividing the estimated variance of the nonparametric estimator across all simulated datasets without covariate adjustment by the estimated variance of the estimator with covariate adjustment. The nonparametric unadjusted estimator is obtained by  $\hat{\mu}_C^{\text{np}}(a) = \frac{1}{M} \sum_{i=1}^M \left\{ \frac{I(A_i=a)\bar{Y}_i}{\pi^a(1-\pi)^{1-a}} \right\}$ , and  $\hat{\mu}_I^{\text{np}}(a) = \frac{1}{\sum_{i=1}^M N_i} \sum_{i=1}^M N_i \left\{ \frac{I(A_i=a)\bar{Y}_i}{\pi^a(1-\pi)^{1-a}} \right\}$ . This metric quantifies the percentage improvement in efficiency gained through covariate adjustment using either the g-computation or model-robust estimator. This metric also facilitates efficiency comparison between the g-computation and model-robust estimators.

## 4.2 Simulation results

### 4.2.1 Covariate adjustment in g-computation under potential model misspecification

The first simulation experiment focuses on the model-based g-computation estimator and the uncalibrated method for generating credible intervals, which use quantiles of the posterior samples. For the fitted model, we compare the performance of a linear mixed model that adjusts for covariates with main effects and treatment-covariate interactions, as specified in Equation (1), to the performance of an unadjusted linear mixed model. Table 2 summarizes the results for the first simulation experiment. Under data-generating process (i), the cluster-ATE  $\Delta_C = 0.590$  and the individual-ATE  $\Delta_I = 0.625$ , whereas under data-generating process (ii), the cluster-ATE  $\Delta_C = 0.679$  and the individual-ATE  $\Delta_I = 0.842$ .

Specifically, across both scenarios, the proposed g-computation estimator with covariate adjustment consistently outperforms its unadjusted counterpart in terms of efficiency. This improvement highlights how leveraging covariate information can refine outcome prediction, leading to tighter posterior samples and more efficient estimation. In scenario (i), where the linear mixed model is the true underlying outcome model, both g-computation estimators,

Table 2: Results in the first simulation experiment from 500 simulations.  $\Delta_C$  denotes true value of cluster-average treatment effect.  $\Delta_I$  denotes true value of individual-average treatment effect. “Adjusted” indicates that the fitted model is a linear mixed model that includes covariate main effects and treatment–covariate interactions. “Unadjusted” indicates that the linear mixed model includes only a treatment indicator.  $M$  denotes the number of clusters. “Rel. Bias” refers to the relative bias of the estimated value expressed as a percentage, with standard errors shown in parentheses. “Coverage” refers to the 95% credible interval coverage across all simulated datasets expressed as a percentage. “MCSD” denotes Monte Carlo standard deviation. “AESE” denotes average of estimated standard error.

Data-Generating Model	Covariate	M	Estimand	Rel. Bias (SE)	Coverage	MCSD	AESE
Scenario (i) $\Delta_C = 0.590$ $\Delta_I = 0.625$	Adjusted	30	cluster-ATE	-2.1 (1.7)	95.0	0.226	0.247
			individual-ATE	-1.5 (1.7)	95.4	0.244	0.255
		60	cluster-ATE	-2.8 (1.1)	95.0	0.145	0.146
			individual-ATE	-3.1 (1.1)	94.4	0.151	0.154
		90	cluster-ATE	-1.9 (0.8)	95.0	0.112	0.116
	Unadjusted		individual-ATE	-1.5 (0.9)	94.0	0.121	0.122
		30	cluster-ATE	-1.9 (4.9)	95.6	0.652	0.681
			individual-ATE	-7.5 (4.7)	95.6	0.652	0.681
		60	cluster-ATE	0.9 (3.4)	95.0	0.449	0.471
			individual-ATE	-4.8 (3.2)	95.2	0.449	0.471
Scenario (ii) $\Delta_C = 0.679$ $\Delta_I = 0.842$	Adjusted	30	cluster-ATE	0.9 (2.9)	93.6	0.388	0.383
			individual-ATE	-4.8 (2.8)	93.6	0.388	0.383
		60	cluster-ATE	-3.8 (2.3)	82.0	0.354	0.260
			individual-ATE	-3.8 (2.0)	82.8	0.372	0.266
		90	cluster-ATE	-2.6 (1.5)	83.2	0.222	0.156
	Unadjusted		individual-ATE	-3.6 (1.2)	83.6	0.227	0.162
		30	cluster-ATE	-1.8 (1.2)	83.0	0.181	0.124
			individual-ATE	-1.9 (1.0)	80.2	0.193	0.129
		60	cluster-ATE	-2.6 (2.9)	93.4	0.435	0.428
			individual-ATE	-21.5 (2.3)	91.2	0.435	0.428
	Unadjusted	90	cluster-ATE	0.7 (1.9)	94.0	0.281	0.294
			individual-ATE	-18.8 (1.5)	92.2	0.281	0.294
		30	cluster-ATE	1.3 (1.6)	93.2	0.250	0.240
			individual-ATE	-18.3 (1.3)	87.2	0.250	0.240

with and without covariate adjustment, show negligible bias and achieve nominal coverage. However, in scenario (ii), the fitted linear mixed models fail to account for the non-linear transformations and interactions inherent in the outcome-generating model, leading to a misspecified parametric working model. In this case, the unadjusted linear mixed model yields a biased point estimator but occasionally achieves coverage closer to the nominal level. In contrast, the covariate adjusted model produces a less biased point estimator, yet its credible intervals consistently fall short of nominal coverage. It suggests that in the presence of informative cluster size, adjusting for covariates, even with a misspecified model, can improve efficiency but may compromise nominal coverage. Conversely, not adjusting for covariates may achieve nominal coverage in some cases by chance, at the cost of efficiency and accuracy. Therefore, without further modification of the estimators and the inferential approach, these commonly used methods fail to fulfill the promise of estimand-aligned inference under model misspecification in CRTs.

#### 4.2.2 Improving estimation and inference under model misspecification: comparison of point estimator and posterior variance correction combinations

The second simulation experiment compares four combinations of point estimators and posterior variance estimation methods: the model-based g-computation and model-robust estimators, each paired with either the uncalibrated or the calibrated variance estimation method. A linear mixed model with covariate adjustment is used as the fitted model throughout this experiment. Our goal is to resolve the issue shown in the first experiment by combining a model-robust estimator and a calibrated posterior variance estimation method. To summarize the simulation results, Table A in Web Appendix 3 reports numerical results for data-generating process (i), under which the linear mixed model is the true underlying outcome model. Tables B and C in Web Appendix 3 present the precise numerical results for data-generating processes (ii) and (iii), which involve non-linear relationships between the covariates and the outcomes, and potentially more complicated interactions, such as exponential transformations and trigonometric functions. To facilitate comparison, Figure 1 displays the visualization of simulation results across all three scenarios. It includes three panels summarizing relative bias, the ratio of AESE to MCSD, and empirical coverage of 95% credible intervals. The relative bias panel includes results only for the two point estimators, as bias is not affected by the posterior variance estimation. Web Appendix 6 provides the relative efficiency comparisons for the proposed point estimators across all three scenarios, evaluating the efficiency gains of covariate adjusted g-computation and model-robust estimators relative to the unadjusted moment estimator.

For scenario (i), Figure 1 shows that both the g-computation and model-robust estimators yield negligible bias and achieve nominal coverage for both the cluster-ATE and individual-ATE when paired with calibrated variance estimation. In contrast, the uncalibrated variance estimation method for constructing credible intervals, based on posterior quantiles, fails to achieve nominal frequentist coverage for the model-robust estimator, although it performs adequately when paired with the g-computation estimator. This difference arises because the g-computation estimator depends solely on the outcome model parameters. When the outcome model is correctly specified, the posterior distribution from Bayesian estimation appropriately reflects the uncertainty. However, the model-robust estimator involves standardization that depends on both model parameters and the observed data, introducing additional variability that the uncalibrated posterior variance fails to capture. Regarding the AESE-to-MCSD ratio, all combinations yield values close to one under moderate sample sizes, except for the model-robust estimator paired with uncalibrated posterior inference, which systematically underestimates uncertainty. Notably, the model-robust estimator with calibrated posterior inference achieves AESE-to-MCSD ratios near one even in small-sample settings, demonstrating accurate calibration of uncertainty despite limited data.

For scenarios (ii) and (iii), where the working model is incorrectly specified, the model-robust estimator shows smaller, or at least comparable, relative bias in estimating both the individual-ATE and cluster-ATE compared to the g-computation estimator. Furthermore, only the model-robust estimator, when paired with the calibrated variance estimation method, consistently achieves nominal coverage for both cluster-ATE and individual-ATE, thus supporting robust estimand-aligned inference in CRTs under outcome model misspecification. This is also reflected in the AESE-to-MCSD ratios in Figure 1, where only the model-robust estimator with calibrated inference consistently yields values near one under scenarios (ii) and (iii), indicating well-calibrated uncertainty estimates that align with the empirical variability observed across replications. The model-based g-computation estimator depends solely on model parameters

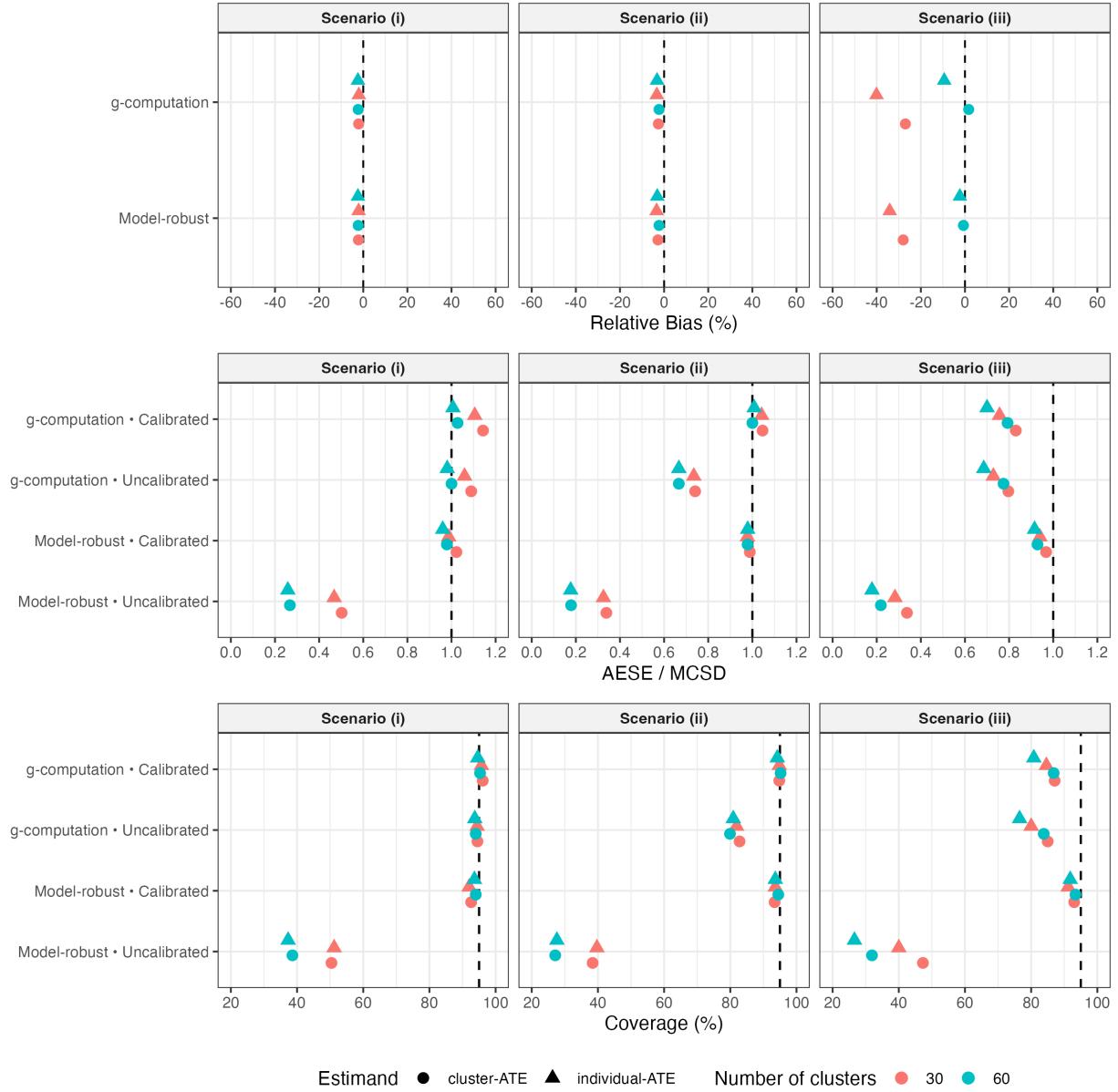


Figure 1: Comparison of relative bias, AESE-to-MCSD ratio, and coverage of 95% credible intervals from the second simulation experiment based on 1000 replications for scenario (i), (ii), and (iii), estimating the cluster-average treatment effect (cluster-ATE) and individual-average treatment effect (individual-ATE). A linear mixed model, which adjusts for covariates with main effects and treatment–covariate interactions, is used as the fitted model. The estimation methods are categorized into g-computation and model-robust approaches, each with calibrated and uncalibrated posterior inference. For the calibrated variance estimation, 100 resampled datasets are used within each simulation to compute posterior variance. Shapes distinguish estimands (circles for cluster-ATE and triangles for individual-ATE), and colors indicate the number of clusters ( $M = 30$  in red,  $M = 60$  in blue). Reference lines are included for each metric: dashed at 0 for relative bias (no systematic over or underestimation), dashed at 1 for AESE-to-MCSD ratio (perfect calibration where AESE equals MCSD), and dashed at 95% for coverage (indicating the nominal 95% level).

and not on the observed outcome data. Therefore, when the outcome model is correctly specified, the uncalibrated posterior inference can still provide valid uncertainty quantification. However, the g-computation estimator becomes invalid when the outcome model is misspecified. In more complex data-generating processes, this misspecification leads to undercoverage of the credible intervals. To maintain small bias and achieve nominal coverage under complex data-generating processes, the model-robust estimator should be paired with calibrated posterior variance estimation to account for both the observed data variability and potential parametric model misspecification.

Finally, as shown in Web Appendix 6, the relative efficiency of both the model-based and model-robust estimators relative to the unadjusted nonparametric estimator is always greater than 1 across all data-generating processes, further confirming that incorporating covariate information improves estimation efficiency. For scenario (i), Web Appendix 6 demonstrates that both the g-computation and model-robust estimators achieve substantial efficiency gains relative to the unadjusted moment estimator, with comparable levels of improvement. When the working model is incorrectly specified, i.e, scenario (ii) and (iii), the relative efficiency of the model-robust estimator consistently outperforms or is at least comparable to that of the g-computation estimator.

