Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Model-Robust and Efficient Covariate Adjustment for Cluster-Randomized Experiments

Bingkai Wang, Chan Park, Dylan S. Small & Fan Li

To cite this article: Bingkai Wang, Chan Park, Dylan S. Small & Fan Li (2024) Model-Robust and Efficient Covariate Adjustment for Cluster-Randomized Experiments, Journal of the American Statistical Association, 119:548, 2959-2971, DOI: 10.1080/01621459.2023.2289693

To link to this article: https://doi.org/10.1080/01621459.2023.2289693

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2024, VOL. 119, NO. 548, 2959–2971: Theory and Methods
https://doi.org/10.1080/01621459.2023.2289693

Taylor & Francis
Taylor & Francis Group

Model-Robust and Efficient Covariate Adjustment for Cluster-Randomized Experiments

Bingkai Wang[a], Chan Park[a], Dylan S. Small[a], and Fan Li[b]

[a]The Statistics and Data Science Department of the Wharton School, University of Pennsylvania, Philadelphia, PA; [b]Department of Biostatistics and Center for Methods in Implementation and Prevention Science, Yale School of Public Health, New Haven, CT

ABSTRACT

Cluster-randomized experiments are increasingly used to evaluate interventions in routine practice conditions, and researchers often adopt model-based methods with covariate adjustment in the statistical analyses. However, the validity of model-based covariate adjustment remains unclear when the working models are misspecified, leading to ambiguity of estimands and risk of bias. In this article, we first adapt two model-based methods—generalized estimating equations and linear mixed models—with weighted g-computation to achieve robust inference for cluster-average and individual-average treatment effects. To further overcome the limitations of model-based covariate adjustment methods, we propose efficient estimators for each estimand that allow for flexible covariate adjustment and additionally address cluster size variation dependent on treatment assignment and other cluster characteristics. Such cluster size variations often occur post-randomization and, if ignored, can lead to bias of model-based estimators. For our proposed covariate-adjusted estimators, we prove that when the nuisance functions are consistently estimated by machine learning algorithms, the estimators are consistent, asymptotically normal, and efficient. When the nuisance functions are estimated via parametric working models, the estimators are triply-robust. Simulation studies and analyses of three real-world cluster-randomized experiments demonstrate that the proposed methods are superior to existing alternatives. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received December 2022
Accepted November 2023

KEYWORDS
Causal inference; Cluster-randomized trial; Covariate adjustment; Efficient influence function; Estimands; Machine learning

1. Introduction

Cluster-randomized experiments refer to study designs that randomize treatment at the cluster level; clusters can be villages, hospitals, or worksites (Murray 1998; Donner and Klar 2000). Cluster randomization is often used to study group-level interventions or to prevent treatment contamination under individual randomization, and is increasingly adopted in pragmatic clinical trials evaluating interventions in routine practice conditions. For example, among 10 demonstration projects across different disease areas (from 2012 to 2017) supported by the United States *National Institute of Health Pragmatic Clinical Trials Collaboratory*, eight studies adopted cluster randomization (Table 1 in Weinfurt et al. 2017).

In the analyses of cluster-randomized experiments, covariate adjustment is essential to address baseline chance imbalance and improve the precision of the treatment effect estimator. However, challenges in covariate adjustment persist due to the complexity associated with the multilevel data structure under cluster randomization. First, although model-based methods, including generalized estimating equations (GEE, Liang and Zeger 1986) and generalized linear mixed models (Breslow and Clayton 1993), are commonly used to perform covariate adjustment in cluster-randomized experiments, it remains unclear whether the treatment effect coefficient corresponds to a clearly-defined estimand of interest, especially when the working model is misspecified. Even in the absence of covariates, Wang et al. (2022) has demonstrated that the treatment coefficient estimator from the GEE with an exchangeable working correlation structure corresponds to an ambiguous estimand when the cluster size is informative for causal effects. Second, participants are frequently sampled after cluster randomization such that the observed cluster size may depend on the treatment assignment and other cluster attributes. Failure to address such cluster-dependent sampling schemes, as in standard techniques, can lead to bias in estimating the treatment effect. For example, when the observed cluster size depends on cluster-level covariates, Bugni et al. (2023) demonstrated that the standard difference-in-means estimator is biased for typical estimands of interest. To date, robust causal inference methods that can simultaneously maximize the precision gain from covariate adjustment and address cluster-dependent sampling are unavailable for cluster-randomized experiments.

In this article, our primary contribution is to provide new covariate-adjusted estimators for cluster-randomized experiments that target clearly-defined estimands with minimal model assumptions, improve precision over standard methods,

CONTACT Bingkai Wang bingkai.w@gmail.com The Statistics and Data Science Department of the Wharton School, University of Pennsylvania, Philadelphia, PA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

and address cluster-dependent sampling. We focus on two classes of causal estimands: the cluster-average treatment effect and the individual-average treatment effect. The former estimand addresses the question of "what is the expected change in outcome associated with treatment for a typical cluster with its natural source population?" and gives equal weight to each cluster, whereas the latter estimand addresses the question of "what is the expected change in outcome associated with treatment for a typical individual?" and gives equal weight to each individual. The two estimands represent treatment effects at different levels and differ when there is treatment effect heterogeneity by cluster size (Kahan et al. 2023). To estimate both estimands, we first adapt GEE and linear mixed models through weighted g-computation, and provide a set of sufficient conditions to achieve model-robust inference, that is, consistency and asymptotic normality under arbitrary working model misspecification. These new insights help clarify when conventional multilevel regression models, which are routinely used in standard practice (Turner et al. 2017), provide valid average causal effect estimators even if the regression model formulation differs from the unknown and potentially complex data-generating process. To the best of our knowledge, this entire set of sufficient conditions for typical model-based methods to achieve model-robust inference has not been elucidated in the prior literature, and can inform current practice in analyzing cluster-randomized experiments.

The weighted g-computation estimator based on GEE or linear mixed models, however, is subject to two potential limitations. First, it can be biased when the sufficient conditions for model robustness fail to hold, for example, when the observed cluster size depends on treatment assignment and cluster-level covariates. Second, such an estimator, albeit robust under certain types of model misspecification, is not guaranteed to improve estimation precision through covariate adjustment, and the linear working models may be too restrictive to maximize the precision gain. These two limitations motivate us to develop more principled estimators for covariate adjustment. To achieve this goal, we characterize the efficient influence function for both the cluster-average and the individual-average causal effect estimands and propose efficient estimators that allow for flexible covariate adjustment and simultaneously address cluster-dependent sampling. When the nuisance functions are estimated by parametric models, such as GEE or generalized linear mixed models, the proposed estimators are triply-robust; that is, they are consistent to the specified causal estimand if two out of the three nuisance functions are consistently estimated. When the nuisance working models are all consistently estimated, for example, by machine learning algorithms with cross-fitting, the proposed estimators achieve the semiparametric efficiency lower bounds given our causal models. Compared to model-based covariate adjustment, the efficient estimators notably increase the flexibility in covariate adjustment and substantially enrich the toolbox for analyzing cluster-randomized experiments.

Our results build on but differ from the existing literature on causal inference for cluster-randomized experiments. Imai, King, and Nall (2009) and Middleton and Aronow (2015) proposed cluster-level methods for estimating the average treatment effects but did not consider covariate adjustment to improve efficiency. Schochet et al. (2022) established the finite-population central limit theorem for linearly-adjusted estimators under blocked cluster randomization, and Su and Ding (2021) extended their results by providing a unified theory for a class of weighted average treatment effect estimands. These prior efforts considered a finite-population framework with linear working models and an independence working correlation structure. In contrast, we study causal effect estimators under a super-population framework and address robust and efficient estimation under a much wider class of working models that are not limited to linear specifications. Balzer et al. (2023) and Benitez et al. (2021) applied targeted maximum likelihood estimation under hierarchical structural models for cluster-randomized experiments. However, they did not consider cluster size variation arising from cluster-dependent sampling. Bugni et al. (2023) adapted moment-based estimators to address cluster-dependent sampling but did not consider covariate adjustment. None of these prior results have addressed efficient causal inference under cluster-dependent sampling. We therefore fill in this important gap by proposing efficient estimators to achieve flexible covariate adjustment and accommodate post-randomization cluster-dependent sampling.

The remainder of the article is organized as follows. In Section 2, we formalize our super-population framework, present the causal estimands of interest, and structural assumptions for identification. Section 3 adapts GEE and linear mixed models to target our estimands. In Section 4, we develop the efficient influence functions and propose our efficient estimators. In Sections 5 and 6, we demonstrate our theoretical results via simulation studies and reanalyses of three cluster-randomized experiments. Section 7 concludes with a discussion.

2. Notation, Estimands, and Assumptions

We consider a cluster-randomized experiment with m clusters. For each cluster i, we let N_i denote the total number of individuals in the underlying source population, M_i be the observed number of individuals sampled into the study, $A_i \in \{0,1\}$ be the cluster-level treatment indicator, and C_i be a q-dimensional vector of cluster-level covariates. For each individual j in cluster i, we define Y_{ij} as the outcome, X_{ij} as a p-dimensional vector of individual-level covariates, and $S_{ij} \in \{0,1\}$ as the sampling indicator, that is, recruited into the experiment. By definition, the observed cluster size $M_i = \sum_{j=1}^{N_i} S_{ij} \leq N_i$.