#### 4.2.3 Leveraging BART for estimating cluster-ATE and individual-ATE

While linear mixed models provide a widely used parametric framework for estimating treatment effects in CRTs, their validity can be sensitive to model misspecification. In complex settings involving non-linear relationships or high-dimensional covariates, misspecification of the outcome model can lead to biased or inefficient inference. BART (Chipman et al., 2010) is a nonparametric Bayesian ensemble method that constructs a sum-of-trees model and mitigates concerns about parametric misspecification. It consists of two key components: an ensemble of regression trees summed to model outcome mean, and a regularization prior that constrains the tree parameters to prevent overfitting (Chipman et al., 2010). Recent extensions such as mixed-effects BART (Spanbauer and Sparapani, 2021) incorporate cluster-specific random intercepts into the BART framework to account for within-cluster correlation in clustered data. It combines the flexibility of BART to model the fixed effect portion with hierarchical modeling, allowing it to accommodate both non-linearity and within-cluster dependence in CRTs. Suppose that the final outcome  $Y_{ij}$  is related to generic covariates  $A_i$ ,  $\mathbf{X}_{ij}$ , and  $N_i$ , the mixed-effects BART is given by  $Y_{ij} = \sum_{l=1}^L h(A_i, \mathbf{X}_{ij}, N_i; T_l, \Lambda_l) + \phi_i + \epsilon_{ij}$ , where  $\phi_i \sim \mathcal{N}(0, \sigma_\phi^2)$  represents the cluster-level random effect, and  $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$  denotes the individual-level residual error. The mean function is modeled as a sum of  $L$  regression trees, expressed as  $\sum_{l=1}^L h(A_i, \mathbf{X}_{ij}, N_i; T_l, \Lambda_l)$ , where each  $T_l$  represents the structure of the  $l$ -th binary tree, and  $\Lambda_l$  denotes a set of parameters values associated with each of the terminal nodes in tree  $T_l$  (Chipman et al., 2010). The function  $h(\cdot)$  assigns a terminal node value in  $\Lambda_l$  to  $\{A_i, \mathbf{X}_{ij}, N_i\}$  according to binary tree  $T_l$ . Increasing  $L$  allows for greater model flexibility, though it also imposes higher computational costs. A common default of  $L = 50$  trees has been found to provide adequate performance in practice, as suggested by Bleich et al. (2014). This formulation mirrors the linear mixed model structure while replacing the fixed component with a data-adaptive ensemble model, allowing for flexible modeling of fixed effects within the hierarchical framework. In the following, we implement this mixed-effects BART model using the `mxbart` function from the R package `mxBART` (Spanbauer and Sparapani, 2020).

Using mixed-effects BART with covariate adjustment as the working outcome models, Figure 2 displays results from BART-based methods under scenarios (ii) and (iii), characterized by non-linear outcome models and informative cluster size. Results are shown for g-computation with uncalibrated posterior inference and for model-robust standardization

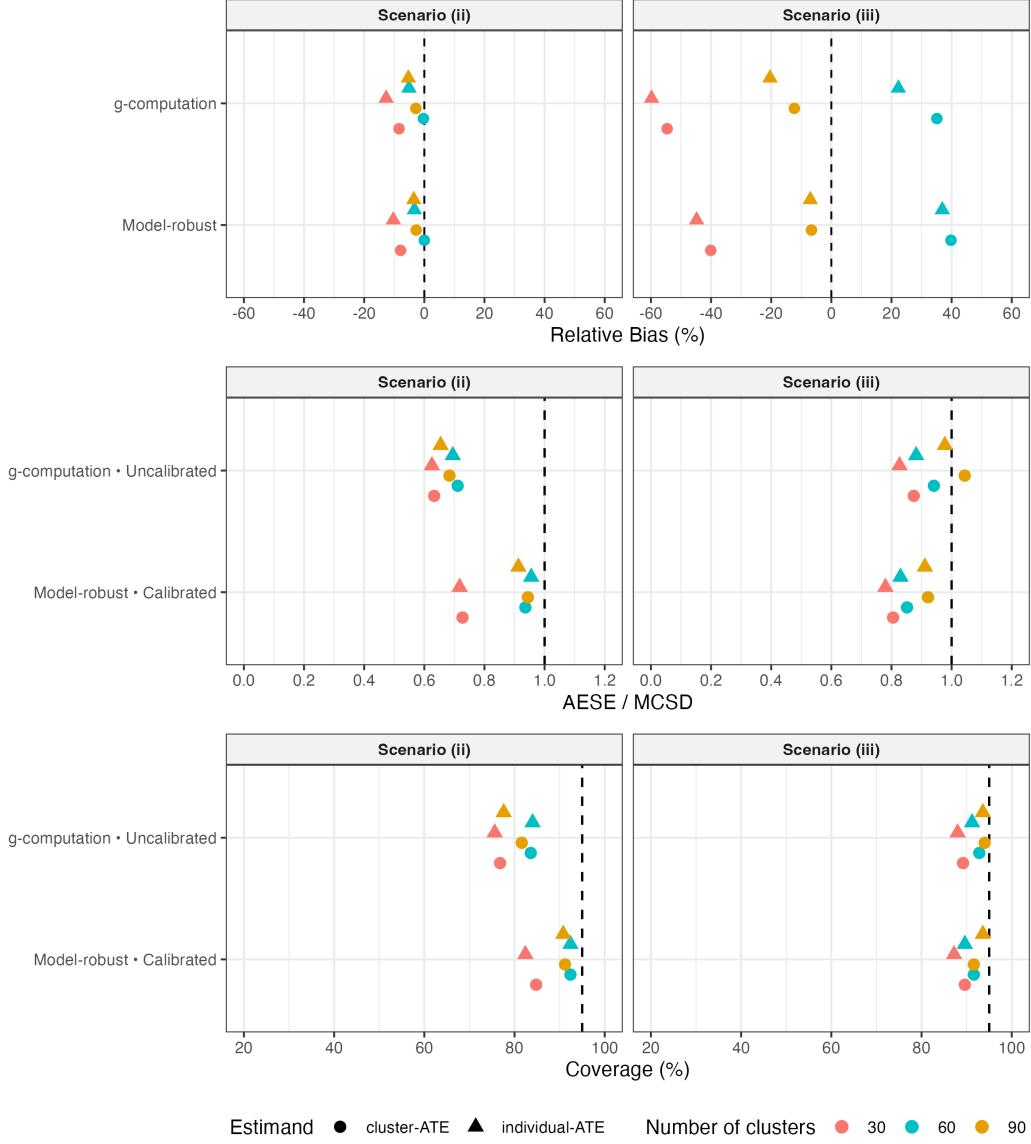


Figure 2: Comparison of relative bias, AESE-to-MCSD ratio, and coverage of 95% credible intervals from the third simulation experiment based on 250 replications for scenario (ii), and (iii), estimating the cluster-average treatment effect (cluster-ATE) and individual-average treatment effect (individual-ATE). A mixed-effects BART is used as the fitted model. The estimation methods are categorized into g-computation with uncalibrated posterior inference and model-robust standardization with calibrated posterior inference. For the calibrated posterior inference, 100 resampled datasets are used within each simulation to compute variance. Shapes distinguish estimands (circles for cluster-ATE and triangles for individual-ATE), and colors indicate the number of clusters ( $M = 30$  in red,  $M = 60$  in blue,  $M = 90$  in orange). Reference lines are included for each metric: dashed at 0 for relative bias (no systematic over or underestimation), dashed at 1 for AESE-to-MCSD ratio (perfect calibration where AESE equals MCSD), and dashed at 95% for coverage (indicating the nominal 95% level).

with calibrated posterior inference using 100 resampled datasets. We include a larger number of clusters ( $M = 90$ ) in this simulation, as tree-based ensemble methods like BART usually require larger sample sizes to stabilize posterior inference and reduce variability in predictions. Web Appendix 4 presents the precise numerical results corresponding to these BART-based simulations.

We first compare the performance of the linear mixed model and BART, each combined with the model-based g-computation estimator and uncalibrated posterior inference. While BART paired with g-computation and uncalibrated intervals occasionally attains close to nominal coverage, this is not consistently guaranteed. In contrast, the linear mixed model under the same estimation strategy consistently fails to achieve nominal coverage under model misspecification. Broadly speaking, when combined with g-computation and uncalibrated intervals, neither the linear mixed model nor BART supports nominal coverage under non-linear data-generating processes, under informative cluster size. Moreover, in small-sample settings,  $M = 30$  or  $M = 60$ , BART shows larger relative bias compared to the linear mixed model, and does not consistently demonstrate efficiency gains as measured by MCSD shown in Web Appendix 4. Thus, while BART combined with g-computation and uncalibrated inference may offer some improvement in coverage over its linear mixed model counterpart, it still falls short when compared to the linear mixed model paired with the model-robust estimator and calibrated variance estimation, which delivers more reliable coverage performance.

We next compare BART and the linear mixed model when both are used in conjunction with model-robust standardization and calibrated variance estimation. In the presence of informative cluster size and non-linear data-generating processes, BART combined with model-robust standardization and calibrated intervals can approach nominal coverage when the number of clusters is at least moderate. However, the linear mixed model paired with this combination more consistently achieves nominal coverage, even in small-sample settings. Although both approaches are embedded under the model-robust standardization formula, the linear mixed model's parametric structure can avoid over-fitting and helps stabilize variance estimation. In contrast, BART's flexible and data-adaptive nature may lead to overfitting in small-sample settings, resulting in underestimated uncertainty. This issue is addressed as the number of clusters (i.e., independent units) increases to  $M = 90$ .

Focusing specifically on scenarios (ii) and (iii), which involve informative cluster size, comparing BART combined with g-computation and uncalibrated inference to BART combined with model-robust approach and calibrated inference, the latter usually achieves coverage close to the nominal level in cases where the former does not. We then explore whether informative cluster size (effect modification at the cluster level) plays a key role in this simulation comparison. Hence we design and evaluate a scenario without informative cluster size, with details of the data-generating process and corresponding mixed-effects BART results provided in Web Appendix 5. Interestingly, when the cluster size is non-informative, both approaches yield small bias and coverage close to the nominal coverage, even with a small number of clusters. These results demonstrate the importance of using model-robust standardization with calibrated inference to properly account for informative cluster size and to achieve estimand-aligned inference, even when using flexible machine learning models such as BART. In the presence of informative cluster size, the model-robust standardization formula coupled with linear mixed model may be preferred as the calibrated posterior inference is more likely to achieve nominal coverage in finite samples.

Table 3: Analysis of the Pain Program for Active Coping and Training (PPACT) data. Each analysis using the linear mixed model (LMM) as the fitted model was performed with a total of 2,000 MCMC iterations. The first 1,000 iterations were discarded as burn-in. Each analysis using BART as the fitted model was performed with a total of 10,000 MCMC iterations, with the first 5,000 discarded as burn-in. Convergence was assessed using traceplots of the cluster-ATE and individual-ATE as well as the Geweke diagnostic.

Fitted Model	Covariate	Estimator	Posterior Inference	cluster-ATE		individual-ATE	
				Estimate	95% C.I.	Estimate	95% C.I.
LMM	Unadjusted	g-computation	Uncalibrated	-0.627	(-0.977, -0.274)	-0.627	(-0.977, -0.274)
			Calibrated	-0.627	(-0.984, -0.269)	-0.627	(-0.984, -0.269)
	Adjusted	Model-robust	Uncalibrated*	-0.661	(-0.661, -0.661)	-0.616	(-0.622, -0.610)
			Calibrated	-0.661	(-1.082, -0.241)	-0.616	(-0.990, -0.242)
BART	Adjusted	g-computation	Uncalibrated	-0.612	(-0.980, -0.254)	-0.530	(-0.878, -0.188)
			Calibrated	-0.612	(-0.991, -0.233)	-0.530	(-0.888, -0.171)
	Unadjusted	Model-robust	Uncalibrated	-0.594	(-0.658, -0.530)	-0.520	(-0.566, -0.472)
			Calibrated	-0.594	(-0.954, -0.235)	-0.520	(-0.849, -0.191)

\* When using the model-robust estimator without covariate adjustment and without posterior variance correction, the estimated cluster-ATE remains identical across all posterior draws. This occurs because, given that the observed data have equal numbers of clusters in each intervention arm, the model-robust estimator reduces to the nonparametric unadjusted estimator specified in Section 4.1. This explains the row marked with an asterisk, where the 95% credible interval has zero width, with upper and lower bounds coinciding with the point estimate.

## 5 Illustrative data application

We demonstrate our methods to analyze data from the Pain Program for Active Coping and Training (PPACT) (DeBar et al., 2022). PPACT is a pragmatic CRT evaluating the effect of a cognitive behavioral therapy intervention embedded in primary care versus usual care for treating long-term opioid users with chronic pain. The study included adult health plan members from participating primary care providers with at least six months of continuous enrollment, a pain diagnosis within the past year, and long-term opioid use, defined as a supply of short-acting opioids for more than 90 days over a 120-day period, or at least two dispensations of long-acting opioids in the past 180 days. The study involves a total of 106 clinic-based clusters, with 53 randomized to the intervention and 53 to usual care in a 1:1 ratio, and cluster sizes vary from 3 to 13. The primary hypothesis is that patients receiving the intervention would have a greater reduction in pain impact compared to those receiving usual care. The primary outcome is the self-reported PEGS score at 12 months, a continuous measure that assesses pain impact as a composite of pain intensity and interference with enjoyment of life, general activity, and sleep. The baseline covariates we consider include cluster size, patient age at randomization, gender, disability, current smoking status, body mass index, alcohol and drug abuse diagnosis in 6 months prior to randomization, comorbidity defined as having two or more specified chronic medical conditions, depression diagnosis in 6 months prior to randomization, number of different pain types, average morphine milligram equivalents dose per day in 6 months prior to randomization, and an indicator for whether this average daily dose exceeded 90. Our analysis includes a total of 705 individuals from the 106 clusters, selected for having complete data on all relevant baseline covariates and outcomes.

We apply twelve combinations of Bayesian methods to analyze the PPACT data, including g-computation and model-robust estimators, with and without posterior variance correction, under both linear mixed models and mixed-effects BART. For linear mixed models, both covariate-adjusted and unadjusted specifications are considered, whereas BART uses only the adjusted specification, since excluding covariates would hinder its ability to capture nonlinear and interaction effects. Convergence is evaluated using traceplots of the cluster-ATE and individual-ATE and the Geweke diagnostic. The results from all twelve methods are summarized in Table 3. Across all twelve methods, we observe that the 95% credible intervals for both the cluster-ATE and individual-ATE consistently exclude zero, suggesting that the cognitive behavioral therapy is effective in reducing pain impact among long-term opioid users with chronic pain.

We next interpret the findings presented in Table 3. The unadjusted g-computation approach, the least reliable in our simulations, gives the same point estimates for the cluster-ATE and individual-ATE. This masks the distinction between the definitions of the two estimands, clearly indicating that the method is not estimand-aligned under informative cluster size. In contrast, the most trustworthy strategy supported by our simulation is the model-robust standardization estimator with covariate adjustment and calibrated posterior variance. Both linear mixed models and BART can be appropriate outcome models, especially given the relatively large number of clusters, which supports stable estimation even with BART. Under these preferred methods, the difference between the estimated cluster-ATE and individual-ATE becomes more apparent.

Focusing on the best-performing strategy, model-robust estimator with covariate adjustment and calibrated posterior inference, we find that BART and linear mixed model yield comparable results. Specifically, while the 95% credible intervals are slightly wider under the linear mixed model, the point estimates and interval bounds are quite close to those obtained using BART, suggesting that a linear specification may adequately approximate the true underlying data-generating process. Using a linear mixed model, the estimated cluster-ATE is  $-0.594$  with a 95% credible interval of  $(-0.954, -0.235)$ , and the individual-ATE is  $-0.520$  with a 95% credible interval of  $(-0.849, -0.191)$ . With BART, the estimated cluster-ATE is  $-0.588$  with a 95% credible interval of  $(-0.926, -0.250)$ , and the individual-ATE is  $-0.526$  with a 95% credible interval of  $(-0.837, -0.215)$ . These results indicate that, when accounting for the appropriate unit of inference, the expected reduction in pain impact, measured by the PEGS score at 12 months due to the cognitive behavioral therapy intervention, is  $0.588$  at the cluster level and  $0.526$  at the individual level. Both reductions are statistically significant and show the effectiveness of the cognitive behavioral therapy among long-term opioid users with chronic pain.

## 6 Discussion

Despite the growing importance of CRTs in health research, Bayesian analyses of such trials have often focused on model-based parameters of hierarchical models rather than on explicitly defined causal estimands. This disconnect can lead to ambiguity in interpretation, as posterior inferences may implicitly target quantities that depend on model specification rather than on the intended unit of inference. To address this methodological gap, we develop estimand-aligned Bayesian inferential strategies for CRTs that preserve the causal interpretation of treatment effects. Building on the distinctions proposed by Kahan et al. (2023), we formalize two target estimands, the cluster-ATE and the individual-ATE, which diverge in the presence of informative cluster size. We then survey different Bayesian estimators that integrates the g-computation formula and model-robust standardization to estimate these estimands, thereby

improving the calibration of Bayesian inference to achieve frequentist-valid coverage. Through extensive simulation studies, we evaluate the finite-sample performance of the proposed estimators under varying outcome models, covariate adjustment strategies, and posterior variance estimation methods. Our findings highlight the appeal of combining Bayesian hierarchical models with model-robust standardization formula to target model-free estimands, particularly in the presence of informative cluster size. By leveraging this formulation, we extended the approach in [Antonelli et al. \(2022\)](#) to generate a calibrated variance estimator that jointly accounts for parameter uncertainty and sampling variability, thereby improving the frequentist calibration of Bayesian credible intervals. Simulation results demonstrate that, even under misspecified parametric working models, the model-robust estimator derived from a Bayesian linear mixed model achieves valid estimand-aligned inference for both the cluster-ATE and individual-ATE when combined with the calibrated variance approach. For applied analysts, these findings suggest that standard Bayesian hierarchical models can be extended to deliver causally interpretable and well-calibrated inference simply by post-processing posterior samples through model-robust standardization and variance correction, offering a practical route toward principled estimand-aligned Bayesian analysis of CRTs.

In our numerical studies, we also consider Bayesian nonparametric models to enhance flexibility in modeling outcomes and dependencies among observations. Specifically, mixed-effects BART ([Spanbauer and Sparapani, 2021](#)) combines the standard BART framework for fixed effects with hierarchical random effects, providing a natural extension for CRTs. Our simulation studies show that mixed-effects BART, when combined with model-robust standardization and calibrated variance estimation, achieves small bias and near-nominal coverage for both cluster-ATE and individual-ATE, even under model misspecification and informative cluster size, but particularly when the number of clusters is reasonably large. However, in small-sample settings, the Bayesian linear mixed model with the same estimation strategy tends to achieve more stable frequentist coverage, likely due to its parametric structure stabilizing variance estimation, whereas BART's flexibility may induce mild overfitting and underestimated uncertainty. By contrast, when g-computation is used with uncalibrated posterior inference, neither the linear mixed model nor BART attains nominal coverage under model misspecification. These findings further underscore the importance of coupling model-robust standardization with calibrated variance estimation to ensure estimand-aligned inference under informative cluster size, even when leveraging flexible Bayesian nonparametric models. Future work should further investigate the strengths of mixed-effects BART for CRTs with larger numbers of clusters or high-dimensional covariates, and explore alternative posterior calibration methods, such as the Bayesian bootstrap approach ([Yiu et al., 2025](#)), that provide valid inference for low-dimensional functionals while retaining the adaptivity of nonparametric Bayesian models.

The broader motivation of this work resonates with the Calibrated Bayes framework articulated by [Rubin \(1984\)](#) and [Little \(2006\)](#), which advocates Bayesian reasoning in principle yet calibration to real-world frequencies in practice. As Rubin discussed in his 1984 article ([Rubin, 1984](#)), “*frequency calculations are useful for making Bayesian statements scientific, scientific in the sense of capable of being shown wrong by empirical test; here the technique is the calibration of Bayesian probabilities to the frequencies of actual events... The applied statistician should be Bayesian in principle and calibrated to the real world in practice—appropriate frequency calculations help to define such a tie*”. [Little \(2006\)](#) later formalized this synthesis as a statistical philosophy that “*uses frequentist methods for model assessment and Bayesian methods for inference under the model*”. In this spirit, our proposed estimand-aligned Bayesian approach offers a calibrated option that reconciles Bayesian hierarchical inference with frequentist performance guarantees, rather than prescribing calibration as a universal tool, in the context of CRTs. By explicitly aligning posterior inference with

model-free estimands and empirical coverage properties, we achieve model-robust analysis under cluster randomization within a Bayesian framework. We believe this work enriches the Bayesian presence in the realm of model-assisted analysis of randomized experiments—an area where Bayesian methods are sparsely discussed but can be valued for their flexibility, coherence, and ability to integrate prior and other design-based information.

## Acknowledgements

Research in this article was supported by a Patient-Centered Outcomes Research Institute Award® (PCORI® Award ME-2022C2-27676). The statements presented in this article are solely the responsibility of the authors and do not necessarily represent the official views of PCORI®, its Board of Governors, or the Methodology Committee.

## Data Availability Statement

The sample R code to implement the proposed methods can be found at the GitHub website at <https://github.com/Ruyi-Liu/Model-robust-Bayesian-inference-in-cluster-randomized-trials>. The PPACT data that were used to illustrate the proposed approach in Section 5 are publicly available at <https://github.com/PainResearch/PPACT>.

## References

- Antonelli, J., Papadogeorgou, G., and Dominici, F. (2022). Causal inference in high dimensions: a marriage between bayesian modeling and good frequentist properties. *Biometrics*, 78(1):100–114.
- Balzer, L. B., van der Laan, M., Ayieko, J., Kamya, M., Chamie, G., Schwab, J., Havlir, D. V., and Petersen, M. L. (2023). Two-stage tmle to reduce bias and improve efficiency in cluster randomized trials. *Biostatistics*, 24(2):502–517.
- Balzer, L. B., Zheng, W., van der Laan, M. J., and Petersen, M. L. (2019). A new approach to hierarchical data analysis: Targeted maximum likelihood estimation for the causal effect of a cluster-level exposure. *Statistical Methods in Medical Research*, 28(6):1761–1780.
- Benitez, A., Petersen, M. L., van der Laan, M. J., Santos, N., Butrick, E., Walker, D., Ghosh, R., Otieno, P., Waiswa, P., and Balzer, L. B. (2023). Defining and estimating effects in cluster randomized trials: a methods comparison. *Statistics in Medicine*, 42(19):3443–3466.
- Bleich, J., Kapelner, A., George, E. I., and Jensen, S. T. (2014). Variable selection for bart: an application to gene regulation. *The Annals of Applied Statistics*, 8(3):1750–1781.
- Carlin, B. P. and Louis, T. A. (1997). Bayes and empirical bayes methods for data analysis.
- Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.
- Clark, A. B. and Bachmann, M. O. (2010). Bayesian methods of analysis for cluster randomized trials with count outcome data. *Statistics in Medicine*, 29(2):199–209.

- DeBar, L., Mayhew, M., Benes, L., Bonifay, A., Deyo, R. A., Elder, C. R., Keefe, F. J., Leo, M. C., McMullen, C., Owen-Smith, A., et al. (2022). A primary care-based cognitive behavioral therapy intervention for long-term opioid users with chronic pain: a randomized pragmatic trial. *Annals of Internal Medicine*, 175(1):46–55.
- Donner, A., Klar, N., and Klar, N. S. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*, volume 27. Arnold London.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.
- Grantham, K. L., Kasza, J., Heritier, S., Carlin, J. B., and Forbes, A. B. (2022). Evaluating the performance of bayesian and restricted maximum likelihood estimation for stepped wedge cluster randomized trials with a small number of clusters. *BMC Medical Research Methodology*, 22(1):112.
- Greenland, S. (1982). Interpretation and estimation of summary ratios under heterogeneity. *Statistics in Medicine*, 1(3):217–227.
- Hong, L. and Martin, R. (2020). Model misspecification, bayesian versus credibility estimation, and gibbs posteriors. *Scandinavian Actuarial Journal*, 2020(7):634–649.
- Jones, B. G., Streeter, A. J., Baker, A., Moyeed, R., and Creanor, S. (2021). Bayesian statistics in the design and analysis of cluster randomised controlled trials and their reporting quality: a methodological systematic review. *Systematic Reviews*, 10:1–14.
- Kadane, J. B. (1995). Prime time for bayes. *Controlled Clinical Trials*, 16(5):313–318.
- Kahan, B. C., Blette, B. S., Harhay, M. O., Halpern, S. D., Jairath, V., Copas, A., and Li, F. (2024). Demystifying estimands in cluster-randomised trials. *Statistical Methods in Medical Research*, 33(7):1211–1232.
- Kahan, B. C., Li, F., Copas, A. J., and Harhay, M. O. (2023). Estimands in cluster-randomized trials: choosing analyses that answer the right question. *International Journal of Epidemiology*, 52(1):107–118.
- Li, F., Ding, P., and Mealli, F. (2023). Bayesian causal inference: a critical review. *Philosophical Transactions of the Royal Society A*, 381(2247):20220153.
- Li, F., Tong, J., Cheng, C., Fang, X., Kahan, B. C., and Wang (2025). Model-robust standardization in cluster-randomized trials. *Statistics in Medicine*.
- Little, R. J. (2006). Calibrated bayes: a bayes/frequentist roadmap. *The American Statistician*, 60(3):213–223.
- Murray, D. M. (1998). *Design and Analysis of Group-Randomized Trials*, volume 29. Monographs in Epidemiology and.
- Nugent, J. R., Marquez, C., Charlebois, E. D., Abbott, R., and Balzer, L. B. (2024). Blurring cluster randomized trials and observational studies: Two-stage tmle for subsampling, missingness, and few independent units. *Biostatistics*, 25(3):599–616.
- Ohnishi, Y. and Li, F. (2025). A bayesian nonparametric approach to mediation and spillover effects with multiple mediators in cluster-randomized trials. *Journal of the American Statistical Association*, (just-accepted):1–20.
- Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American statistical Association*, 82(398):387–394.

- Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, pages 1151–1172.
- Spanbauer, C. and Sparapani, R. (2020). *mxBART: Bayesian Additive Regression Trees with Mixed Effects*. R package version 1.1.
- Spanbauer, C. and Sparapani, R. (2021). Nonparametric machine learning for precision medicine with longitudinal clinical trials and bayesian additive regression trees with mixed models. *Statistics in Medicine*, 40(11):2665–2691.
- Spiegelhalter, D. J. (2001). Bayesian methods for cluster randomized trials with continuous responses. *Statistics in Medicine*, 20(3):435–452.
- Spiegelhalter, D. J., Freedman, L. S., and Parmar, M. K. (1994). Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 157(3):357–387.
- Su, F. and Ding, P. (2021). Model-assisted analyses of cluster-randomized experiments. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):994–1015.
- Thompson, S. G., Warn, D. E., and Turner, R. M. (2004). Bayesian methods for analysis of binary outcome data in cluster randomized trials on the absolute risk scale. *Statistics in Medicine*, 23(3):389–410.
- Tong, G., Tong, J., Jiang, Y., Esserman, D., Harhay, M. O., and Warren, J. L. (2024). Hierarchical bayesian modeling of heterogeneous outcome variance in cluster randomized trials. *Clinical Trials*, 21(4):451–460.
- Turner, R. M., Omar, R. Z., and Thompson, S. G. (2001). Bayesian methods of analysis for cluster randomized trials with binary outcome data. *Statistics in Medicine*, 20(3):453–472.
- van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes. With Applications to Statistics*. New York: Springer.
- Van der Vaart, A. W. (2000). *Asymptotic Statistics*, volume 3. Cambridge university press.
- Wang, B., Harhay, M. O., Tong, J., Small, D. S., Morris, T. P., and Li, F. (2025). On the mixed-model analysis of covariance in cluster-randomized trials. *Statistical Science*.
- Wang, B., Li, F., and Wang, R. (2024a). Handling incomplete outcomes and covariates in cluster-randomized trials: doubly-robust estimation, efficiency considerations, and sensitivity analysis. *arXiv preprint arXiv:2401.11278*.
- Wang, B., Park, C., Small, D. S., and Li, F. (2024b). Model-robust and efficient covariate adjustment for cluster-randomized experiments. *Journal of the American Statistical Association*, pages 1–13.
- Weinfurt, K. P., Hernandez, A. F., Coronado, G. D., DeBar, L. L., Dember, L. M., Green, B. B., Heagerty, P. J., Huang, S. S., James, K. T., Jarvik, J. G., et al. (2017). Pragmatic clinical trials embedded in healthcare systems: generalizable lessons from the nih collaboratory. *BMC Medical Research Methodology*, 17:1–10.
- Yang, S., Li, F., Starks, M. A., Hernandez, A. F., Mentz, R. J., and Choudhury, K. R. (2020). Sample size requirements for detecting treatment effect heterogeneity in cluster randomized trials. *Statistics in Medicine*, 39(28):4218–4237.
- Yiu, A., Fong, E., Holmes, C., and Rousseau, J. (2025). Semiparametric posterior corrections. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkaf005.
- Zhu, A. Y., Mitra, N., Hemming, K., Harhay, M. O., and Li, F. (2024). Leveraging baseline covariates to analyze small cluster-randomized trials with a rare binary outcome. *Biometrical Journal*, 66(1):2200135.