We proceed under the potential outcomes framework (Neyman, Dabrowska, and Speed 1990) and define $Y_{ij}(a)$ as the potential outcome and $S_{ij}(a)$ as the potential sampling indicator if cluster i were assigned to treatment group $a \in \{0,1\}$. We assume consistency such that $Y_{ij} = A_i Y_{ij}(1) + (1-A_i)Y_{ij}(0)$ and $S_{ij} = A_i S_{ij}(1) + (1-A_i)S_{ij}(0)$. As a result, the potential observed cluster size in a treated cluster, denoted as $M_i(1) = \sum_{j=1}^{N_i} S_{ij}(1)$, can be different from its counterpart in a control cluster, denoted as $M_i(0) = \sum_{j=1}^{N_i} S_{ij}(0)$. Defining $Y_i(a) = \{Y_{i1}(a), \ldots, Y_{iN_i}(a)\} \in \mathbb{R}^{N_i}$ as the collection of potential outcomes, $S_i(a) = \{S_{i1}(a), \ldots, S_{iN_i}(a)\} \in \mathbb{R}^{N_i}$ as the collection of potential sampling indicators, and $\mathrm{X}_i = (X_{i1}, \ldots, X_{iN_i})^\top \in$

$\mathbb{R}^{N_i \times p}$, we write the collection of random variables in a cluster as $W_i = \{Y_i(1), Y_i(0), S_i(1), S_i(0), N_i, C_i, \mathrm{X}_i\}$. We next introduce the following assumptions on the complete, while not fully observed, data $\{(W_1, A_1), \ldots, (W_m, A_m)\}$.

Assumption 1 (Super-population). (a) Random variables $W_1, \ldots, W_m$ are mutually independent. (b) The source population size N_i follows an unknown distribution $\mathcal{P}^N$ over a finite support on $\mathbb{N}^+$. (c) Given N_i, W_i follows an unknown distribution $\mathcal{P}^{W|N}$ with finite second moments.

Assumption 2 (Cluster randomization). The treatment indicator A_i for each cluster is an independent realization from a Bernoulli distribution $\mathcal{P}^A$ with $pr(A = 1) = \pi \in (0, 1)$. Furthermore, $(A_1, \ldots, A_m)$ is independent of $(W_1, \ldots, W_m)$.

Assumption 3 (Cluster-dependent sampling). For $a \in \{0, 1\}$, the potential observed cluster size $M(a) = h_a(N, C, \epsilon_a)$ for some unknown function h_a and exogenous random noise ϵ_a that is independent of $\{N, C, \mathrm{X}, Y(a)\}$. Furthermore, for each possible N-dimensional binary vector s with $M(a)$ ones, $\mathrm{pr}\{S(a) = s \mid Y(a), M(a), \mathrm{X}, N, C\} = \binom{N}{M(a)}^{-1}$.

Assumption 1(a) implies that the data vectors from different clusters are independent, while the outcomes and covariates in the same cluster can be arbitrarily correlated. Assumption 1(b)–(c) formalize the condition that $W_1, \ldots, W_m$ are marginally identically distributed according to a mixture distribution, $\mathcal{P}^{W|N} \times \mathcal{P}^N$. This technical condition is useful for handling the varying dimension of W_i across clusters. Assumption 2 describes the simple cluster randomization design, which we use as a starting place to present our main results. In Sections 3 and 4, we also discuss how our results can be applied to stratified cluster randomization (Zelen 1974) and biased-coin cluster randomization (Efron 1971), which are two typical restricted cluster randomization schemes. Given Assumptions 1–2, the expectation on (W_i, A_i) is taken with respect to $\mathcal{P}^{W|N} \times \mathcal{P}^N \times \mathcal{P}^A$. Assumption 3 implies that the number of sampled individuals can depend on the assignment A and cluster characteristics (source population size N and cluster covariates C), but the sampling process is completely random given the number of sampled individuals and the source population size. This assumption relaxes the setting of Bugni et al. (2023) to allow for arm-specific sampling. Important special cases of Assumptions 3 include full enrollment (or focusing on the population of sampled individuals) such that $M(a) = N$; random sampling with an arm-specific sample size such that $M(a) = m_a$ for some integer $m_a \leq N$; and independent cluster-specific sampling such that each $S_{ij}(a)$ is independently determined by flipping a cluster-specific coin. Assumption 3 can be violated when sampling additionally depends on individual-level covariates, which are generally unobserved for nonparticipants. This sampling mechanism leads to post-randomization selection bias (Li et al. 2022), and we leave this form of selection bias for separate work.

We define the class of cluster-average treatment effect (Δ_C) and individual-average treatment effect (Δ_I) estimands as

$$\Delta_C = f\{\mu_C(1), \mu_C(0)\}, \quad \Delta_I = f\{\mu_I(1), \mu_I(0)\},$$

where f is a pre-specified smooth function determining the scale of effect measure and, for $a = 0, 1$,

$$\mu_C(a) = E\left\{\frac{\sum_{j=1}^{N_i} Y_{ij}(a)}{N_i}\right\}, \quad \mu_I(a) = \frac{E\left\{\sum_{j=1}^{N_i} Y_{ij}(a)\right\}}{E(N_i)}.$$

For example, $f(x, y) = x - y$ leads to the difference estimands, $f(x, y) = x/y$ leads to the relative risk estimands, and $f(x, y) = x(1 - y)/\{y(1 - x)\}$ leads to the odds ratio estimands. These two classes of estimands Δ_C and Δ_I differ based on the corresponding treatment-specific mean potential outcomes, $\mu_C(a)$ versus $\mu_I(a)$. The former represents the average potential outcome associated with treatment a for a typical cluster along with its natural source population, while the latter represents the average potential outcome associated with treatment a for a typical individual. Depending on the nature of the intervention and the endpoint, either or both estimands may be relevant in a given cluster-randomized experiment. These two estimands can be considered as the super-population analogs to those studied in Su and Ding (2021) and Kahan et al. (2023) under the finite-population framework and a generalization of the difference estimands in Bugni et al. (2023). When the source population size $N_i = n$ is a constant and each element in $\{Y_{ij}, j = 1, \ldots, n\}$ is assumed to be marginally identically distributed, we have $\mu_C(a) = \mu_I(a) = E[Y_{ij}(a)]$, a special case considered in Wang et al. (2021) where the two types of estimands coincide and bear the same interpretation as the typical estimand in an individual-randomized experiment. We refer to Kahan et al. (2023) for a more elaborate discussion on differentiating these two estimands and their example applications to cluster-randomized experiments.

Finally, we write the observed data for cluster i as $\mathcal{O}_i = \{Y_i^o, M_i, A_i, N_i, C_i, \mathrm{X}_i^o\}$, where $Y_i^o = \{Y_{ij} : S_{ij} = 1, j = 1, \ldots, N_i\} \in \mathbb{R}^{M_i}$ and $\mathrm{X}_i^o = \{X_{ij} : S_{ij} = 1, j = 1, \ldots, N_i\} \in \mathbb{R}^{M_i \times p}$. The relationships among these random variables are visualized in Figure S1 in the supplementary material. The central task is to estimate Δ_C and Δ_I with $\mathcal{O}_1, \ldots, \mathcal{O}_m$. Of note, we assume that the source population size N_i is available or can be elicited from either historical data or cluster stakeholders; this is often feasible for cluster-randomized experiments conducted within schools, worksites, or healthcare delivery systems. When the source population size in each cluster is unknown, we will discuss how to estimate Δ_C in Section 4, but Δ_I is generally not identifiable unless we equate the observed cluster size with the source population size by setting $N = M$.

3. Model-Based Covariate Adjustment

3.1. Generalized Estimating Equations (GEE)

One popular approach to analyze cluster-randomized experiments is through GEE, often specified through the marginal mean model, $g\{E(Y_{ij} \mid U_{ij})\} = U_{ij}^\top \beta$, where g is the link function, $U_{ij} = (1, A_i, L_{ij}^\top)^\top$ for user-specified covariates L_{ij} as an arbitrary function of (N_i, C_i, X_{ij}), and $\beta = (\beta_0, \beta_A, \beta_L^\top)^\top$. We assume that g is the canonical link, for example, $g(x) = x$ for continuous outcomes and $g(x) = \log\{x/(1 - x)\}$ for binary outcomes. The parameters β are estimated by solving the following estimating equations:

$$\sum_{i=1}^{m} \mathrm{D}_i^\top \mathrm{V}_i^{-1}(Y_i^o - \mu_i^o) = 0, \tag{1}$$

where $\mu_i^o = \{E(Y_{ij} \mid U_{ij}) : S_{ij} = 1\} \in \mathbb{R}^{M_i}$, $\mathrm{D}_i = \partial \mu_i^o / \partial \beta^\top$, $\mathrm{V}_i = \mathrm{Z}_i^{1/2} \mathrm{R}_i(\rho) \mathrm{Z}_i^{1/2}$ with $\mathrm{R}_i(\rho) \in \mathbb{R}^{M_i \times M_i}$ being a working correlation matrix and $\mathrm{Z}_i = \mathrm{diag}\{v(Y_{ij}|U_{ij}) : S_{ij} = 1\} \in \mathbb{R}^{M_i \times M_i}$ for some known variance function v. We consider two working correlation structures that are commonly used for analyzing cluster-randomized experiments: the independence correlation, $\mathrm{R}_i(\rho) = \mathrm{I}_{M_i}$, and the exchangeable correlation, $\mathrm{R}_i(\rho) = (1-\rho)\mathrm{I}_{M_i} + \rho 1_{M_i} 1_{M_i}^\top$, where $\mathrm{I}_{M_i} \in \mathbb{R}^{M_i \times M_i}$ is the identity matrix and $1_{M_i} \in \mathbb{R}^{M_i}$ is a vector of ones. The estimator for β is denoted by $\widehat{\beta}$, and the correlation parameter ρ is estimated by a moment estimator $\widehat{\rho}$ described in Example 4 of Liang and Zeger (1986). For analyzing cluster-randomized experiments, a conventional practice is to directly use the coefficient $\widehat{\beta}_A$ along with the robust sandwich variance for inference. However, this practice can lead to an ambiguous treatment effect estimand even in the absence of covariate adjustment (Wang et al. 2022).