## Appendix 1. Consistency of the calibrated variance estimator

We follow the decomposition and approach of [Antonelli et al. \(2022\)](#) and adapt the proof for CRT settings. Throughout, we assume the treatment assignment model is known due to randomization. Under a more general representation, the model-robust estimator for  $\Delta_\omega$ ,  $\omega \in \{C, I\}$ , takes the form of

$$\Delta_\omega^{\text{mr}} = f(\mu_\omega^{\text{mr}}(1), \mu_\omega^{\text{mr}}(0)),$$

with

$$\hat{\mu}_\omega^{\text{mr}}(a) = \left( \sum_{i=1}^M \frac{\omega_i}{\bar{\omega}} \right)^{-1} \sum_{i=1}^M \frac{\omega_i}{\bar{\omega}} \left\{ \frac{\mathbb{I}(A_i = a) (\bar{Y}_i - \hat{\mathbb{E}}(\bar{Y}_i | A_i = a, \mathbf{X}_i, N_i))}{\pi^a (1 - \pi)^{1-a}} + \hat{\mathbb{E}}(\bar{Y}_i | A_i = a, \mathbf{X}_i, N_i) \right\},$$

where  $\bar{\omega} = M^{-1} \sum_{i=1}^M \omega_i$  is the average weight across clusters,  $M$  is the number of clusters,  $A_i$  is the treatment assignment for cluster  $i$ ,  $\bar{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} Y_{ij}$  is the observed mean outcome in cluster  $i$ ,  $N_i$  denotes the size of cluster  $i$ , and  $\mathbf{X}_i$  represents the collection of all covariates within cluster  $i$ . Specifically, setting  $\omega_i = 1$  yields the cluster-ATE, while  $\omega_i = N_i$  corresponds to the individual-ATE. We will focus on the proof for  $\mu_\omega(a)$  and it is trivial to then extend to  $\Delta_\omega$ . Then,

$$\hat{\mu}_\omega^{\text{mr}}(a) - \mu_\omega(a) = \left( \sum_{i=1}^M \omega_i \right)^{-1} \sum_{i=1}^M \omega_i \left[ \frac{\mathbb{I}(A_i = a)}{\pi_a} (\bar{Y}_i - \hat{m}_{ai}) + \hat{m}_{ai} - \mu_\omega(a) \right],$$

where  $\pi_a \equiv \mathbb{P}(A_i = a) = \pi^a (1 - \pi)^{1-a}$  for simplicity,  $\hat{m}_{ai} = \hat{\mathbb{E}}(\bar{Y}_i | A_i = a, \mathbf{X}_i, N_i)$  is the estimated conditional mean (from the outcome model), and  $m_{ai}$  is the true conditional mean. We assume a parametric working model for the outcome regression and apply Taylor series expansions for  $\hat{m}_{ai}$  in the proof.

We use the following decomposition

$$\hat{\mu}_\omega^{\text{mr}}(a) - \mu_\omega(a) = A + B$$

where

$$\begin{aligned} A &= \left( \sum_{i=1}^M \omega_i \right)^{-1} \sum_{i=1}^M \omega_i (\hat{m}_{ai} - m_{ai}) \left( 1 - \frac{\mathbb{I}(A_i = a)}{\pi_a} \right), \\ B &= \left( \sum_{i=1}^M \omega_i \right)^{-1} \sum_{i=1}^M \omega_i \left[ \frac{\mathbb{I}(A_i = a)}{\pi_a} (\bar{Y}_i - m_{ai}) + m_{ai} - \mu_\omega(a) \right]. \end{aligned}$$

Let  $\Delta(\mathbf{D}, \boldsymbol{\beta})$  denote the doubly robust estimator as a function of data and model parameters. The variance estimator is

$$\hat{V} = \text{Var}_{\mathbf{D}^{(k)}} \left\{ \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} [\Delta(\mathbf{D}^{(k)}, \boldsymbol{\beta})] \right\} + \text{Var}_{\boldsymbol{\beta}|\mathbf{D}} [\Delta(\mathbf{D}, \boldsymbol{\beta})],$$

and our target is

$$V = \text{Var}_{\mathbf{D}} \left\{ \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} [\Delta(\mathbf{D}, \boldsymbol{\beta})] \right\}.$$

Our goal is to show that  $\hat{V} - V = o_p(M^{-1})$  in a CRT, the outcome model contracts at a sufficiently fast rate, and the cluster weights are bounded away from zero and infinity so that no individual cluster receives vanishingly small or disproportionately large influence. To achieve this, we analyze the asymptotic behavior of two main components separately. First, we show that  $\text{Var}_{\boldsymbol{\beta}|\mathbf{D}} [\Delta(\mathbf{D}, \boldsymbol{\beta})] = o_p(M^{-1})$ . Next, we show that  $\text{Var}_{\mathbf{D}^{(k)}} \left\{ \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} [\Delta(\mathbf{D}^{(k)}, \boldsymbol{\beta})] \right\} = \text{Var}_{\mathbf{D}}(B) + o_p(M^{-1})$ , and that

$\text{Var}_{\mathbf{D}} \{ \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} [\Delta(\mathbf{D}, \boldsymbol{\beta})] \} = \text{Var}_{\mathbf{D}}(B) + o_p(M^{-1})$ . Together, these results imply that the proposed variance estimator is consistent.

**Assumptions** Let  $P_0 \in \mathcal{P}$  denote the true (unknown) probability distribution that generates the observed data. In our setting, this corresponds to the distribution over all cluster-level data tuples:  $(A_i, \mathbf{X}_i, N_i, \mathbf{Y}_i)$ , where  $A_i$  is the treatment assignment,  $\mathbf{X}_i$  denotes the matrix of covariates for cluster  $i$ ,  $N_i$  is the cluster size, and  $\mathbf{Y}_i$  is the vector of observed outcomes for individuals in cluster  $i$ . The collection of random variables in cluster  $i$  is  $\tilde{\mathbf{O}}_i^{full} = \{\mathbf{Y}_i(1), \mathbf{Y}_i(0), N_i, \mathbf{X}_i\}$ . The complete data  $\{(\tilde{\mathbf{O}}_1^{full}, A_1), \dots, (\tilde{\mathbf{O}}_M^{full}, A_M)\}$  is only partially observed.

**Assumption 1 (Super-population):**

- (a) Random variables  $\tilde{\mathbf{O}}_1^{full}, \dots, \tilde{\mathbf{O}}_M^{full}$  are mutually independent.
- (b) The total number of individuals in cluster  $i$ ,  $N_i$ , follows an unknown distribution  $\mathcal{P}^N$  over a finite support on  $\mathbb{N}^+$ .
- (c) Given  $N_i$ ,  $\tilde{\mathbf{O}}_i^{full}$  follows an unknown distribution  $\mathcal{P}^{\tilde{\mathbf{O}}|N}$  with finite second moments.
- (d)  $\sup_{P_0} \mathbb{E}_{P_0} \left[ (\bar{Y}_i - m_{ai})^2 \right] \leq K_y < \infty$ .

According to this Assumption, the random variable vectors from different clusters are independent of each other, but within the same cluster, there can be correlation among outcomes, covariates and cluster size. Meanwhile, the collection of random variables in each cluster,  $\{\tilde{\mathbf{O}}_1^{full}, \dots, \tilde{\mathbf{O}}_M^{full}\}$ , are marginally identically distributed according to  $\mathcal{P}^N \times \mathcal{P}^{\tilde{\mathbf{O}}|N}$ . This technical condition is useful for managing the differences in the dimensions of  $\tilde{\mathbf{O}}_i^{full}$  among various clusters.

**Assumption 2 (Bounds on the error of posterior distributions):**  $\sup_{P_0} \mathbb{E}_{P_0} \text{Var}_M (\hat{m}_{ai} - m_{ai} | \mathbf{D}_i) \leq K_m < \infty$ .

It states that the posterior distribution of the conditional mean of the cluster-level outcome has bounded variance around its true value. This assumption guarantees that the posterior does not assign excessive mass far from the truth and is typically mild in applications, especially when the outcome is categorical or bounded.

**Assumption 3 (Posterior contraction of the outcome model):** There exists a sequence  $\epsilon_M \rightarrow 0$  and a constant  $Q > 0$ , independent of  $\epsilon_M$ , such that

$$\sup_{P_0} \mathbb{E}_{P_0} \mathbb{P}_M \left( \frac{1}{\sqrt{M}} \|\hat{\mathbf{m}}_a - \mathbf{m}_a\| > Q\epsilon_M \mid \mathbf{D} \right) \rightarrow 0,$$

where  $\|\cdot\|$  denotes the Euclidean norm taken over the  $M$  clusters. This assumption states that the posterior distribution of the outcome model contracts to the true conditional mean function at rate  $\epsilon_M$ . We will use a Taylor series expansion for  $\hat{m}_{ai}$  of the following form

$$\hat{m}_{ai} = m_{ai} + d_1(\mathbf{X}_i, N_i, \boldsymbol{\beta}^*)^\top (\boldsymbol{\beta} - \boldsymbol{\beta}^*) + (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top H_1(\mathbf{X}_i, N_i, \tilde{\boldsymbol{\beta}})(\boldsymbol{\beta} - \boldsymbol{\beta}^*),$$

where  $\boldsymbol{\beta}^*$  is the true value of  $\boldsymbol{\beta}$ ,  $\tilde{\boldsymbol{\beta}}$  is a point between  $\boldsymbol{\beta}$  and  $\boldsymbol{\beta}^*$ , and  $m(\cdot)$  has bounded first and second derivatives.

**Assumption 4 (Moments of the posterior distribution):**

- (a)  $\mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} [\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^4] = o_p(M^{-1})$
- (b)  $\mathbb{E}_{\mathbf{D}} [\mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} [\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^8]]^{1/2} = o_p(M^{-1})$

These assumptions assume the convergence rates of the outcome regression models are slightly faster than  $M^{-1/4}$ .

**Assumption 5 (Regularity conditions on cluster weights).** There exist constants  $0 < b_\omega < B_\omega < \infty$ , independent of  $M$ , such that for all clusters  $i = 1, \dots, M$ ,

$$b_\omega \leq \omega_i \leq B_\omega.$$

This ensures that no individual cluster receives vanishingly small or disproportionately large weight.

**Asymptotic Behavior of  $\text{Var}_{\beta|D}[\Delta(D, \beta)]$**  For the CRT setting with known treatment mechanism, we have the decomposition

$$\hat{\mu}_\omega^{\text{mr}}(a) - \mu_\omega(a) = A + B.$$

We will analyze  $\text{Var}_{\beta|D}[A]$  and  $\text{Var}_{\beta|D}[B]$ . If each of these two components is  $o_p(M^{-1})$ , then the entire variance is of the same order, since the sum of variances and covariances remains within this rate. First, note that  $B$  contains no unknown parameters (i.e., it depends only on observed outcomes and the true conditional mean  $m_{ai}$ ), so  $\text{Var}_{\beta|D}[B] = 0$ . Next, we analyze  $\text{Var}_{\beta|D}[A]$ , which takes the form

$$\text{Var}_{\beta|D}[A] = \text{Var}_{\beta|D} \left[ \left( \sum_{i=1}^M \omega_i \right)^{-1} \sum_{i=1}^M \omega_i \left( 1 - \frac{\mathbb{I}(A_i = a)}{\pi_a} \right) (\hat{m}_{ai} - m_{ai}) \right].$$

Using the Taylor expansion of  $\hat{m}_{ai}$  around  $\beta^*$  (from Assumption 3), we get:

$$\hat{m}_{ai} - m_{ai} = d_1(\mathbf{X}_i, N_i, \beta^*)^\top (\beta - \beta^*) + (\beta - \beta^*)^\top H_1(\mathbf{X}_i, N_i, \tilde{\beta})(\beta - \beta^*).$$

Substituting this into the variance and using the arithmetic mean and geometric mean inequality :

$$\begin{aligned} \text{Var}_{\beta|D}[A] &= \text{Var}_{\beta|D} \left[ \left( \sum_{i=1}^M \omega_i \right)^{-1} \sum_{i=1}^M \omega_i \left( 1 - \frac{\mathbb{I}(A_i = a)}{\pi_a} \right) \times \left( d_1(\mathbf{X}_i, N_i, \beta^*)^\top (\beta - \beta^*) + (\beta - \beta^*)^\top H_1(\mathbf{X}_i, N_i, \tilde{\beta})(\beta - \beta^*) \right) \right] \\ &\leq 2 \text{Var}_{\beta|D} \left[ \left( \sum_{i=1}^M \omega_i \right)^{-1} \sum_{i=1}^M \omega_i \left( 1 - \frac{\mathbb{I}(A_i = a)}{\pi_a} \right) d_1(\mathbf{X}_i, N_i, \beta^*)^\top (\beta - \beta^*) \right] \\ &\quad + 2 \text{Var}_{\beta|D} \left[ \left( \sum_{i=1}^M \omega_i \right)^{-1} \sum_{i=1}^M \omega_i \left( 1 - \frac{\mathbb{I}(A_i = a)}{\pi_a} \right) (\beta - \beta^*)^\top H_1(\mathbf{X}_i, N_i, \tilde{\beta})(\beta - \beta^*) \right]. \end{aligned}$$

We analyze the first variance term,  $\text{Var}_{\beta|D} \left[ \left( \sum_{i=1}^M \omega_i \right)^{-1} \sum_{i=1}^M \omega_i \left( 1 - \frac{\mathbb{I}(A_i = a)}{\pi_a} \right) d_1(\mathbf{X}_i, N_i, \beta^*)^\top (\beta - \beta^*) \right]$ . By Assumption 5, the cluster weights satisfy  $b_\omega \leq \omega_i \leq B_\omega$  for all  $i$ , where  $0 < b_\omega < B_\omega < \infty$  are constants independent of  $M$ . Therefore,

$$b_\omega M \leq \sum_{i=1}^M \omega_i \leq B_\omega M, \quad \text{which implies} \quad \frac{1}{(B_\omega M)^2} \leq \left( \sum_{i=1}^M \omega_i \right)^{-2} \leq \frac{1}{(b_\omega M)^2}.$$

Conditioning on the observed data  $D$ , the terms  $\left( 1 - \frac{\mathbb{I}(A_i = a)}{\pi_a} \right)$  and  $d_1(\mathbf{X}_i, N_i, \beta^*)$  are nonrandom. Applying Assumption 4, we can factor out  $\text{Var}_{\beta|D}(\beta - \beta^*)$  and bound the variance by

$$\frac{\lambda}{(b_\omega M)^2} \cdot \left\| \sum_{i=1}^M \omega_i \left( 1 - \frac{\mathbb{I}(A_i = a)}{\pi_a} \right) d_1(\mathbf{X}_i, N_i, \beta^*) \right\|^2,$$

where  $\lambda$  denotes the largest eigenvalue of  $\text{Var}_{\beta|D}(\beta - \beta^*)$ . Since the weighted sum inside the norm is  $O_p(\sqrt{M})$  by the multivariate central limit theorem, the entire expression is

$$\text{Var}_{\beta|D} \left[ \left( \sum_{i=1}^M \omega_i \right)^{-1} \sum_{i=1}^M \omega_i \left( 1 - \frac{\mathbb{I}(A_i = a)}{\pi_a} \right) d_1(\mathbf{X}_i, N_i, \beta^*)^\top (\beta - \beta^*) \right] = o_p(M^{-1}).$$

We now handle the second term.

$$\begin{aligned} & \text{Var}_{\beta|D} \left[ \left( \sum_{i=1}^M \omega_i \right)^{-1} \sum_{i=1}^M \omega_i \left( 1 - \frac{\mathbb{I}(A_i = a)}{\pi_a} \right) (\beta - \beta^*)^\top H_1(\mathbf{X}_i, N_i, \tilde{\beta})(\beta - \beta^*) \right] \\ & \leq \mathbb{E}_{\beta|D} \left[ \left( \left( \sum_{i=1}^M \omega_i \right)^{-1} \sum_{i=1}^M \omega_i \left( 1 - \frac{\mathbb{I}(A_i = a)}{\pi_a} \right) (\beta - \beta^*)^\top H_1(\mathbf{X}_i, N_i, \tilde{\beta})(\beta - \beta^*) \right)^2 \right]. \end{aligned}$$

Since  $H_1(\cdot)$  has bounded operator norm and  $\left(1 - \frac{\mathbb{I}(A_i = a)}{\pi_a}\right)$  is bounded by a positive constant, the term is bounded by a constant times  $\|\beta - \beta^*\|_2^2$ . Therefore,  $\left((\beta - \beta^*)^\top H_1(\mathbf{X}_i, N_i, \tilde{\beta})(\beta - \beta^*)\right)^2 \leq C \|\beta - \beta^*\|_2^4$  for some constant  $C > 0$ . Substituting and applying Cauchy-Schwarz inequality,

$$\begin{aligned} & \mathbb{E}_{\beta|D} \left[ \left( \left( \sum_{i=1}^M \omega_i \right)^{-1} \sum_{i=1}^M \omega_i \left( 1 - \frac{\mathbb{I}(A_i = a)}{\pi_a} \right) (\beta - \beta^*)^\top H_1(\mathbf{X}_i, N_i, \tilde{\beta})(\beta - \beta^*) \right)^2 \right] \\ & \leq \frac{1}{(b_\omega M)^2} \mathbb{E}_{\beta|D} \left[ \left( \sum_{i=1}^M \omega_i^2 \left( 1 - \frac{\mathbb{I}(A_i = a)}{\pi_a} \right)^2 \right) \sum_{i=1}^M \left( (\beta - \beta^*)^\top H_1(\mathbf{X}_i, N_i, \tilde{\beta})(\beta - \beta^*) \right)^2 \right] \\ & \leq \frac{C}{(b_\omega M)^2} \left( \sum_{i=1}^M \omega_i^2 \left( 1 - \frac{\mathbb{I}(A_i = a)}{\pi_a} \right)^2 \right) \sum_{i=1}^M \mathbb{E}_{\beta|D} [\|\beta - \beta^*\|_2^4] \\ & \leq \frac{C'}{M} \sum_{i=1}^M \mathbb{E}_{\beta|D} [\|\beta - \beta^*\|_2^4] \\ & = C' \mathbb{E}_{\beta|D} [\|\beta - \beta^*\|_2^4]. \end{aligned}$$

By Assumption 4,  $\mathbb{E}_{\beta|D} [\|\beta - \beta^*\|_2^4] = o_p(M^{-1})$ , so the entire expression is  $o_p(M^{-1})$ . Therefore,

$$\text{Var}_{\beta|D} \left[ \left( \sum_{i=1}^M \omega_i \right)^{-1} \sum_{i=1}^M \omega_i \left( 1 - \frac{\mathbb{I}(A_i = a)}{\pi_a} \right) (\beta - \beta^*)^\top H_1(\mathbf{X}_i, N_i, \tilde{\beta})(\beta - \beta^*) \right] = o_p(M^{-1}).$$

Putting these two terms together shows each variance term is  $o_p(M^{-1})$ , so  $\text{Var}_{\beta|D}[A] = o_p(M^{-1})$ . Since  $\text{Var}_{\beta|D}[B] = 0$ , we conclude that

$$\text{Var}_{\beta|D}[\Delta(D, \beta)] = o_p(M^{-1}).$$

**Asymptotic Behavior of  $\text{Var}_{D^{(k)}} \left\{ \mathbb{E}_{\beta|D}[\Delta(D^{(k)}, \beta)] \right\}$**  Recall that we have

$$\hat{\mu}_\omega^{\text{mr}}(a) - \mu_\omega(a) = A + B.$$

Our aim is to show that

$$\text{Var}_{\mathbf{D}^{(k)}} \{ \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}}[B] \} \approx \text{Var}_{\mathbf{D}}[B], \quad \text{and} \quad \text{Var}_{\mathbf{D}^{(k)}} \{ \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}}[A] \} = o_p(M^{-1}).$$

Together, these can imply that the total variance is  $o_p(M^{-1})$ . For  $\text{Var}_{\mathbf{D}^{(k)}} \{ \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}}[B] \}$ , we have

$$\text{Var}_{\mathbf{D}^{(k)}} \{ \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}}[B] \} = \text{Var}_{\mathbf{D}^{(k)}}[B] = \text{Var}_{\mathbf{D}^{(k)}} \left[ \left( \sum_{i=1}^M \omega_i^{(k)} \right)^{-1} \sum_{i=1}^M \omega_i^{(k)} \left[ \frac{\mathbb{I}(A_i^{(k)} = a)}{\pi_a} (\bar{Y}_i^{(k)} - m_{ai}^{(k)}) + m_{ai}^{(k)} - \mu_\omega(a) \right] \right],$$

where  $(\omega_i^{(k)}, A_i^{(k)}, \bar{Y}_i^{(k)}, m_{ai}^{(k)})$  come from the cluster-level bootstrap sample  $\mathbf{D}^{(k)}$ . As  $M \rightarrow \infty$ , resampling from  $\mathbf{D}^{(k)}$  converges to sampling from  $\mathbf{D}$ , so

$$\text{Var}_{\mathbf{D}^{(k)}}[B] \approx \text{Var}_{\mathbf{D}} \left[ \left( \sum_{i=1}^M \omega_i \right)^{-1} \sum_{i=1}^M \omega_i \left[ \frac{\mathbb{I}(A_i = a)}{\pi_a} (\bar{Y}_i - m_{ai}) + m_{ai} - \mu_\omega(a) \right] \right] = \text{Var}_{\mathbf{D}}[B].$$

Hence,

$$\text{Var}_{\mathbf{D}^{(k)}} \{ \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}}[B] \} \approx \text{Var}_{\mathbf{D}}[B].$$

This approximation follows from the cluster bootstrap's consistency. For

$$\text{Var}_{\mathbf{D}^{(k)}} \{ \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}}[A] \} = \text{Var}_{\mathbf{D}^{(k)}} \left\{ \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} \left[ \left( \sum_{i=1}^M \omega_i^{(k)} \right)^{-1} \sum_{i=1}^M \omega_i^{(k)} \left( 1 - \frac{\mathbb{I}(A_i^{(k)} = a)}{\pi_a} \right) (\hat{m}_{ai}^{(k)} - m_{ai}^{(k)}) \right] \right\},$$

using a Taylor expansion of  $\hat{m}_{ai}^{(k)}$  around  $m_{ai}^{(k)}$ , together with Assumptions 3 (posterior contraction) and 4 (bounded posterior moments), we can obtain that  $\text{Var}_{\mathbf{D}^{(k)}} \{ \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}}[A] \} = o_p(M^{-1})$  as demonstrated in the steps below.

$$\begin{aligned} \text{Var}_{\mathbf{D}^{(k)}} \{ \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}}[A] \} &= \text{Var}_{\mathbf{D}^{(k)}} \left\{ \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} \left[ \left( \sum_{i=1}^M \omega_i^{(k)} \right)^{-1} \sum_{i=1}^M \omega_i^{(k)} \left( 1 - \frac{\mathbb{I}(A_i^{(k)} = a)}{\pi_a} \right) (\hat{m}_{ai}^{(k)} - m_{ai}^{(k)}) \right] \right\} \\ &= \text{Var}_{\mathbf{D}^{(k)}} \left\{ \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} \left[ \left( \sum_{i=1}^M \omega_i^{(k)} \right)^{-1} \sum_{i=1}^M \omega_i^{(k)} \left( 1 - \frac{\mathbb{I}(A_i^{(k)} = a)}{\pi_a} \right) \right. \right. \\ &\quad \times \left. \left. \left( d_1(\mathbf{X}_i^{(k)}, N_i^{(k)}, \boldsymbol{\beta}^*)^\top (\boldsymbol{\beta} - \boldsymbol{\beta}^*) + (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top H_1(\mathbf{X}_i^{(k)}, N_i^{(k)}, \tilde{\boldsymbol{\beta}})(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \right) \right] \right\} \\ &\leq 2 \text{Var}_{\mathbf{D}^{(k)}} \left\{ \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} \left[ \left( \sum_{i=1}^M \omega_i^{(k)} \right)^{-1} \sum_{i=1}^M \omega_i^{(k)} \left( 1 - \frac{\mathbb{I}(A_i^{(k)} = a)}{\pi_a} \right) \times d_1(\mathbf{X}_i^{(k)}, N_i^{(k)}, \boldsymbol{\beta}^*)^\top (\boldsymbol{\beta} - \boldsymbol{\beta}^*) \right] \right\} \\ &\quad + 2 \text{Var}_{\mathbf{D}^{(k)}} \left\{ \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} \left[ \left( \sum_{i=1}^M \omega_i^{(k)} \right)^{-1} \sum_{i=1}^M \omega_i^{(k)} \left( 1 - \frac{\mathbb{I}(A_i^{(k)} = a)}{\pi_a} \right) \times (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top H_1(\mathbf{X}_i^{(k)}, N_i^{(k)}, \tilde{\boldsymbol{\beta}})(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \right] \right\}. \end{aligned}$$

Then,

$$\begin{aligned}
& \text{Var}_{\mathbf{D}^{(k)}} \left\{ \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} \left[ \left( \sum_{i=1}^M \omega_i^{(k)} \right)^{-1} \sum_{i=1}^M \omega_i^{(k)} \left( 1 - \frac{\mathbb{I}(A_i^{(k)} = a)}{\pi_a} \right) \times d_1(\mathbf{X}_i^{(k)}, N_i^{(k)}, \boldsymbol{\beta}^*)^\top (\boldsymbol{\beta} - \boldsymbol{\beta}^*) \right] \right\} \\
&= \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} [(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top] \text{Var}_{\mathbf{D}^{(k)}} \left\{ \left( \sum_{i=1}^M \omega_i^{(k)} \right)^{-1} \sum_{i=1}^M \omega_i^{(k)} \left( 1 - \frac{\mathbb{I}(A_i^{(k)} = a)}{\pi_a} \right) \times d_1(\mathbf{X}_i^{(k)}, N_i^{(k)}, \boldsymbol{\beta}^*)^\top \right\} \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} [\boldsymbol{\beta} - \boldsymbol{\beta}^*] \\
&= \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} [(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top] \left\{ \mathbb{E}_{\mathbf{D}^{(k)}} \left[ \left( \sum_{i=1}^M \omega_i^{(k)} \right)^{-2} \right] \text{Var}_{\mathbf{D}^{(k)}} \left\{ \sum_{i=1}^M \omega_i^{(k)} \left( 1 - \frac{\mathbb{I}(A_i^{(k)} = a)}{\pi_a} \right) d_1(\mathbf{X}_i^{(k)}, N_i^{(k)}, \boldsymbol{\beta}^*)^\top \right\} \right. \\
&\quad \left. + \text{Var}_{\mathbf{D}^{(k)}} \left[ \left( \sum_{i=1}^M \omega_i^{(k)} \right)^{-1} \right] \left\| \mathbb{E}_{\mathbf{D}^{(k)}} \left\{ \sum_{i=1}^M \omega_i^{(k)} \left( 1 - \frac{\mathbb{I}(A_i^{(k)} = a)}{\pi_a} \right) d_1(\mathbf{X}_i^{(k)}, N_i^{(k)}, \boldsymbol{\beta}^*)^\top \right\} \right\|^2 \right\} \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} [\boldsymbol{\beta} - \boldsymbol{\beta}^*] \\
&= M \cdot \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} [(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top] \mathbb{E}_{\mathbf{D}^{(k)}} \left[ \left( \sum_{i=1}^M \omega_i^{(k)} \right)^{-2} \right] \underbrace{\text{Var}_{\mathbf{D}^{(k)}} \left\{ \frac{1}{\sqrt{M}} \sum_{i=1}^M \omega_i^{(k)} \left( 1 - \frac{\mathbb{I}(A_i^{(k)} = a)}{\pi_a} \right) d_1(\mathbf{X}_i^{(k)}, N_i^{(k)}, \boldsymbol{\beta}^*)^\top \right\}}_{\Sigma^{(k)}} \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} [\boldsymbol{\beta} - \boldsymbol{\beta}^*] \\
&\quad + M^2 \cdot \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} [(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top] \text{Var}_{\mathbf{D}^{(k)}} \left[ \left( \sum_{i=1}^M \omega_i^{(k)} \right)^{-1} \right] \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} [\boldsymbol{\beta} - \boldsymbol{\beta}^*] \\
&\quad \cdot \left\| \mathbb{E}_{\mathbf{D}^{(k)}} \left\{ \frac{1}{M} \sum_{i=1}^M \omega_i^{(k)} \left( 1 - \frac{\mathbb{I}(A_i^{(k)} = a)}{\pi_a} \right) d_1(\mathbf{X}_i^{(k)}, N_i^{(k)}, \boldsymbol{\beta}^*)^\top \right\} \right\|^2 \\
&\leq M \cdot \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} [(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top] \frac{1}{(b_\omega M)^2} \lambda^{(k)} \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} [\boldsymbol{\beta} - \boldsymbol{\beta}^*] + M^2 \cdot O(M^{-3}) \cdot O(1) \cdot \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} [(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top] \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} [\boldsymbol{\beta} - \boldsymbol{\beta}^*] \\
&\leq \left( \frac{\lambda^{(k)}}{b_\omega^2 M} + O(M^{-1}) \right) \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} [(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top] \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} [\boldsymbol{\beta} - \boldsymbol{\beta}^*] \\
&= o_p(M^{-1}).
\end{aligned}$$