To estimate our target estimands Δ_C and Δ_I, we propose weighted g-computation estimators, defined as $\widehat{\Delta}_C^{\text{GEE-g}} = f\left\{\widehat{\mu}_C^{\text{GEE-g}}(1), \widehat{\mu}_C^{\text{GEE-g}}(0)\right\}$ and $\widehat{\Delta}_I^{\text{GEE-g}} = f\left\{\widehat{\mu}_I^{\text{GEE-g}}(1), \widehat{\mu}_I^{\text{GEE-g}}(0)\right\}$, where, for $a = 0, 1$,

$$\widehat{\mu}_C^{\text{GEE-g}}(a) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{M_i} \sum_{j:S_{ij}=1} g^{-1}\left(\widehat{\beta}_0 + \widehat{\beta}_A a + \widehat{\beta}_L^\top L_{ij}\right),$$

$$\widehat{\mu}_I^{\text{GEE-g}}(a) = \frac{1}{\sum_{i=1}^{m} N_i} \sum_{i=1}^{m} \frac{N_i}{M_i} \sum_{j:S_{ij}=1} g^{-1}\left(\widehat{\beta}_0 + \widehat{\beta}_A a + \widehat{\beta}_L^\top L_{ij}\right).$$

Here, the g-computation step refers to the average of model predictions for each cluster after setting the treatment assignment to be a, and has been commonly used as a population standardization technique for estimating marginal estimands in observational studies (Rosenbaum 1987). To target each specific estimand, the weighting for each cluster is taken to be $1/m$ for estimating $\mu_c(a)$ (which gives equal weight to each cluster) and $N_i / \sum_{i=1}^{m} N_i$ for estimating $\mu_I(a)$ (which gives equal weight to each individual). The variance of $\widehat{\Delta}_C^{\text{GEE-g}}$ and $\widehat{\Delta}_I^{\text{GEE-g}}$ can be consistently estimated by the robust sandwich variance estimator after applying the delta method, denoted by $\widehat{V}_C^{\text{GEE-g}}$ and $\widehat{V}_I^{\text{GEE-g}}$, respectively. Their explicit expressions are developed in the supplementary material.

To proceed, we additionally make the following assumption, which implies that the observed cluster size within each arm is only subject to exogenous randomness.

Assumption 4 (Arm-specific random sampling). For $a \in \{0, 1\}$, the potential observed cluster size $M(a) = h_a^*(\epsilon_a)$ for some unknown function h_a^* and exogenous random noise ϵ_a that is independent of $\{N, C, \mathrm{X}, Y(a)\}$. Furthermore, for each possible N-dimensional binary vector s with $M(a)$ ones, $pr\{S(a) = s \mid Y(a), M(a), \mathrm{X}, N, C\} = \binom{N}{M(a)}^{-1}$.

Assumption 4 is a special case of Assumption 3 by substituting $h_a^*(\epsilon_a)$ for $h_a(N, C, \epsilon_a)$, and is plausible when the number of individuals observed in each cluster can at most depend on the cluster treatment status, regardless of other cluster-level characteristics. As one example, $M(1) = M(0)$ may follow a truncated Poisson distribution with mean 50 and taking values within the range of $[5, 200]$, in which case the observed cluster size is completely random; such a within-cluster random sampling assumption has been previously made for sample size and power calculations in cluster-randomized experiments (Shi and Lee 2018). As another example, $M(a) \sim \text{Poisson}(50 + 20a)$ indicates an increased recruitment effort in the treated clusters. Intuitively, Assumption 4 is needed because GEE does not model the sampling procedure and may cause bias when sampling is related to covariates as in Assumption 3. Assumption 4 resolves this bias by making the sampling procedure ignorable within each treatment group. The requirement for Assumption 4 indicates an inherent limitation of standard GEE in handling within-cluster sampling. In Remark 1, we provide special cases where Assumption 4 can be relaxed to Assumption 3 without comprising robustness.

Under arm-specific random sampling, Theorem 1 articulates several model specifications of GEE such that $\widehat{\Delta}_I^{\text{GEE-g}}$ and $\widehat{\Delta}_C^{\text{GEE-g}}$ are consistent and asymptotically normal, leading to valid statistical inference for our estimands defined under the potential outcomes framework. The regularity conditions required are moment and continuity assumptions similar to those invoked in Theorem 5.31 of van der Vaart (1998) for M-estimators. Particularly, model specifications (S2)–(S4) in Theorem 1 make no assumption about the underlying distribution for the potential outcomes, and hence allow for model-robust estimation.

Theorem 1. (a) Under Assumptions 1, 2, 4 and regularity conditions in the supplementary material, $\left(\widehat{V}_C^{\text{GEE-g}}\right)^{-1/2} (\widehat{\Delta}_C^{\text{GEE-g}} - \Delta_C) \xrightarrow{d} \mathcal{N}(0, 1)$ if any of the following conditions holds: (S1) the mean model $g\{E(Y_{ij} \mid U_{ij})\} = U_{ij}^\top \beta$ is correctly specified; (S2) an independence working correlation structure is used; (S3) g is the identity link function and the working variance is constant with $v(Y_{ij} \mid U_{ij}) = \sigma^2$; (S4) L_{ij} does not vary within each cluster and thus is only a function of cluster-level covariates (N_i, C_i). (b) If the estimating equations (1) are additionally weighted by the source population size N_i for each cluster i, then $\left(\widehat{V}_I^{\text{GEE-g}}\right)^{-1/2} (\widehat{\Delta}_I^{\text{GEE-g}} - \Delta_I) \xrightarrow{d} \mathcal{N}(0, 1)$ if any of (S1)–(S4) holds.

For model-robust inference via GEE, (S1) is trivial as a correctly specified mean model is expected to yield a valid causal effect estimator. Without a correct mean model, (S2) indicates that an independence working correlation structure admits valid causal inference, and may differ from the classic recommendation that the intracluster correlations should be modeled for analyzing cluster-randomized experiments (Murray 1998; Donner and Klar 2000). (S3) is the default specification for analyzing continuous outcomes and ensures model-robustness by proceeding with an ordinary least squares estimator (Su and Ding 2021). Finally, (S4) provides a more elaborate mean model specification adjusting for only cluster-level covariates. When none of (S2)–(S4) holds, for example, logistic GEE with an exchangeable working correlation adjusting for individual-level covariates, the resulting weighted g-computation approach is not guaranteed to be consistent for our causal estimands.

Remark 1. In Theorem 1, Assumption 4 can be replaced by Assumption 3 under (S1), or under (S2)–(S4) if each cluster is further weighted by $\{1 + (M_i - 1)\widehat{\rho}\}/M_i$ in solving the estimating equations (1). Under (S2)–(S4), this weighting serves two purposes: the numerator $\{1 + (M_i - 1)\widehat{\rho}\}$ removes the undesired weighting from the exchangeable working correlation, whereas the denominator offsets the observed cluster size effect. With such weights, the weighted g-computation approach is asymptotically equivalent to that based on an independence working correlation structure, thereby trading the estimation of intracluster correlation for model robustness.

Remark 2. Assumption 2 does not hold when stratified cluster randomization or biased-coin cluster randomization is used. However, the consistency and asymptotic normality results in Theorem 1 (and also Theorem 2) still hold, with the only difference being that the variance estimators $\widehat{V}_C^{\text{GEE-g}}$ and $\widehat{V}_I^{\text{GEE-g}}$ may be conservative, in the sense that the asymptotic normal distribution has a variance smaller than 1. This result, as well as the variance difference, was provided in Theorem 1 of Wang et al. (2023). An important special case where the variance estimators are consistent is $\pi = 0.5, f(x,y) = x - y$, and the strata variables are adjusted for as covariates. Beyond these two randomization schemes, similar asymptotic results can also be obtained for other covariate-adaptive randomization schemes following the theory of Bugni, Canay, and Shaikh (2018), Jiang et al. (2022), and Rafi (2023).