$\lambda^{(k)}$  is the maximum eigenvalue of  $\Sigma^{(k)}$ . The third line follows from the law of total variance. The fifth line follows from the Delta method, Mean Value Theorem, and the boundedness of the relevant terms. The last line is implied by Assumption 4. Using

the fact that the Hessian matrix is bounded, and Jensen's inequality, we have the following

$$\begin{aligned}
& \text{Var}_{\mathbf{D}^{(k)}} \left\{ \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} \left[ \left( \sum_{i=1}^M \omega_i^{(k)} \right)^{-1} \sum_{i=1}^M \omega_i^{(k)} \left( 1 - \frac{\mathbb{I}(A_i^{(k)} = a)}{\pi_a} \right) (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top H_1(\mathbf{X}_i^{(k)}, N_i^{(k)}, \tilde{\boldsymbol{\beta}})(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \right] \right\} \\
&= \text{Var}_{\mathbf{D}^{(k)}} \left\{ \left( \sum_{i=1}^M \omega_i^{(k)} \right)^{-1} \sum_{i=1}^M \omega_i^{(k)} \left( 1 - \frac{\mathbb{I}(A_i^{(k)} = a)}{\pi_a} \right) \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} \left( (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top H_1(\cdot)(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \right) \right\} \\
&\leq \mathbb{E}_{\mathbf{D}^{(k)}} \left[ \left( \sum_{i=1}^M \omega_i^{(k)} \right)^{-2} \left( \sum_{i=1}^M \omega_i^{(k)} \left( 1 - \frac{\mathbb{I}(A_i^{(k)} = a)}{\pi_a} \right) \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} \left( (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top H_1(\cdot)(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \right) \right)^2 \right] \\
&\leq \frac{1}{(b_\omega M)^2} \mathbb{E}_{\mathbf{D}^{(k)}} \left[ \left( \sum_{i=1}^M \omega_i^{(k)} \left( 1 - \frac{\mathbb{I}(A_i^{(k)} = a)}{\pi_a} \right) \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} \left( (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top H_1(\cdot)(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \right) \right)^2 \right] \\
&\leq \frac{1}{(b_\omega M)^2} \mathbb{E}_{\mathbf{D}^{(k)}} \left[ \sum_{i=1}^M \left( \omega_i^{(k)} \right)^2 \left( 1 - \frac{\mathbb{I}(A_i^{(k)} = a)}{\pi_a} \right)^2 \sum_{i=1}^M \left[ \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} \left( (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top H_1(\cdot)(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \right) \right]^2 \right] \\
&\leq \frac{1}{(b_\omega M)^2} \mathbb{E}_{\mathbf{D}^{(k)}} \left[ \sum_{i=1}^M \left( \omega_i^{(k)} \right)^2 \left( 1 - \frac{\mathbb{I}(A_i^{(k)} = a)}{\pi_a} \right)^2 \sum_{i=1}^M \left[ \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} \left( \lambda(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top (\boldsymbol{\beta} - \boldsymbol{\beta}^*) \right) \right]^2 \right] \\
&\leq \frac{C}{M^2} \left( \sum_{i=1}^M \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} [\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2]^2 \right) \mathbb{E}_{\mathbf{D}^{(k)}} \left[ \sum_{i=1}^M \left( \omega_i^{(k)} \right)^2 \left( 1 - \frac{\mathbb{I}(A_i^{(k)} = a)}{\pi_a} \right)^2 \right] \\
&= C' \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} [\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2]^2 \\
&= o_p(M^{-1}).
\end{aligned}$$

$\lambda$  denotes the largest eigenvalue of the Hessian matrix, and  $C$  and  $C'$  are positive constants. Combining this result with our preceding bounds yields

$$\text{Var}_{\mathbf{D}^{(k)}} \left\{ \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} [\Delta(\mathbf{D}^{(k)}, \boldsymbol{\beta})] \right\} = \text{Var}_{\mathbf{D}} [B] + o_p(M^{-1}).$$

**Asymptotic behavior of  $\text{Var}_{\mathbf{D}} \{ \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} [\Delta(\mathbf{D}, \boldsymbol{\beta})] \}$**  Reuse the expansion, we have

$$\begin{aligned}
& \text{Var}_{\mathbf{D}} \{ \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} [A] \} \\
&= \text{Var}_{\mathbf{D}} \left\{ \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} \left[ \left( \sum_{i=1}^M \omega_i \right)^{-1} \sum_{i=1}^M \omega_i \left( 1 - \frac{\mathbb{I}(A_i = a)}{\pi_a} \right) (\hat{m}_{ai} - m_{ai}) \right] \right\} \\
&= \text{Var}_{\mathbf{D}} \left\{ \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} \left[ \left( \sum_{i=1}^M \omega_i \right)^{-1} \sum_{i=1}^M \omega_i \left( 1 - \frac{\mathbb{I}(A_i = a)}{\pi_a} \right) \left( d_1(\mathbf{X}_i, N_i, \boldsymbol{\beta}^*)^\top (\boldsymbol{\beta} - \boldsymbol{\beta}^*) + (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top H_1(\mathbf{X}_i, N_i, \tilde{\boldsymbol{\beta}})(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \right) \right] \right\} \\
&\leq 2 \text{Var}_{\mathbf{D}} \left\{ \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} \left[ \left( \sum_{i=1}^M \omega_i \right)^{-1} \sum_{i=1}^M \omega_i \left( 1 - \frac{\mathbb{I}(A_i = a)}{\pi_a} \right) d_1(\mathbf{X}_i, N_i, \boldsymbol{\beta}^*)^\top (\boldsymbol{\beta} - \boldsymbol{\beta}^*) \right] \right\} \\
&\quad + 2 \text{Var}_{\mathbf{D}} \left\{ \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} \left[ \left( \sum_{i=1}^M \omega_i \right)^{-1} \sum_{i=1}^M \omega_i \left( 1 - \frac{\mathbb{I}(A_i = a)}{\pi_a} \right) (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top H_1(\mathbf{X}_i, N_i, \tilde{\boldsymbol{\beta}})(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \right] \right\}
\end{aligned}$$

We now handle the two variance terms in the upper bound of  $\text{Var}_{\mathbf{D}} \{ \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}}[A] \}$  separately. For the first term, we have

$$\begin{aligned}
& \text{Var}_{\mathbf{D}} \left\{ \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} \left[ \left( \sum_{i=1}^M \omega_i \right)^{-1} \sum_{i=1}^M \omega_i \left( 1 - \frac{\mathbb{I}(A_i = a)}{\pi_a} \right) d_1(\mathbf{X}_i, N_i, \boldsymbol{\beta}^*)^\top (\boldsymbol{\beta} - \boldsymbol{\beta}^*) \right] \right\} \\
& \leq \mathbb{E}_{\mathbf{D}} \left\{ \left\| \left( \sum_{i=1}^M \omega_i \right)^{-1} \sum_{i=1}^M \omega_i \left( 1 - \frac{\mathbb{I}(A_i = a)}{\pi_a} \right) d_1(\mathbf{X}_i, N_i, \boldsymbol{\beta}^*)^\top \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \right\|^2 \right\} \\
& \leq \frac{1}{(b_\omega M)^2} \mathbb{E}_{\mathbf{D}} \left\{ \left\| \sum_{i=1}^M \omega_i \left( 1 - \frac{\mathbb{I}(A_i = a)}{\pi_a} \right) d_1(\mathbf{X}_i, N_i, \boldsymbol{\beta}^*)^\top \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \right\|^2 \right\} \\
& = \frac{1}{(b_\omega M)^2} \mathbb{E}_{\mathbf{D}} \left\{ \left( \sum_{i=1}^M \omega_i \left( 1 - \frac{\mathbb{I}(A_i = a)}{\pi_a} \right) d_1(\mathbf{X}_i, N_i, \boldsymbol{\beta}^*)^\top \right) \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \right. \\
& \quad \times \left. \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top \left( \sum_{i=1}^M \omega_i \left( 1 - \frac{\mathbb{I}(A_i = a)}{\pi_a} \right) d_1(\mathbf{X}_i, N_i, \boldsymbol{\beta}^*) \right) \right\} \\
& \leq \frac{\lambda}{b_\omega^2 M} \mathbb{E}_{\mathbf{D}} \left\{ \left( \frac{1}{\sqrt{M}} \sum_{i=1}^M \omega_i \left( 1 - \frac{\mathbb{I}(A_i = a)}{\pi_a} \right) d_1(\mathbf{X}_i, N_i, \boldsymbol{\beta}^*)^\top \right) \left( \frac{1}{\sqrt{M}} \sum_{i=1}^M \omega_i \left( 1 - \frac{\mathbb{I}(A_i = a)}{\pi_a} \right) d_1(\mathbf{X}_i, N_i, \boldsymbol{\beta}^*) \right) \right\} \\
& = \frac{\lambda}{b_\omega^2 M} \mathbb{E}_{\mathbf{D}} \left\| \frac{1}{\sqrt{M}} \sum_{i=1}^M \omega_i \left( 1 - \frac{\mathbb{I}(A_i = a)}{\pi_a} \right) d_1(\mathbf{X}_i, N_i, \boldsymbol{\beta}^*)^\top \right\|^2 \\
& \leq \frac{\lambda}{b_\omega^2 M} o_p(1) \\
& = o_p(M^{-1}),
\end{aligned}$$

where  $\lambda$  denotes the maximum eigenvalue of  $\mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top$ . The final result follows by applying Assumption 4 and the multivariate Central Limit Theorem. Now we focus on the second of the two terms.

$$\begin{aligned}
& \text{Var}_{\mathbf{D}} \left\{ \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} \left[ \left( \sum_{i=1}^M \omega_i \right)^{-1} \sum_{i=1}^M \omega_i \left( 1 - \frac{\mathbb{I}(A_i = a)}{\pi_a} \right) (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top H_1(\mathbf{X}_i, N_i, \tilde{\boldsymbol{\beta}})(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \right] \right\} \\
&= \text{Var}_{\mathbf{D}} \left\{ \left( \sum_{i=1}^M \omega_i \right)^{-1} \sum_{i=1}^M \omega_i \left( 1 - \frac{\mathbb{I}(A_i = a)}{\pi_a} \right) \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} \left( (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top H_1(\mathbf{X}_i, N_i, \tilde{\boldsymbol{\beta}})(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \right) \right\} \\
&\leq \mathbb{E}_{\mathbf{D}} \left\{ \left[ \left( \sum_{i=1}^M \omega_i \right)^{-1} \sum_{i=1}^M \omega_i \left( 1 - \frac{\mathbb{I}(A_i = a)}{\pi_a} \right) \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} \left( (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top H_1(\mathbf{X}_i, N_i, \tilde{\boldsymbol{\beta}})(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \right) \right]^2 \right\} \\
&\leq \frac{1}{b_\omega^2 M^2} \mathbb{E}_{\mathbf{D}} \left\{ \left( \sum_{i=1}^M \omega_i^2 \left( 1 - \frac{\mathbb{I}(A_i = a)}{\pi_a} \right)^2 \right) \left( \sum_{i=1}^M \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}}^2 \left[ (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top H_1(\mathbf{X}_i, N_i, \tilde{\boldsymbol{\beta}})(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \right] \right) \right\} \\
&\leq \frac{1}{b_\omega^2 M^2} \mathbb{E}_{\mathbf{D}} \left\{ \left[ \sum_{i=1}^M \omega_i^2 \left( 1 - \frac{\mathbb{I}(A_i = a)}{\pi_a} \right)^2 \right]^2 \right\}^{1/2} \mathbb{E}_{\mathbf{D}} \left[ \left( \sum_{i=1}^M \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}}^2 \left[ (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top H_1(\mathbf{X}_i, N_i, \tilde{\boldsymbol{\beta}})(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \right] \right)^2 \right]^{1/2} \\
&\leq \frac{C}{M} \mathbb{E}_{\mathbf{D}} \left[ \left( \sum_{i=1}^M \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}}^2 \left[ (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top H_1(\mathbf{X}_i, N_i, \tilde{\boldsymbol{\beta}})(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \right] \right)^2 \right]^{1/2} \\
&\leq \frac{C}{M} \mathbb{E}_{\mathbf{D}} \left[ \left( \sum_{i=1}^M \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}}^2 [\lambda \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2] \right)^2 \right]^{1/2} \\
&= \frac{C'}{M} \mathbb{E}_{\mathbf{D}} \left[ \left( \sum_{i=1}^M \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}}^2 [\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2] \right)^2 \right]^{1/2} \\
&= C' \mathbb{E}_{\mathbf{D}} \left[ \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}}^4 [\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2] \right]^{1/2} \\
&\leq C' \mathbb{E}_{\mathbf{D}} [\mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}} [\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^8]]^{1/2} \\
&= o_p(M^{-1})
\end{aligned}$$

Therefore, we have established that  $\text{Var}_{\mathbf{D}}\{\mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}}[A]\}$  is  $o_p(M^{-1})$ , while  $\text{Var}_{\mathbf{D}}\{\mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}}[B]\} = \text{Var}_{\mathbf{D}}[B]$ . Therefore,

$$\text{Var}_{\mathbf{D}} \{\mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}}[\Delta(D, \boldsymbol{\beta})]\} = \text{Var}_{\mathbf{D}}[B] + o_p(M^{-1}).$$

Combining this with the results from previous sections, we conclude that

$$\hat{V} - V = \text{Var}_{\mathbf{D}^{(k)}} \left\{ \mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}}[\Delta(\mathbf{D}^{(k)}, \boldsymbol{\beta})] \right\} + \text{Var}_{\boldsymbol{\beta}|\mathbf{D}}[\Delta(D, \boldsymbol{\beta})] - \text{Var}_{\mathbf{D}} \{\mathbb{E}_{\boldsymbol{\beta}|\mathbf{D}}[\Delta(\mathbf{D}, \boldsymbol{\beta})]\} = o_p(M^{-1}),$$

which completes the proof.

## Appendix 2. Consistency of the model-robust point estimator

Utilizing the decomposition and the assumptions listed above in Web Appendix 1, we have  $\hat{\mu}_\omega^{\text{mr}}(a) - \mu_\omega(a) = A + B$ , where

$$\begin{aligned} A &= \left( \sum_{i=1}^M \omega_i \right)^{-1} \sum_{i=1}^M \omega_i (\hat{m}_{ai} - m_{ai}) \left( 1 - \frac{\mathbb{I}(A_i = a)}{\pi_a} \right), \\ B &= \left( \sum_{i=1}^M \omega_i \right)^{-1} \sum_{i=1}^M \omega_i \left[ \frac{\mathbb{I}(A_i = a)}{\pi_a} (\bar{Y}_i - m_{ai}) + m_{ai} - \mu_\omega(a) \right]. \end{aligned}$$

We examine the contraction rates for each component of  $\hat{\mu}_\omega^{\text{mr}}(a) - \mu_\omega(a) = A + B$ , and show that, even under outcome model misspecification, the estimator remains root- $M$  consistent, that is,  $\hat{\mu}_\omega^{\text{mr}}(a) - \mu_\omega(a)$  contracts at rate  $M^{-1/2}$ .

**Contraction rate of component  $B$**  We begin with component  $B$ , which does not depend on the posterior distribution of the outcome model,

$$\begin{aligned} &\sup_{P_0} \mathbb{E}_{P_0} \mathbb{P}_M (|B| > C\epsilon_M \mid \mathbf{D}) \\ &= \sup_{P_0} \mathbb{E}_{P_0} \mathbb{P}_M \left( \left| \left( \sum_{i=1}^M \omega_i \right)^{-1} \sum_{i=1}^M \omega_i \left[ \frac{\mathbb{I}(A_i = a)}{\pi_a} (\bar{Y}_i - m_{ai}) + m_{ai} - \mu_\omega(a) \right] \right| > C\epsilon_M \mid \mathbf{D} \right) \\ &= \sup_{P_0} \mathbb{E}_{P_0} \mathbb{P} \left( \left| \left( \sum_{i=1}^M \omega_i \right)^{-1} \sum_{i=1}^M \omega_i \left[ \frac{\mathbb{I}(A_i = a)}{\pi_a} (\bar{Y}_i - m_{ai}) + m_{ai} - \mu_\omega(a) \right] \right| > C\epsilon_M \right) \\ &= \sup_{P_0} \mathbb{P}_{P_0} \left( \left| \left( \sum_{i=1}^M \omega_i \right)^{-1} \sum_{i=1}^M \omega_i \left[ \frac{\mathbb{I}(A_i = a)}{\pi_a} (\bar{Y}_i - m_{ai}) + m_{ai} - \mu_\omega(a) \right] \right| > C\epsilon_M \right) \\ &\leq \sup_{P_0} \frac{\text{Var}_{P_0} \left( \left( \sum_{i=1}^M \omega_i \right)^{-1} \sum_{i=1}^M \omega_i \left[ \frac{\mathbb{I}(A_i = a)}{\pi_a} (\bar{Y}_i - m_{ai}) \right] \right)}{C^2 \epsilon_M^2} \\ &\leq \sup_{P_0} \frac{1}{C^2 \epsilon_M^2 b_\omega^2 M^2} \text{Var}_{P_0} \left( \sum_{i=1}^M \omega_i \left[ \frac{\mathbb{I}(A_i = a)}{\pi_a} (\bar{Y}_i - m_{ai}) \right] \right) \\ &\leq \sup_{P_0} \frac{B_\omega^2}{C^2 \epsilon_M^2 b_\omega^2 M^2} \sum_{i=1}^M \text{Var}_{P_0} \left( \frac{\mathbb{I}(A_i = a)}{\pi_a} (\bar{Y}_i - m_{ai}) \right) \\ &\leq \sup_{P_0} \frac{B_\omega^2}{C^2 \epsilon_M^2 b_\omega^2 M^2} \sum_{i=1}^M \mathbb{E}_{P_0} \left( \frac{\mathbb{I}(A_i = a)}{\pi_a^2} (\bar{Y}_i - m_{ai})^2 \right) \\ &\leq \sup_{P_0} \frac{B_\omega^2}{C^2 \epsilon_M^2 b_\omega^2 M^2} \sum_{i=1}^M \frac{1}{\pi_a^2} \mathbb{E}_{P_0} \left( (\bar{Y}_i - m_{ai})^2 \right) \\ &\leq \sup_{P_0} \frac{B_\omega^2}{C^2 \epsilon_M^2 b_\omega^2 M^2} \sum_{i=1}^M \frac{1}{\pi_a^2} \mathbb{E}_{P_0} \left( (\bar{Y}_i - m_{ai})^2 \right) \\ &\leq \frac{B_\omega^2 M K_y}{C^2 \epsilon_M^2 b_\omega^2 M^2 \pi_a^2} \\ &= \frac{B_\omega^2 K_y}{C^2 \epsilon_M^2 b_\omega^2 \pi_a^2} \cdot \frac{1}{M} \end{aligned}$$

The second equality holds because all components of  $B$  are functions of the observed data  $\mathbf{D}$  and are therefore constants when conditioning on  $\mathbf{D}$ . For any threshold  $\epsilon_M > 0$ , we can control the probability by Chebyshev's inequality (line 5). Following Assumption 1, we obtain a bound on the variance and conclude that, if  $\sqrt{M} \epsilon_M \rightarrow \infty$  (for example, if  $\epsilon_M > M^{-1/2}$ ), this probability converges to zero.

**Contraction rate of component  $A$**  Recall  $A = \left(\sum_{i=1}^M \omega_i\right)^{-1} \sum_{i=1}^M \omega_i (\hat{m}_{ai} - m_{ai}) \left(1 - \frac{\mathbb{I}(A_i=a)}{\pi_a}\right)$ , where the CRT randomization makes  $\pi_a \in (0, 1)$  known and the cluster weights satisfy  $0 < b_\omega \leq \omega_i \leq B_\omega < \infty$ . Let  $S_\omega := \sum_{i=1}^M \omega_i$ ,  $\mathcal{F}$  be a pointwise measurable class of functions  $f : \mathcal{X} \rightarrow \mathbb{R}$  with square-integrable envelope  $F$ . Further, define the centered class  $\mathcal{G} := \{g = f - m_a : f \in \mathcal{F}\}$  where  $m_a := \mathbb{E}(\bar{Y}_i | A_i = a, \mathbf{X}_i, N_i)$ .  $\mathcal{G}$  is  $P_0$ -Donsker with envelope  $G \in L_2(P_0)$ . For notational convenience, we further let  $\xi_i := \omega_i/S_\omega$  so  $\sum_i \xi_i = 1$  and  $\frac{b_\omega}{B_\omega M} \leq \xi_i \leq \frac{B_\omega}{b_\omega M}$ . Define the weighted and unweighted empirical/true measures:

$$P_M^\omega f := \sum_{i=1}^M \xi_i f(W_i), \quad P_M f := \frac{1}{M} \sum_{i=1}^M f(W_i), \quad P_0 f := \mathbb{E}[f(W)], \quad W_i = (A_i, \mathbf{X}_i, N_i),$$

and the weighted empirical process  $\mathbb{G}_M^\omega f := \sqrt{M}\{(P_M^\omega - P_0)f\}$ , as well as the unweighted empirical process  $\mathbb{G}_M f := \sqrt{M}\{(P_M - P_0)f\}$ . Let  $Z_i := 1 - \frac{\mathbb{I}(A_i=a)}{\pi_a}$ , then  $\mathbb{E}[Z_i | X_i] = 0$ ,  $\text{Var}[Z_i | X_i] = \frac{1-\pi_a}{\pi_a}$ , and  $|Z_i| \leq \pi_a^{-1}$ . Write  $g_{\hat{m}} := \hat{m}_a - m_a \in \mathcal{G}$  and define the index class:

$$\mathcal{H} := \{h_g(w) := zg(x) : g \in \mathcal{G}, w = (x, a), z = 1 - \mathbb{I}(a = a_0)/\pi_a\}.$$

By  $\mathbb{E}[Z_i | X_i] = 0$ , we have  $P_0 h_g = \mathbb{E}\{\mathbb{E}[Z|X]g(X)\} = 0$  for all  $g \in \mathcal{G}$ .

Given this setup,  $A$  is expressed as an empirical process at a random index as follows:

$$A = \sum_{i=1}^M \xi_i g_{\hat{m}}(X_i) Z_i = (P_M^\omega - P_0) h_{g_{\hat{m}}} = M^{-1/2} \mathbb{G}_M^\omega h_{g_{\hat{m}}}.$$

Since  $\mathcal{G}$  is  $P_0$ -Donsker, and  $|Z| \leq \pi_a^{-1}$  is bounded, the product class  $\mathcal{H} = \{Z g : g \in \mathcal{G}\}$  is  $P_0$ -Donsker with envelope  $H := |Z| G$  and  $\|H\|_2 \leq \pi_a^{-1} \|G\|_2 < \infty$ . Moreover, because  $b_\omega/(B_\omega M) \leq \xi_i \leq B_\omega/(b_\omega M)$ , the weighted process is dominated (up to constants) by the unweighted process:

$$\sup_{h \in \mathcal{H}} |\mathbb{G}_M^\omega h| \leq \frac{B_\omega^2}{b_\omega^2} \sup_{h \in \mathcal{H}} |\mathbb{G}_M h|.$$

By the standard symmetrization inequality (e.g., Lemma 2.3.1 by [van der Vaart and Wellner \(1996\)](#))

$$\mathbb{E} \left[ \sup_{h \in \mathcal{H}} |\mathbb{G}_M h| \right] \leq 2 \mathbb{E} \left[ \sup_{h \in \mathcal{H}} \left| \frac{1}{\sqrt{M}} \sum_{i=1}^M \varepsilon_i h(W_i) \right| \right],$$

where  $\{\varepsilon_i\}$  are the Rademacher variables independent of data. By the uniform entropy integral bound for  $P_0$ -Donsker class (e.g., Theorem 2.14.1 by [van der Vaart and Wellner \(1996\)](#)), for all  $M$ , we have

$$\mathbb{E} \left[ \sup_{h \in \mathcal{H}} |\mathbb{G}_M^\omega h| \right] \leq \frac{B_\omega^2}{b_\omega^2} \mathbb{E} \left[ \sup_{h \in \mathcal{H}} |\mathbb{G}_M h| \right] \leq 2 \frac{B_\omega^2}{b_\omega^2} \mathbb{E} \left[ \sup_{h \in \mathcal{H}} \left| \frac{1}{\sqrt{M}} \sum_{i=1}^M \varepsilon_i h(W_i) \right| \right] \leq 2 \frac{B_\omega^2}{b_\omega^2} C_{\mathcal{H}},$$

for some constant  $C_{\mathcal{H}} < \infty$ . Therefore, by Markov's inequality, we have

$$\sup_{h \in \mathcal{H}} |\mathbb{G}_M^\omega h| = O_p(1),$$

which proves  $A = M^{-1/2} \mathbb{G}_M^\omega h_{g_{\widehat{m}}} = O_p(M^{-1/2})$ . Combining the results for components  $A$  and  $B$ , we conclude that  $\widehat{\mu}_\omega^{\text{mr}}(a) - \mu_\omega(a) = A + B = O_p(M^{-1/2})$ , which establishes the overall root- $M$  consistency of the model-robust estimator.

### Appendix 3. Simulation Results for the Second Experiment Across Three Scenarios

Table A: Results from the second simulation experiment based on 1000 replications for scenario (i), estimating the cluster-average treatment effect,  $\Delta_C = 0.590$ , and the individual-average treatment effect,  $\Delta_I = 0.625$ . A linear mixed model, which adjusts for covariates with main effects and treatment-covariate interactions, is used as the fitted model. For the calibrated variance estimation method, 100 resampled datasets are used within each simulation to compute posterior variance.  $M$  denotes the number of clusters. “Relative Bias” refers to the relative bias of the estimated value expressed as a percentage, with standard errors shown in parentheses. “Coverage” refers to the 95% credible interval coverage across all simulated datasets expressed as a percentage. “MCSD” denotes Monte Carlo standard deviation. “AESE” denotes average of estimated standard error. “RE” refers to the relative efficiency of the proposed estimator (either the g-computation estimator or the model-robust estimator) with covariate adjustment, compared to the unadjusted moment estimator.

Estimator	Posterior Inference	M	Estimand	Relative Bias (SE)	Coverage	MCSD	AESE	RE
g-computation	Uncalibrated	30	cluster-ATE	-2.1 (1.2)	94.5	0.224	0.244	15.666
			individual-ATE	-2.0 (1.2)	94.4	0.237	0.251	20.035
		60	cluster-ATE	-2.3 (0.8)	94.0	0.145	0.145	19.643
			individual-ATE	-2.4 (0.8)	93.7	0.155	0.152	24.731
g-computation	Calibrated	30	cluster-ATE	-2.1 (1.2)	96.1	0.224	0.256	15.666
			individual-ATE	-2.0 (1.2)	95.7	0.237	0.262	20.035
		60	cluster-ATE	-2.3 (0.8)	95.3	0.145	0.149	19.643
			individual-ATE	-2.4 (0.8)	94.6	0.155	0.156	24.731
Model-robust	Uncalibrated	30	cluster-ATE	-2.2 (1.2)	50.4	0.223	0.112	15.806
			individual-ATE	-2.1 (1.2)	51.2	0.235	0.110	20.378
		60	cluster-ATE	-2.3 (0.8)	38.6	0.142	0.038	20.482
			individual-ATE	-2.4 (0.8)	37.3	0.151	0.039	26.059
Model-robust	Calibrated	30	cluster-ATE	-2.2 (1.2)	92.6	0.223	0.228	15.806
			individual-ATE	-2.1 (1.2)	92.0	0.235	0.232	20.378
		60	cluster-ATE	-2.3 (0.8)	94.0	0.142	0.139	20.482
			individual-ATE	-2.4 (0.8)	93.6	0.151	0.145	26.059

Table B: Results from the second simulation experiment based on 1000 replications for scenario (ii), estimating the cluster-average treatment effect,  $\Delta_C = 0.679$ , and the individual-average treatment effect,  $\Delta_I = 0.842$ . A linear mixed model, which adjusts for covariates with main effects and treatment-covariate interactions, is used as the fitted model. For the calibrated variance estimation method, 100 resampled datasets are used within each simulation to compute posterior variance.  $M$  denotes the number of clusters. “Relative Bias” refers to the relative bias of the estimated value expressed as a percentage, with standard errors shown in parentheses. “Coverage” refers to the 95% credible interval coverage across all simulated datasets expressed as a percentage. “MCSD” denotes Monte Carlo standard deviation. “AESE” denotes average of estimated standard error. “RE” refers to the relative efficiency of the proposed estimator (either the g-computation estimator or the model-robust estimator) with covariate adjustment, compared to the unadjusted moment estimator.

Estimator	Posterior Inference	M	Estimand	Relative Bias (SE)	Coverage	MCSD	AESE	RE
g-computation	Uncalibrated	30	cluster-ATE	-2.6 (1.6)	82.8	0.347	0.257	1.404
			individual-ATE	-3.3 (1.3)	81.9	0.358	0.263	1.517
		60	cluster-ATE	-2.3 (1.1)	79.9	0.231	0.154	1.629
			individual-ATE	-3.2 (0.9)	80.9	0.240	0.160	1.738
g-computation	Calibrated	30	cluster-ATE	-2.6 (1.6)	94.8	0.347	0.363	1.404
			individual-ATE	-3.3 (1.3)	94.8	0.358	0.373	1.517
		60	cluster-ATE	-2.3 (1.1)	95.2	0.231	0.231	1.629
			individual-ATE	-3.2 (0.9)	94.2	0.240	0.242	1.738
Model-robust	Uncalibrated	30	cluster-ATE	-2.8 (1.6)	38.4	0.346	0.117	1.408
			individual-ATE	-3.4 (1.3)	39.7	0.357	0.116	1.524
		60	cluster-ATE	-2.3 (1.1)	27.1	0.230	0.041	1.649
			individual-ATE	-3.1 (0.9)	27.6	0.238	0.042	1.763
Model-robust	Calibrated	30	cluster-ATE	-2.8 (1.6)	93.4	0.346	0.342	1.408
			individual-ATE	-3.4 (1.3)	93.6	0.357	0.349	1.524
		60	cluster-ATE	-2.3 (1.1)	94.5	0.230	0.225	1.649
			individual-ATE	-3.1 (0.9)	93.6	0.238	0.233	1.763

Table C: Results from the second simulation experiment based on 1000 replications for scenario (iii), estimating the cluster-average treatment effect,  $\Delta_C = 0.637$ , and the individual-average treatment effect,  $\Delta_I = 0.754$ . A linear mixed model, which adjusts for covariates with main effects and treatment-covariate interactions, is used as the fitted model. For the calibrated variance estimation method, 100 resampled datasets are used within each simulation to compute posterior variance.  $M$  denotes the number of clusters. “Relative Bias” refers to the relative bias of the estimated value expressed as a percentage, with standard errors shown in parentheses. “Coverage” refers to the 95% credible interval coverage across all simulated datasets expressed as a percentage. “MCSD” denotes Monte Carlo standard deviation. “AESE” denotes average of estimated standard error. “RE” refers to the relative efficiency of the proposed estimator (either the g-computation estimator or the model-robust estimator) with covariate adjustment, compared to the unadjusted moment estimator.