3.2. (Generalized) Linear Mixed Models

Generalized linear mixed models are another popular method for analyzing cluster-randomized experiments. If we write b_i as the random effect for cluster i, a typical generalized linear mixed model applied to cluster-randomized experiment often includes the following assumptions: (i) $b_1, \ldots, b_m$ are independent, identically distributed from $\mathcal{N}(0, \tau^2)$, where τ^2 is an unknown variance component; (ii) $Y_{i1}, \ldots, Y_{iN_i}$ are conditionally independent given $(U_{i1}, \ldots, U_{iN_i}, b_i)$; and (iii) the distribution $Y_{ij} \mid (U_{i1}, \ldots, U_{iN_i}, b_i)$ is a member of the exponential family with $E(Y_{ij} \mid U_{i1}, \ldots, U_{iN_i}, b_i) = E(Y_{ij} \mid U_{ij}, b_i) = g^{-1}(U_{ij}^\top \alpha + b_i)$, where g is the canonical link and $\alpha = (\alpha_0, \alpha_A, \alpha_L^\top)^\top$. Given the above specifications, we estimate model parameters by maximizing the likelihood function, and a common practice is to consider α_A as the treatment effect parameter. However, the interpretation of α is conditional on random effects, and if the mixed model is misspecified, α lacks a direct connection to our marginal estimands, with an important exception that we detail below.

An interesting special case where the mixed model provides model-robust inference in cluster-randomized experiments is when a linear mixed model is considered as the working model. In this case, we can define the weighted g-computation estimators as $\widehat{\Delta}_C^{\text{LMM-g}} = f\left\{\widehat{\mu}_C^{\text{LMM-g}}(1), \widehat{\mu}_C^{\text{LMM-g}}(0)\right\}$ and $\widehat{\Delta}_I^{\text{LMM-g}} = f\left\{\widehat{\mu}_I^{\text{LMM-g}}(1), \widehat{\mu}_I^{\text{LMM-g}}(0)\right\}$, where

$$\widehat{\mu}_C^{\text{LMM-g}}(a) = \widehat{\alpha}_0 + \widehat{\alpha}_A a + \widehat{\alpha}_L^\top \frac{\sum_{i=1}^m \overline{L}_i^o}{m},$$

$$\widehat{\mu}_I^{\text{LMM-g}}(a) = \widehat{\alpha}_0 + \widehat{\alpha}_A a + \widehat{\alpha}_L^\top \frac{\sum_{i=1}^m N_i \overline{L}_i^o}{\sum_{i=1}^m N_i}$$

with $\overline{L}_i^o = M_i^{-1} \sum_{j=1}^{N_i} S_{ij} L_{ij}$ being the average covariate value for cluster i among the observed individuals. These two weighted g-computation estimators are constructed in a similar fashion to those in Section 3.1. Furthermore, when the interest lies in the difference estimands with $f(x,y) = x - y$, we simply have $\widehat{\Delta}_C^{\text{LMM-g}} = \widehat{\alpha}_A$ and $\widehat{\Delta}_I^{\text{LMM-g}} = \widehat{\alpha}_A$ such that the treatment effect estimator from the linear mixed model can be taken as the average causal effect estimator, but these two coefficients are estimated via different weights applied to the log-likelihood as we explain in Theorem 2. For $\widehat{\Delta}_C^{\text{LMM-g}}$ and $\widehat{\Delta}_I^{\text{LMM-g}}$, we construct sandwich variance estimators $\widehat{V}_C^{\text{LMM-g}}$ and $\widehat{V}_I^{\text{LMM-g}}$ with their expressions given in the supplementary material.

Theorem 2 shows that $\widehat{\Delta}_C^{\text{LMM-g}}$ and $\widehat{\Delta}_I^{\text{LMM-g}}$ are asymptotically valid if the observed cluster size is only subject to exogenous variation within each arm, even when the linear mixed working model is arbitrarily misspecified.

Theorem 2. Under Assumptions 1, 2, 4 and regularity conditions provided in the supplementary material, $\left(\widehat{V}_C^{\text{LMM-g}}\right)^{-1/2} (\widehat{\Delta}_C^{\text{LMM-g}} - \Delta_C) \xrightarrow{d} \mathcal{N}(0,1)$. If each cluster is weighted by N_i in the log-likelihood function of the working linear mixed model, then $\left(\widehat{V}_I^{\text{LMM-g}}\right)^{-1/2} (\widehat{\Delta}_I^{\text{LMM-g}} - \Delta_I) \xrightarrow{d} \mathcal{N}(0,1)$.

Theorem 2 is the counterpart of Theorem 1 for linear mixed models. In the special case that $M_i(1) = M_i(0)$, $N_i = n$ is constant, $f(x,y) = x - y$, and Y_{ij} are marginally identically distributed, Theorem 2 reduces to Theorem 1(a) of Wang et al. (2021). Compared to GEE with an identity link function and working exchangeable correlation, linear mixed models yield similar estimating equations, but the estimators for nuisance parameters are different, that is, moment estimation for ρ in GEE versus maximum likelihood estimation for σ^2, τ^2 in linear mixed models, leading to slightly different asymptotic variances. For noncontinuous outcomes, although linear mixed models can still provide valid inference, they are likely misspecified due to the Normal assumptions on random effects and noises. In this sense, GEE could be more flexible regarding the choice of link function and variance function, and hence more natural for handling noncontinuous outcomes.

Both GEE and linear mixed models provide means to adjust for covariates, which may improve the precision over the unadjusted analysis. However, without further restrictions on the data-generating process, neither method is guaranteed to be equally or more efficient than an unadjusted analysis in general. For instance, Wang et al. (2021) has constructed an example (Scenario 1 of Section 4) showing that adjusting for covariates can even decrease the precision of $\widehat{\alpha}_A$ based on linear mixed models. With an independence correlation structure, Su and Ding (2021) showed that GEE with identity link may also lose precision by covariate adjustment. These inherent limitations of model-based methods serve as a strong motivation for deriving more principled, and statistically efficient estimators that can maximize the precision gain from covariate adjustment in cluster-randomized experiments.

Finally, while GEE and generalized linear mixed models are the most widely used approach for analyzing cluster-randomized experiments, there are two alternative estimators, the augmented GEE (Stephens, Tchetgen Tchetgen, and Gruttola 2012) and targeted maximum likelihood estimation (Benitez et al. 2021; Balzer et al. 2023) that can also provide robust estimation under Assumptions 1, 2, and 4 when certain aspects of working models are misspecified. We provide discussions of those approaches in the supplementary material.

4. Efficient Covariate Adjustment

4.1. Efficient Influence Functions

For model-based covariate adjustment to provide model-robust inference, Theorems 1 and 2 largely require the observed cluster size to be independent of cluster characteristics (N, C). This is a rather strong assumption and may be violated if the observed cluster size is proportional to the source population size (a common scenario in healthcare research as hospital volume is often associated with patient recruitment results) or if the observed cluster size depends on geographical location or other cluster characteristics. Additionally, the structure of the GEE and the linear mixed model estimators can limit their ability to leverage baseline covariates for maximum precision gain in cluster-randomized experiments, especially when the parametric modeling assumptions are incorrect. To address such limitations, we develop more principled estimators to simultaneously optimize covariate adjustment and accommodate variable cluster sizes arising from post-randomization cluster-dependent sampling schemes. The proposed estimators are motivated by the efficient influence function, which is a nonparametric functional of observed data that characterizes the target estimand (Hines et al. 2022). With the efficient influence function, one can derive the semiparametric efficiency lower bound for an estimand. That is, the asymptotic variance of all regular and asymptotically linear estimators over the underlying causal model is lower bounded by the variance of the efficient influence function. More importantly, recent advances in causal inference (e.g., van der Laan and Rose 2011; Chernozhukov et al. 2018) showed how to use the efficient influence functions to construct an efficient estimator, that is, achieving the semiparametric efficiency bound, by incorporating machine learning algorithms. Theorem 3 provides the efficient influence functions for $\mu_C(a)$ and $\mu_I(a)$, from which the efficient influence functions for Δ_C and Δ_I can be obtained by the chain rule.

Theorem 3. (a) Given Assumptions 1–3, the efficient influence function for $\mu_C(a)$ is

$$\mathrm{EIF}_C(a) = \frac{I\{A = a\}}{\pi^a(1-\pi)^{1-a}} \left\{ \overline{Y}^o - E\left(\overline{Y}^o \mid A = a, \mathrm{X}^o, M, N, C\right) \right\} + \frac{pr(A = a \mid M, N, C)}{\pi^a(1-\pi)^{1-a}} \left\{ \begin{array}{c} E\left(\overline{Y}^o \mid A = a, \mathrm{X}^o, M, N, C\right) \\ -E\left(\overline{Y}^o \mid A = a, N, C\right) \end{array} \right\} + E\left(\overline{Y}^o \mid A = a, N, C\right) - \mu_C(a), \quad a \in \{0, 1\},$$

where $I\{A = a\}$ is an indicator function of $A = a$ and $\overline{Y}^o = M^{-1}\sum_{j=1}^N S_{\cdot j} Y_{\cdot j}$ refers to the average outcome among the observed participants in each cluster. Here, $S_{\cdot j} = AS_{\cdot j}(1) + (1-A)S_{\cdot j}(0)$ and $Y_{\cdot j} = AY_{\cdot j}(1) + (1-A)Y_{\cdot j}(0)$, where $S_{\cdot j}(a)$ and $Y_{\cdot j}(a)$ are the jth element of $S(a), Y(a)$ defined in $\mathcal{P}^W$, respectively. (b) Given Assumptions 1–3, the efficient influence function for $\mu_I(a)$ is

$$\mathrm{EIF}_I(a) = \frac{N}{E(N)} \left\{ \mathrm{EIF}_C(a) + \mu_C(a) - \mu_I(a) \right\}, \quad a \in \{0, 1\}.$$

The efficient influence functions in Theorem 3 involve three nuisance functions, which we denote by $\eta_a^* = E\left(\overline{Y}^o \mid A = a, \mathrm{X}^o, M, N, C\right)$, $\zeta_a^* = E\left(\overline{Y}^o \mid A = a, N, C\right)$, and $\kappa_a^* = pr(A = a \mid M, N, C)$. Compared to $\mathrm{EIF}_C(a)$, $\mathrm{EIF}_I(a)$ additionally includes a weight, $N/E(N)$, to target the individual-average treatment effect. When M is identical to the source population size N, the number of sampled individuals can be treated as a pre-randomization variable, and the efficient influence function for $\mu_C(a)$ reduces to that in Balzer et al. (2019), where the only nuisance function is η_a^*. In more general settings as we consider here, two additional nuisance functions, ζ_a^* and κ_a^*, are needed to leverage X^o, which involves post-randomization information S, for addressing bias and improving efficiency. Next, we construct new estimators for Δ_C and Δ_I based on the derived efficient influence functions.

4.2. Efficient Estimators based on the Efficient Influence Functions

Based on Theorem 3, we propose the following estimator for $\mu_C(a)$ and $\mu_I(a)$:

$$\widehat{\mu}_C^{\mathrm{Eff}}(a) = \frac{1}{m}\sum_{i=1}^m D_i(a, \widehat{h}_a),$$

$$\widehat{\mu}_I^{\mathrm{Eff}}(a) = \frac{1}{\sum_{i=1}^m N_i}\sum_{i=1}^m N_i D_i(a, \widehat{h}_a), \tag{2}$$

where $\widehat{h}_a = (\widehat{\eta}_a, \widehat{\zeta}_a, \widehat{\kappa}_a)$ are user-specified estimators for nuisance functions $h_a^* = (\eta_a^*, \zeta_a^*, \kappa_a^*)$ and

$$D_i(a, \widehat{h}_a) = \frac{I\{A_i = a\}}{\pi^a(1-\pi)^{1-a}} \left\{ \overline{Y}_i^o - \widehat{\eta}_a(\mathrm{X}_i^o, M_i, N_i, C_i) \right\} + \frac{\widehat{\kappa}_a(M_i, N_i, C_i)}{\pi^a(1-\pi)^{1-a}} \left\{ \begin{array}{c} \widehat{\eta}_a(\mathrm{X}_i^o, M_i, N_i, C_i) \\ -\widehat{\zeta}_a(N_i, C_i) \end{array} \right\} + \widehat{\zeta}_a(N_i, C_i).$$

Then the target estimands defined in Section 2 are estimated by $\widehat{\Delta}_C^{\mathrm{Eff}} = f\left\{\widehat{\mu}_C^{\mathrm{Eff}}(1), \widehat{\mu}_C^{\mathrm{Eff}}(0)\right\}$ and $\widehat{\Delta}_I^{\mathrm{Eff}} = f\left\{\widehat{\mu}_I^{\mathrm{Eff}}(1), \widehat{\mu}_I^{\mathrm{Eff}}(0)\right\}$. Among the many possibilities for estimating these nuisance functions, we primarily consider the following two approaches.

The first approach considers parametric working models, that are, $\widehat{\eta}_a = \eta_a(\widehat{\theta}_{\eta,a}), \widehat{\zeta}_a = \zeta_a(\widehat{\theta}_{\zeta,a}), \widehat{\kappa}_a = \kappa_a(\widehat{\theta}_{\kappa,a})$ for pre-specified functions $\eta_a, \zeta_a, \kappa_a$ with finite-dimensional parameters $\theta_{\eta,a}, \theta_{\zeta,a}, \theta_{\kappa,a}$. In this case, we use superscript "Eff-PM", for example, $\widehat{\Delta}_C^{\text{Eff-PM}}$, to highlight the role of parametric models in estimation. Example working models include GEE, generalized linear mixed models, penalized regression with variable selection, among others; in all cases, the associated parameters are estimated by solving estimating equations, and we assume that the estimating equations satisfy regularity conditions provided in the supplementary material such that the nuisance parameter estimators are asymptotically linear. With this method, we can compute the sandwich variance estimators $\widehat{V}_C^{\text{Eff-PM}}$ for $\widehat{\Delta}_C^{\text{Eff-PM}}$ and $\widehat{V}_I^{\text{Eff-PM}}$ for $\widehat{\Delta}_I^{\text{Eff-PM}}$, which are also given in the supplementary material. For the subsequent technical discussions, we denote the probability limit of $(\widehat{\theta}_{\eta,a}, \widehat{\theta}_{\zeta,a}, \widehat{\theta}_{\kappa,a})$ as $(\underline{\theta}_{\eta,a}, \underline{\theta}_{\zeta,a}, \underline{\theta}_{\kappa,a})$.

The second approach exploits machine learning algorithms with cross-fitting to estimate all nuisance functions. For this case, we use superscript "Eff-ML", for example, $\widehat{\Delta}_C^{\text{Eff-ML}}$, to indicate the use of machine learning algorithms. We assume that each nuisance function estimator is consistent to the truth at an $m^{1/4}$ rate such that $\widehat{h}_a - h_a^* = o_p(m^{-1/4})$ in $L_2(\mathcal{P})$-norm. This $m^{1/4}$ rate can be achieved by many methods such as random forests (Wager and Walther 2015), neural networks (Farrell, Liang, and Misra 2021), and boosting (Luo and Spindler 2016); a further discussion on this rate is provided in the supplementary material. In addition, we assume a regularity condition that $\widehat{\kappa}_a$ and $E\{(\eta_a^*)^2 \mid M, N, C\}$ are uniformly bounded; a similar condition is invoked in Chernozhukov et al. (2018) for controlling the remainder term and consistently estimating the variance. In the cross-fitting step, we randomly partition m clusters into K parts with roughly equal sizes (the size difference is at most 1), and denote $\mathcal{O}_k^*$ as the kth part and $\mathcal{O}_{-k}^* = \bigcup_{k' \in \{1,\dots,K\} \setminus \{k\}} \mathcal{O}_{k'}^*$. For each k, we compute $\widehat{h}_{a,k}$, which is the nuisance function trained on $\mathcal{O}_{-k}^*$ and evaluated at $\mathcal{O}_k^*$, and the estimated nuisance function $\widehat{h}_a$ evaluated on all clusters is then the combination of $\widehat{h}_{a,k}$ for all $k = 1, \dots, K$. We then plug in $\widehat{h}_a$ to (2) to compute the estimators for Δ_C and Δ_I. In practice, we recommend choosing K such that $m/K \geq 10$. For variance estimation, we propose the following consistent estimators based on the efficient influence function and cross-fitting:

$$\widehat{V}_C^{\text{Eff-ML}} = \frac{1}{m^2} \sum_{k=1}^{K} \sum_{i \in \mathcal{O}_k} \left[\sum_{a=0}^{1} \dot{f}_{a,C} \left\{ D_i(a, \widehat{h}_{a,k}) - \frac{1}{|\mathcal{O}_k|} \sum_{l \in \mathcal{O}_k} D_l(a, \widehat{h}_{a,k}) \right\} \right]^2,$$

$$\widehat{V}_I^{\text{Eff-ML}} = \frac{1}{(\sum_{i=1}^{m} N_i)^2} \sum_{k=1}^{K} \sum_{i \in \mathcal{O}_k} \left[\sum_{a=0}^{1} \dot{f}_{a,I} \left\{ N_i D_i(a, \widehat{h}_{a,k}) - \frac{1}{|\mathcal{O}_k|} \sum_{l \in \mathcal{O}_k} N_l D_l(\widehat{h}_{a,k}) \right\} \right]^2,$$

where $\dot{f}_{a,C}$ is the partial derivative of f on $\mu_C(a)$ at $\{\widehat{\mu}_C^{\text{Eff-ML}}(1), \widehat{\mu}_C^{\text{Eff-ML}}(0)\}$, $\dot{f}_{a,I}$ is the partial derivative of f on $\mu_I(a)$ at $\{\widehat{\mu}_I^{\text{Eff-ML}}(1), \widehat{\mu}_I^{\text{Eff-ML}}(0)\}$, and $|\mathcal{O}_k|$ is the size of $\mathcal{O}_k$.

Theorem 4 summarizes the asymptotic behaviors of our proposed estimators under both strategies for estimating the nuisance functions.