Estimator	Posterior Inference	M	Estimand	Relative Bias (SE)	Coverage	MCSD	AESE	RE
g-computation	Uncalibrated	30	cluster-ATE	-27.0 (23.9)	85.0	4.810	3.831	1.050
			individual-ATE	-40.1 (24.3)	80.0	5.787	4.216	1.168
		60	cluster-ATE	1.7 (17.1)	83.8	3.446	2.670	1.193
			individual-ATE	-9.4 (17.6)	76.5	4.203	2.880	1.299
g-computation	Calibrated	30	cluster-ATE	-27.0 (23.9)	87.1	4.810	3.995	1.050
			individual-ATE	-40.1 (24.3)	84.6	5.787	4.379	1.168
		60	cluster-ATE	1.7 (17.1)	86.8	3.446	2.733	1.193
			individual-ATE	-9.4 (17.6)	80.8	4.203	2.942	1.299
Model-robust	Uncalibrated	30	cluster-ATE	-28.0 (21.3)	47.3	4.293	1.447	1.318
			individual-ATE	-34.1 (22.7)	40.0	5.419	1.532	1.331
		60	cluster-ATE	-0.7 (15.6)	31.9	3.145	0.688	1.432
			individual-ATE	-2.3 (16.9)	26.6	4.030	0.719	1.413
Model-robust	Calibrated	30	cluster-ATE	-28.0 (21.3)	93.0	4.293	4.154	1.318
			individual-ATE	-34.1 (22.7)	91.2	5.419	5.091	1.331
		60	cluster-ATE	-0.7 (15.6)	93.4	3.145	2.921	1.432
			individual-ATE	-2.3 (16.9)	91.8	4.030	3.689	1.413

#### Appendix 4. BART simulation results with informative cluster size

Table D: BART-based results under varying scenarios with informative cluster sizes, based on 250 simulations, estimating the cluster-average treatment effect,  $\Delta_C$ , and the individual-average treatment effect,  $\Delta_I$ . A BART model, which adjusts for covariates with main effects and treatment-covariate interactions, is used as the fitted model. For the calibrated variance estimation method, 100 resampled datasets are used within each simulation to compute posterior variance. “Relative Bias” refers to the relative bias of the estimated value expressed as a percentage, with standard errors shown in parentheses. “Coverage” refers to the 95% credible interval coverage across all simulated datasets expressed as a percentage. “MCSD” denotes Monte Carlo standard deviation. “AESE” denotes average of estimated standard error.

Scenario	Estimator	Posterior Inference	M	Estimand	Relative Bias (SE)	Coverage	MCSD	AESE
(ii)	g-computation	Uncalibrated	30	cluster-ATE	-8.4 (3.6)	76.8	0.384	0.243
				individual-ATE	-12.7 (3.0)	75.6	0.395	0.247
			60	cluster-ATE	-0.3 (2.0)	83.6	0.218	0.155
				individual-ATE	-5.1 (1.7)	84.0	0.226	0.157
			90	cluster-ATE	-2.8 (1.6)	81.6	0.177	0.121
				individual-ATE	-5.3 (1.4)	77.6	0.188	0.123
	Model-robust	Calibrated	30	cluster-ATE	-7.9 (3.7)	84.8	0.392	0.285
				individual-ATE	-10.3 (3.1)	82.4	0.407	0.292
			60	cluster-ATE	0.0 (2.0)	92.4	0.219	0.205
				individual-ATE	-3.3 (1.7)	92.4	0.225	0.215
			90	cluster-ATE	-2.7 (1.7)	91.2	0.180	0.170
				individual-ATE	-3.5 (1.5)	90.8	0.195	0.178
(iii)	g-computation	Uncalibrated	30	cluster-ATE	-54.6 (37.7)	89.2	3.796	3.318
				individual-ATE	-59.8 (34.0)	88.0	4.057	3.354
			60	cluster-ATE	35.1 (23.1)	92.8	2.327	2.190
				individual-ATE	22.3 (21.1)	91.2	2.513	2.217
			90	cluster-ATE	-12.3 (15.5)	94.0	1.560	1.628
				individual-ATE	-20.4 (14.2)	93.6	1.696	1.658
	Model-robust	Calibrated	30	cluster-ATE	-40.1 (41.5)	89.6	4.176	3.363
				individual-ATE	-44.8 (43.3)	87.2	5.166	4.029
			60	cluster-ATE	39.8 (24.2)	91.6	2.435	2.074
				individual-ATE	36.9 (25.4)	89.6	3.030	2.515
			90	cluster-ATE	-6.6 (16.0)	91.6	1.616	1.489
				individual-ATE	-7.0 (16.6)	93.6	1.979	1.803

## Appendix 5. Simulation results without informative cluster size

To guarantee there is no informative cluster sizes, we defined our data-generating process as follows. Let  $(N_i, C_{i1}, C_{i2})$ , for  $i = 1, \dots, M$ , be independent draws from distribution  $\mathcal{P}^N \times \mathcal{P}^{C_1} \times \mathcal{P}^{C_2|C_1}$ , where  $\mathcal{P}^N$  is uniform over support  $(10, 50)$ ,  $\mathcal{P}^{C_1} = \mathcal{N}(3, 4)$ ,  $\mathcal{P}^{C_2|C_1} = \mathcal{B}[\text{expit}\{\log(3) \times C_1\}]$  is a Bernoulli distribution with  $\text{expit}(x) = (1 + e^{-x})^{-1}$ . Next, for each individual, we generated the two individual-level covariates from  $X_{ij1} \sim \mathcal{B}(0.6)$ , and  $X_{ij2} \sim \mathcal{N}\{X_{ij1} \times (2C_{i2} - 1), 9\}$ . Potential outcomes  $Y_{ij}(A_i)$  are generated from normal distribution  $Y_{ij}(A_i) \sim \mathcal{N}(\psi_{ij}(A_i), 1)$ , where  $\psi_{ij}(A_i) \sim \mathcal{N}(\eta_{ij}(A_i), \sigma_\phi^2)$ , and  $\eta_{ij}(A_i)$  is the fixed effect portion of the outcome-generating model,  $\eta_{ij}(A_i) = -0.1 - \frac{0.3}{1+\exp\{-6\times(30+X_{ij1}+X_{ij2})\}} + 0.2X_{ij1}X_{ij2} + 0.3 \times (3 + X_{ij1})A_i - \frac{1}{1+\exp\{-4(X_{ij1}+X_{ij2})\}}A_i$ . Let  $\sigma_\phi^2 = 0.25$ , which results in a ICC of 0.2. The observed outcomes are generated by independently sampling  $A_i \sim \mathcal{B}(0.5)$  and defining them as  $Y_{ij} = A_i Y_{ij}(1) + (1 - A_i) Y_{ij}(0)$ . Tables E and F summarize the simulation results under this scenario, evaluating the performance of LMM and BART, respectively, in estimating both the cluster-ATE and individual-ATE.

Table E: LMM-based results under the scenario with no informative cluster size, based on 250 simulations, estimating the cluster-average treatment effect,  $\Delta_C = 0.54$ , and the individual-average treatment effect,  $\Delta_I = 0.54$ . A BART model, which adjusts for covariates with main effects and treatment-covariate interactions, is used as the fitted model. For the calibrated variance estimation method, 100 resampled datasets are used within each simulation to compute posterior variance. “Relative Bias” refers to the relative bias of the estimated value expressed as a percentage, with standard errors shown in parentheses. “Coverage” refers to the 95% credible interval coverage across all simulated datasets expressed as a percentage. “MCSD” denotes Monte Carlo standard deviation. “AESE” denotes average of estimated standard error.

Estimator	Posterior Inference	M	Estimand	Relative Bias (SE)	Coverage	MCSD	AESE
g-computation	Uncalibrated	30	cluster-ATE	-3.6 (2.8)	91.6	0.240	0.250
			individual-ATE	-2.3 (3.0)	93.6	0.253	0.261
		60	cluster-ATE	-1.4 (1.7)	95.6	0.146	0.151
			individual-ATE	-2.5 (1.9)	95.2	0.158	0.157
		90	cluster-ATE	-0.9 (1.4)	94.8	0.123	0.123
			individual-ATE	-0.7 (1.5)	95.6	0.126	0.129
Model-robust	Calibrated	30	cluster-ATE	-3.4 (2.8)	91.6	0.238	0.233
			individual-ATE	-2.3 (2.9)	89.6	0.248	0.242
		60	cluster-ATE	-0.8 (1.7)	95.2	0.144	0.146
			individual-ATE	-2.1 (1.8)	94.4	0.156	0.151
		90	cluster-ATE	-0.6 (1.4)	94.0	0.123	0.120
			individual-ATE	-0.4 (1.5)	94.4	0.125	0.124

Table F: BART-based results under the scenario with no informative cluster size, based on 250 simulations, estimating the cluster-average treatment effect,  $\Delta_C = 0.54$ , and the individual-average treatment effect,  $\Delta_I = 0.54$ . A BART model, which adjusts for covariates with main effects and treatment-covariate interactions, is used as the fitted model. For the calibrated variance estimation method, 100 resampled datasets are used within each simulation to compute posterior variance. “Relative Bias” refers to the relative bias of the estimated value expressed as a percentage, with standard errors shown in parentheses. “Coverage” refers to the 95% credible interval coverage across all simulated datasets expressed as a percentage. “MCSD” denotes Monte Carlo standard deviation. “AESE” denotes average of estimated standard error.

Estimator	Posterior Inference	M	Estimand	Relative Bias (SE)	Coverage	MCSD	AESE
g-computation	Uncalibrated	30	cluster-ATE	-2.0 (2.4)	92.8	0.205	0.193
			individual-ATE	-1.9 (2.4)	93.2	0.206	0.193
		60	cluster-ATE	-1.7 (1.5)	96.0	0.131	0.137
			individual-ATE	-1.6 (1.5)	96.0	0.131	0.137
		90	cluster-ATE	-0.3 (1.3)	94.4	0.113	0.113
			individual-ATE	-0.3 (1.3)	94.8	0.113	0.113
Model-robust	Calibrated	30	cluster-ATE	-1.4 (2.4)	91.6	0.207	0.192
			individual-ATE	-0.9 (2.6)	91.2	0.218	0.202
		60	cluster-ATE	-1.1 (1.5)	95.6	0.131	0.137
			individual-ATE	-2.4 (1.7)	94.0	0.144	0.142
		90	cluster-ATE	-0.1 (1.3)	93.2	0.113	0.114
			individual-ATE	0.3 (1.4)	93.6	0.119	0.119

## Appendix 6. Comparison of relative efficiency from the second simulation experiment

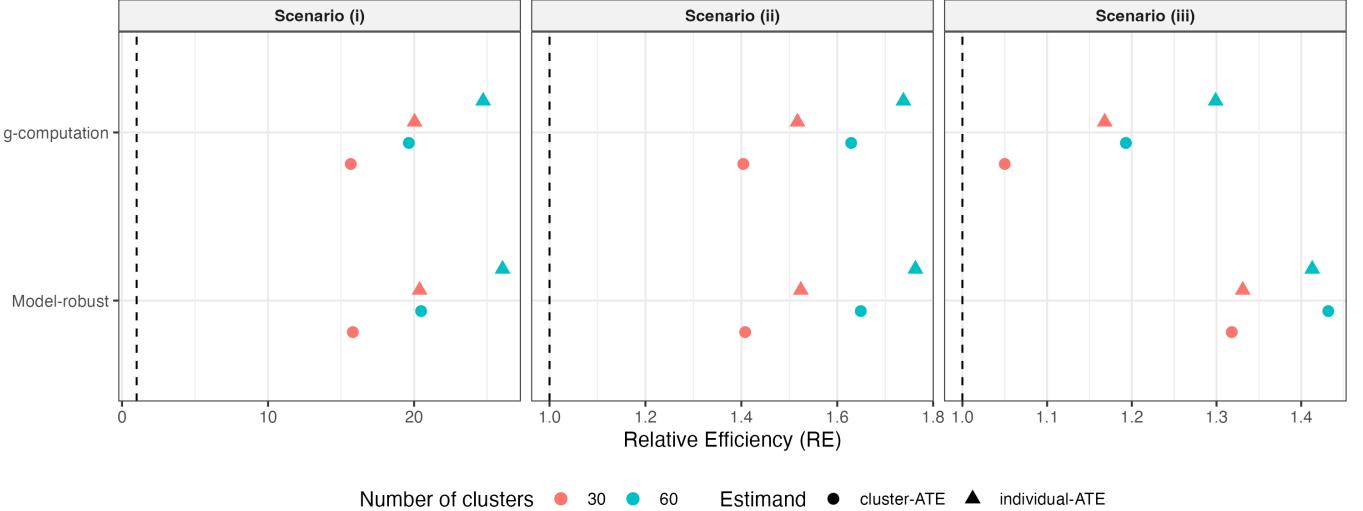


Figure A1: Relative efficiency (RE) from the second simulation experiment based on 1000 replications for scenario (i), (ii), and (iii), estimating the cluster-average treatment effect (cluster-ATE) and individual-average treatment effect (individual-ATE). The point estimation methods are categorized into g-computation and model-robust approaches. A linear mixed model, which adjusts for covariates with main effects and treatment-covariate interactions, is used as the fitted model. Shapes distinguish estimands (circles for cluster-ATE and triangles for individual-ATE), and colors indicate the number of clusters ( $M = 30$  in red,  $M = 60$  in blue). The horizontal dashed line at  $RE = 1$  marks the reference value where the proposed estimator (either the g-computation estimator or the model-robust estimator) with covariate adjustment has the same efficiency as the unadjusted moment estimator; values above 1 indicate higher efficiency.