Theorem 4. Given Assumptions 1–3 and above regularity conditions,

(a) if $\kappa_a(\underline{\theta}_{\kappa,a}) = \kappa_a^*$ or $E\{\eta_a(\underline{\theta}_{\eta,a}) \mid M, N, C\} = \zeta_a(\underline{\theta}_{\zeta,a})$, then $\left(\widehat{V}_C^{\text{Eff-PM}}\right)^{-1/2} (\widehat{\Delta}_C^{\text{Eff-PM}} - \Delta_C) \xrightarrow{d} \mathcal{N}(0,1)$ and $\left(\widehat{V}_I^{\text{Eff-PM}}\right)^{-1/2} (\widehat{\Delta}_I^{\text{Eff-PM}} - \Delta_I) \xrightarrow{d} \mathcal{N}(0,1)$;

(b) if $\widehat{h}_a - h_a^* = o_p(m^{-1/4})$ in $L_2(\mathcal{P})$-norm, then $\left(\widehat{V}_C^{\text{Eff-ML}}\right)^{-1/2} (\widehat{\Delta}_C^{\text{Eff-ML}} - \Delta_C) \xrightarrow{d} \mathcal{N}(0,1)$ and $\left(\widehat{V}_I^{\text{Eff-ML}}\right)^{-1/2} (\widehat{\Delta}_I^{\text{Eff-ML}} - \Delta_I) \xrightarrow{d} \mathcal{N}(0,1)$. Furthermore, $\widehat{V}_C^{\text{Eff-ML}}$ and $\widehat{V}_I^{\text{Eff-ML}}$ converge in probability to the semiparametric efficiency lower bounds of Δ_C and Δ_I, respectively.

For the proposed estimators with parametric working models, Theorem 4(a) implies that the estimator is consistent if κ_a is correctly specified, or the working models η_a and ζ_a are compatible conditioning on M, N, and C. In the special case that all cluster-level covariates are discrete, the latter condition can be satisfied by setting

$$\eta_a(\underline{\theta}_{\eta,a}) = \zeta_a + \underline{\beta}^\top \left(\overline{X}^o - \sum_n \sum_c I\{N = n, C = c\} \underline{\theta}_{n,c} \right),$$

where $\underline{\beta}$ is an arbitrary p-dimensional vector and $\underline{\theta}_{n,c} = E\{\overline{X}^o \mid N = n, C = c\}$; in this special case, the resulting estimators are in fact robust to arbitrary working model misspecification. In more general cases with non-discrete cluster-level covariates, the proposed estimators are at least triply-robust. That is, they are consistent to their respective target estimands as long as two out of the three nuisance functions are correctly modeled, since $E(\eta_a^* \mid M, N, C) = \zeta_a^*$ as proved in the supplementary material.

For the proposed estimators using machine learning algorithms, efficiency can be achieved when all nuisance parameters are consistently estimated, leading to higher asymptotic precision than potentially misspecified parametric working models. For modeling η_a^*, since X^o is a matrix and its dimension may change across clusters, a feasible practical strategy is to fit Y_{ij} on (X_{ij}, M_i, N_i, C_i) and pre-specified summary statistics of X_i^o with fixed dimensions, for example, $\overline{X}_i^o$, and then compute the cluster-average of predictions as the model fit. Alternatively, one can directly model $\overline{Y}_i^o$ on (M_i, N_i, C_i) and functions of X_i^o, which can be potentially high-dimensional due to potentially higher-order associations between X_i^o and $\overline{Y}_i^o$.

To summarize the precision comparison among all considered estimators (referred to as GEE-g, LMM-g, Eff-PM, and Eff-ML for brevity), we first consider consistent estimators under Assumptions 1–3, which include GEE-g (under the specific conditions given by Remark 1), Eff-PM, and Eff-ML. Implied by its efficiency property in Theorem 4, Eff-ML has an equal or smaller asymptotic variance compared to the other two estimators. Furthermore, since GEE-g and Eff-PM both include the special case of no covariate adjustment, covariate adjustment via Eff-ML then guarantees no asymptotic precision loss. However, the variance comparison between Eff-PM and GEE-g, with or without covariate adjustment, is generally indeterminate, as it can depend on the choice of working nuisance models. When we further consider potentially inconsistent estimators under Assumptions 1–3, for example, LMM-g that additionally requires Assumption 4 for consistency, Eff-ML may have a larger asymptotic variance than LMM-g under a model satisfying Assumptions 1–4. This is because estimators can be designed to achieve a smaller asymptotic variance within a stricter model space, without a guarantee of consistency under a larger model space.

Remark 3. In practice, each of the two approaches for nuisance function estimation has its pros and cons. The machine learning methods are asymptotically more precise, but the efficiency gain typically requires at least a moderate number of clusters. In addition, although the cross-fitting procedure yields the desired convergence property, it may be prone to finite-sample bias especially when the sample size is limited (Hines et al. 2022). This finite-sample bias can be potentially alleviated by using parsimonious parametric working models, thereby trading asymptotic precision for better finite-sample performance characteristics. In fact, whether machine learning algorithms outperform parametric methods in a finite-sample setting may deserve a case-by-case evaluation, and may well depend on the sample size, true data-generating distribution, selection of tuning parameters, among others. Generally, we recommend implementing machine learning algorithms when the number of clusters is sufficiently large, for example, $m = 100$, in order to accurately capture aspects of the true data-generating distribution and hence achieve full precision gain. With a small number of clusters, for example, $m = 20$, we caution against using complex working nuisance models, and instead recommend parsimonious parametric nuisance models to avoid over-fitting and ensure numerical stability.

Remark 4. If the observed cluster size $M < N$, our proposed estimators require an accurate estimate of the source population size for efficient estimation. If N is not fully available for all clusters, one can still use the observed data $(\overline{Y}_i^o, A_i, C_i)$ to infer Δ_C, and the efficient influence function for $\mu_C(a)$ becomes

$$\frac{I\{A = a\}}{\pi^a(1-\pi)^{1-a}}(\overline{Y}^o - \zeta_a^*) + \zeta_a^* - \mu_C(a),$$

based on which an estimator for Δ_C can be constructed. For example, we can apply our proposed estimator with $\widehat{\eta}_a$ set to be equal to $\widehat{\zeta}_a$ and get a consistent estimator for Δ_C even when $\widehat{\zeta}_a$ is incorrectly specified. The individual-average treatment effect Δ_I, however, is generally not identifiable without observing N.

Remark 5. Under stratified cluster randomization or biased-coin cluster randomization, Theorem 4(a) still holds except that the variance estimators may be conservative, following the same argument as in Remark 2. In addition, Theorem 4(b) remains unchanged under these two designs with the same assumption on the nuisance function estimators. In particular, the cross-fitting procedure should be modified to achieve treatment balance within each stratum of each fold; see Assumptions 4.2 and 4.3 of Rafi (2023) for an example implementation.

5. Simulation Experiments

5.1. Simulation Design

We conducted two simulation experiments to compare different methods for analyzing cluster-randomized experiments. The first simulation study focused on estimating the difference estimands of the cluster-average and individual-average treatment effect for continuous outcomes, while the second study focused on the relative risk estimands for binary outcomes. In each experiment, we considered a relatively small ($m = 30$) or large ($m = 100$) number of clusters, random observed cluster sizes (i.e., M_i is independent of other variables as a special case of Assumption 4) or cluster-dependent observed cluster sizes (i.e., M_i depends on treatment and cluster-level covariates as in Assumption 3). Combinations of these specifications are labeled as scenarios 1–4 in the simulation results.

In the first simulation experiment, we let (N_i, C_{i1}, C_{i2}), $i = 1, \ldots, m$ 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 $\{10, 50\}$, $\mathcal{P}^{C_1|N} = \mathcal{N}(N/10, 4)$, and $\mathcal{P}^{C_2|N,C_1} = \mathcal{B}[\text{expit}\{\log(N/10)C_1\}]$ is a Bernoulli distribution with $\text{expit}(x) = (1 + e^{-x})^{-1}$. Next, for each individual in the source population, we generated the individual-level covariates from $X_{ij1} \sim \mathcal{B}(N_i/50)$, $X_{ij2} \sim \mathcal{N}\left\{\sum_{j=1}^{N_i} X_{ij1}(2C_{i2} - 1)/N_i, 9\right\}$, and potential outcomes from $Y_{ij}(1) \sim \mathcal{N}\left\{N_i/5 + N_i \sin(C_{i1})(2C_{i2} - 1)/30 + 5e^{X_{ij1}} \mid X_{ij2} \mid, 1\right\}$ and $Y_{ij}(0) \sim \mathcal{N}\{\gamma_i + N_i \sin(C_{i1})(2C_{i2} - 1)/30 + 5e^{X_{ij1}} \mid X_{ij2} \mid, 1\right\}$ where $\gamma_i \sim \mathcal{N}(0, 1)$ is a cluster-level random intercept to allow for a positive residual intracluster correlation. We set $M_i(1) = M_i(0) = 9 + \mathcal{B}(0.5)$ for the random observed cluster size scenario, and $M_i(1) = N_i/5 + 5C_{i2}$, $M_i(0) = 3I\{N_i = 50\} + 3$ for the cluster-dependent observed cluster size scenario. Then, we independently sampled $A_i \sim \mathcal{B}(0.5)$ and defined $Y_{ij} = A_i Y_{ij}(1) + (1 - A_i)Y_{ij}(0)$ and $M_i = A_i M_i(1) + (1 - A_i)M_i(0)$. Finally, for each cluster, we uniformly sampled without replacement M_i individuals, for whom $S_{ij} = 1$, and the observed data in each simulation replicate are $\{Y_{ij}, A_i, M_i, C_{i1}, C_{i2}, X_{ij1}, X_{ij2} : S_{ij} = 1, i = 1, \ldots, m, j = 1, \ldots, N_i\}$. For the second simulation study, the data were generated following the first simulation study, except that the potential outcomes were drawn from the following Bernoulli distributions: $Y_{ij}(1) \sim \mathcal{B}\left[\text{expit}\{-N_i/20 + N_i \sin(C_{i1})(2C_{i2} - 1)/30 + 1.5e^{X_{ij1}} \mid X_{ij2} \mid^{1/2}\right\}\right]$ and $Y_{ij}(0) \sim \mathcal{B}\left[\text{expit}\{\gamma_i + N_i \sin(C_{i1})(2C_{i2} - 1)/30 + 1.5(2X_{ij1} - 1) \mid X_{ij2} \mid^{1/2}\right\}\right]$.

We compared the following methods. The unadjusted method (Bugni et al. 2023) is equivalent to our proposed method setting $\widehat{\eta}_a = \widehat{\zeta}_a$ to be a constant. GEE with weighted g-computation was implemented as described in Section 3.1 with an exchangeable working correlation for continuous outcomes and an independence working correlation for binary outcomes. Linear mixed models with weighted g-computation were implimented as described in Section 3.2 for both continuous and binary outcomes. For each model-based method, all baseline covariates are adjusted for as linear terms, which misspecify the true data-generating distribution. However, the GEE estimator satisfies conditions (S2) for the first simulation and (S3) for the second simulation in Theorem 1, yielding valid asymptotic inference under random observed cluster sizes. Likewise, the working linear mixed model is also misspecified, but the weighted g-computation estimator is consistent under random observed cluster sizes as implied by Theorem 2. For our proposed efficient estimators, we considered both parametric working models and machine learning algorithms to estimate the nuisance functions. The former used generalized linear models for estimating η_a and ζ_a, and we set a correct working model for κ_a such that the conditions in Theorem 4(a) are satisfied. The latter exploited the Super Learner (van der Laan, Polley, and Hubbard 2007) for model-fitting with generalized linear models, regression trees, and neural networks (to obtain consistent estimators for each nuisance function), and facilitates the validation of the results in Theorem 4(b). To summarize, GEE, linear mixed models, and $\widehat{\eta}_a$ adjusted for covariates $(N_i, C_{i1}, C_{i2}, X_{ij1}, X_{ij2})$, whereas $\widehat{\zeta}_a$ adjusted for cluster-level covariates (N_i, C_{i1}, C_{i2}).

For each approach, we used the proposed variance estimator and applied a degrees-of-freedom adjustment, that is, multiplying the variance estimator by a factor $m/(m-5)$, where 5 is the number of adjusted baseline covariates. This adjustment was motivated by a common technique used in the regression context for small-sample bias correction (MacKinnon and White 1985), and we adapted it to potentially improve the finite-sample coverage probability of each variance estimator. In addition, we considered the t-distribution with $m-5$ degrees of freedom, instead of a Normal distribution, to better approximate the distribution of the standardized covariate-adjusted estimator in finite samples. This choice improved the coverage probability (closer to nominal level) by 1%–3% for $m = 30$ and less than 1% for $m = 100$. For each scenario of both simulation experiments, we randomly generated 10,000 datasets and tested the above methods on each dataset. We focused on the following metrics for comparison: bias, empirical standard error (ESE) from the Monte Carlo replications, average of standard error estimates (ASE), and empirical coverage probability (CP) of the 95% confidence interval.

5.2. Simulation Results

Table 1 summarizes the simulation results for continuous outcomes. Since the outcome distribution varies by the source population size, the cluster-average treatment effect $\Delta_C = 6$, which differs from the individual-average treatment effect $\Delta_I = 8.67$. Across all scenarios, the proposed methods with no covariate adjustment, parametric working models, or machine learning algorithms have negligible bias and nominal coverage, while the model-based methods, that is, GEE and linear mixed models, show bias and undercoverage if the observed cluster sizes are generated under the cluster-dependent sampling scheme.

For scenarios 1 and 3, since the observed cluster size is completely random, the model-based estimators perform well, which confirms the theoretical results in Section 3. Among all methods, our method with machine learning algorithms has the highest precision: its variance is 40%–93% and 49%–90% smaller than the other methods for estimating the Δ_C and Δ_I, respectively, demonstrating its potential to flexibly leverage baseline covariates for improving study power. Our estimator that uses parametric models for covariate adjustment has comparable precision to model-based methods and is more efficient than the unadjusted estimator. For scenarios 2 and 4, more individuals are enrolled in treated clusters with a larger source population, leading to bias for methods that used individual-level data without adjusting for this cluster-dependent sampling scheme. Specifically, model-based methods have bias ranging from 0.66 to 1.90, and 1%–13% undercoverage. In contrast, our proposed methods show both validity and precision. The validity is reflected by the negligible bias and nominal coverage, and the precision is borne out by their smaller empirical variance than the unadjusted estimator.

Table S1 in the supplementary material summarizes the simulation results for binary outcomes, and the patterns are generally similar to those for continuous outcomes. In particular, the proposed methods remain valid across scenarios, but the advantage of machine learning algorithms over parametric modeling is less obvious. Finally, comparing across sample size configurations, all methods have less stable performance with a smaller number

Table 1. Results in the first simulation experiment with continuous outcomes.

		Cluster-average treatment effect $\Delta_C = 6$				Individual-average treatment effect $\Delta_I = 8.67$			
Setting	Method	Bias	ESE	ASE	CP	Bias	ESE	ASE	CP
Scenario 1: Small m with random observed cluster sizes	Unadjusted	0.07	4.95	4.88	0.95	−0.09	4.63	4.29	0.93
	GEE-g	0.15	2.78	2.47	0.92	−0.07	3.69	3.13	0.90
	LMM-g	0.15	2.78	2.79	0.96	−0.07	3.68	2.95	0.90
	Eff-PM	−0.01	3.20	2.69	0.93	−0.06	3.96	3.40	0.91
	Eff-ML	0.03	2.16	2.04	0.95	0.01	2.63	2.48	0.95
Scenario 2: Small m with cluster-dependent observed cluster sizes	Unadjusted	−0.05	5.32	5.32	0.95	−0.12	4.79	4.48	0.93
	GEE-g	1.90	3.74	3.24	0.87	0.73	4.41	3.66	0.89
	LMM-g	1.66	3.60	3.68	0.94	0.66	4.28	3.81	0.93
	Eff-PM	−0.01	3.82	3.47	0.93	−0.05	4.30	3.96	0.93
	Eff-ML	−0.21	3.40	3.46	0.96	−0.17	4.17	3.83	0.93
Scenario 3: Large m with random observed cluster sizes	Unadjusted	0.01	2.67	2.65	0.95	−0.03	2.41	2.37	0.94
	GEE-g	0.05	1.42	1.38	0.95	−0.01	1.91	1.83	0.94
	LMM-g	0.05	1.42	1.42	0.95	−0.01	1.91	1.55	0.90
	Eff-PM	0.01	1.42	1.38	0.95	−0.01	1.93	1.92	0.95
	Eff-ML	0.04	0.70	0.71	0.95	0.01	0.78	0.81	0.96
Scenario 4: Large m with cluster-dependent observed cluster sizes	Unadjusted	0.05	2.91	2.89	0.95	0.01	2.53	2.48	0.94
	GEE-g	1.80	1.89	1.82	0.82	0.74	2.21	2.12	0.92
	LMM-g	1.72	1.87	1.89	0.86	0.73	2.20	2.00	0.91
	Eff-PM	−0.01	1.88	1.83	0.95	−0.01	2.20	2.16	0.94
	Eff-ML	−0.01	1.89	1.83	0.94	−0.01	2.20	2.13	0.94

Unadjusted: the unadjusted estimator. GEE-g: GEE with weighted g-computation. LMM-g: linear mixed models with weighted g-computation. Eff-PM: our proposed method with parametric working models. Eff-ML: our proposed method with machine learning algorithms. ESE: empirical standard error. ASE: average of estimated standard error. CP: coverage probability based on t-distribution.

of clusters. Specifically, when $m = 30$, methods with covariate adjustment tend to underestimate the true standard error, causing 0%–5% undercoverage. When m increases to 100, the estimated standard errors match the empirical standard error, thereby implying the validity of our variance estimator.

To test all methods with a higher degree of source population size heterogeneity, we repeated the first and second simulations with N uniformly distributed over integers in $[10, 100]$. Other changes required for this data-generating process are provided in the supplementary material. Tables S2 and S3 in the supplementary material summarize the simulation results, which are similar to the first two simulations, thereby demonstrating the capability of our methods in handling more heterogeneous source population sizes. Of note, an increasing source population size heterogeneity can affect the stability of machine learning algorithms if the number of clusters is small ($m = 30$), as reflected by less accurate standard error estimators. However, the bias in the standard error estimator is due to several outlier point estimates and variance estimates in specific simulation iterations, and therefore does not result in undercoverage. The bias of the standard error estimator vanished as the number of clusters increases to 100.

In the supplementary material, we provide additional simulation results for augmented GEE and targeted maximum likelihood estimation under the same settings. The targeted maximum likelihood estimator was unbiased when it only adjusted for cluster-level covariates, but it had bias for Δ_I if the sampling was cluster-dependent. In most scenarios, both methods were less precise than our proposed method coupled with machine learning estimators for the nuisance functions.

6. Data Applications

6.1. Three Cluster-Randomized Experiments

The Work, Family, and Health Study (WFHS) is a cluster-randomized experiment designed to reduce work-family conflict and improve the health and well-being of employees (Work, Family, and Health Study WFHS). Fifty-six study groups (clusters) were equally randomized to receive a workplace intervention (treatment) or not (control) with each cluster including 3–50 employees. The observed cluster size has mean 11.77 and standard error 7.47. We focused on the control over work hours outcome at the 6-month follow-up, a continuous outcome measure ranging from 1 to 5. We adjusted for the following covariates: cluster sizes and group job functions (core or supporting) at the cluster level, and baseline value of control over work hours and mental health score at the individual level.

The Pain Program for Active Coping and Training study (PPACT), supported by National Institute of Health Pragmatic Clinical Trials Collaboratory, is a pragmatic, cluster-randomized experiment evaluating the effectiveness of a care-based cognitive behavioral therapy intervention for treating long-term opioid users with chronic pain (DeBar et al. 2022). One-hundred-six clusters of primary care providers (clusters) were equally randomized to receive the intervention or usual care. Each cluster contained 1–10 participants enrolled via phone screening. The observed cluster size has mean 2.03 and standard error 4.9. The primary outcome was the PEGS (pain intensity and interference with enjoyment of life, general activity, and sleep) score at 12 months, a continuous scale assessing pain impact as a composite of pain intensity and interference. We adjusted for the cluster sizes and 12 individual-level baseline variables including age, gender, disability, smoking status, body mass index, alcohol abuse, drug abuse, comorbidity, depression, number of pain types, average morphine dose, and heavy opioid usage.

The Improving rational use of artemisinin combination therapies through diagnosis-dependent subsidies (ACTS) study is a cluster-randomized experiment completed in western Kenya (Prudhomme O'Meara et al. 2018). Thirty-two community clusters were randomized in a 1:1 ratio to two arms: malaria rapid diagnostic tests with vouchers of artemisinin combination therapies provided for positive test results (treatment) versus standard package (control). The primary outcome was an indicator of receiving a malaria diagnostic test among fevers in the past four weeks at 12 months. Using survey sampling, the observed cluster sizes ranged from 39 to 129, with mean 56.62 and standard error 16.09. As the source population size was unknown, we focused on inference of the observed population in ACTS. We adjusted for cluster sizes and existence of health facilities at the cluster level, and gender and wealth index at the individual level.

These three cluster-randomized experiments cover different contexts including social science, chronic pain treatment, and infectious disease control; they also feature different sample size configurations, outcome types, and number of covariates. By reanalyzing these datasets, we aim to illustrate our methods in multiple real-world settings. In addition, these three cluster-randomized experiments used three different randomization schemes. WFHS had a biased-coin cluster randomization design (Bray et al. 2013), aiming to balance group job functions, number of vice presidents, and location. PPACT adopted simple randomization as we considered in Assumption 2. ACTS had a stratified cluster randomization design based on six strata defined by subcounties and the existence of health facilities at the cluster level. Although Assumption 2 does not hold under the biased-coin or stratified cluster randomization, they do not affect the consistency of our considered estimators as pointed out in Remarks 2 and 5. To account for the variance reduction under these two restricted randomization designs, we adjusted for all available covariates balanced by the restricted randomization. In particular, since information on the number of vice presidents and location is not available from the WFHS dataset we analyzed, the sandwich variance estimator may be slightly conservative but still valid. For the purpose of demonstrating our theoretical results, we do not further study the variance differences across different randomization designs in the data applications and leave a more systematic study for future research.

6.2. Results of Data Analysis

For each dataset, we estimated the cluster-average and individual-average treatment effects on the difference scale, using the unadjusted estimator, GEE, linear mixed models, and our proposed method with parametric working models and machine learning algorithms. We first set $M = N$ for each study, corresponding to the source population analysis for WFHS and the enrolled

Table 2. Results of full data analyses.

		Cluster-average treatment effect			Individual-average treatment effect		
Study	Method	Estimate	95% C.I.	PVR	Estimate	95% C.I.	PVR
WFHS	Unadjusted	0.17	(0.00, 0.34)	–	0.17	(0.03, 0.31)	–
	GEE-g	0.21	(0.11, 0.31)	65%	0.22	(0.11, 0.32)	44%
	LMM-g	0.21	(0.10, 0.31)	62%	0.22	(0.11, 0.32)	44%
	Eff-PM	0.22	(0.11, 0.33)	58%	0.21	(0.12, 0.30)	58%
	Eff-ML	0.22	(0.10, 0.33)	57%	0.20	(0.11, 0.30)	55%
PPACT	Unadjusted	−0.83	(−1.31, −0.35)	–	−0.70	(−1.12, −0.28)	–
	GEE-g	−0.53	(−0.83, −0.22)	60%	−0.53	(−0.83, −0.22)	47%
	LMM-g	−0.53	(−0.84, −0.21)	56%	−0.52	(−0.81, −0.22)	52%
	Eff-PM	−0.59	(−0.98, −0.21)	35%	−0.51	(−0.81, −0.21)	49%
	Eff-ML	−0.62	(−0.99, −0.24)	39%	−0.54	(−0.84, −0.24)	50%
ACTS	Unadjusted	0.08	(−0.00, 0.16)	–	0.08	(−0.00, 0.16)	–
	GEE-g	0.07	(−0.01, 0.15)	9%	0.07	(−0.01, 0.15)	9%
	LMM-g	0.07	(−0.00, 0.15)	12%	0.07	(−0.00, 0.15)	12%
	Eff-PM	0.08	(0.00, 0.15)	20%	0.07	(−0.00, 0.15)	18%
	Eff-ML	0.08	(−0.00, 0.15)	12%	0.07	(−0.00, 0.15)	18%

Unadjusted: the unadjusted estimator. GEE-g: GEE with weighted g-computation. LMM-g: linear mixed models with weighted g-computation. Eff-PM: our proposed method with parametric working models. Eff-ML: our proposed method with machine learning algorithms. ESE: empirical standard error. 95% C.I.: 95% confidence interval based on t-distribution. PVR: proportional variance reduction compared to the unadjusted estimator.

population analysis for PPACT and ACTS. For each estimator, we reported the point estimates, 95% confidence interval based on t-distribution, and proportion variance reduction compared to the unadjusted estimator (PVR).

Table 2 summarizes the results of our data analyses. Across all studies and for both estimands, the unadjusted analysis has the widest confidence interval, and covariate adjustment can offer variance reduction as high as 65%. While the weighted g-computation estimators have higher PVR than our proposed estimators in WFHS and PPACT for estimating the cluster-average treatment effect, they may be biased since Assumption 4 is violated under $M = N$; in contrast, our proposed methods can remove such bias and are hence more reliable. In the analysis of these three cluster-randomized experiments, machine learning algorithms did not show an apparent advantage over parametric working models.

To further illustrate our methods under the setting of cluster-dependent sampling, we performed a simulation study in the supplementary material based on the WFHS data. This analysis mimics the real-world setting, that is, the distribution of outcome and covariates under control is based on real data. The simulation results, summarized in Table S5 in the supplementary material, showed consistent findings with our theoretical results.

7. Concluding Remarks

Under the overarching goal to improve the current practice of covariate adjustment in cluster-randomized experiments, our contributions to the literature are 2-fold. Above all, we clarified a set of sufficient conditions under which two model-based regression estimators, when combined with weighted g-computation, are robust for estimating the cluster-average treatment effect and individual-average treatment effect, even when their working models are arbitrarily misspecified. Given the frequency of their use in practice, our results serve as important clarifications for existing practice in cluster-randomized experiments and provide simple recipes for robust covariate adjustment through GEE and linear mixed models. Despite the simplicity and accessibility of these model-based estimators, their model-robustness property largely hinges on the arm-specific random sampling assumption. Furthermore, an incorrectly specified working model limits the ability to maximally leverage the precision gain from covariate adjustment. These limitations have motivated us to search for more principled and efficient strategies for covariate adjustment without compromising the model-robustness property for the two classes of estimands. Therefore, as a second contribution, we have derived the efficient influence functions and proposed efficient estimators for the two classes of estimands that allow for efficient covariate adjustment and additionally accommodate cluster-dependent sampling. The efficient estimators open the door for using a wider class of parametric working models or machine learning algorithms to learn the potentially complex data-generating mechanisms without affecting the validity of causal inference in cluster-randomized experiments.

Our asymptotic framework assumes that the source population size of each cluster is bounded. This is a convenient and yet practice assumption that avoids challenges in defining our causal estimands and in addressing the potentially high dimensionality of W_i. Although this assumption may appear strong, we can set the upper bound of the source population size to be large enough and accommodate most real-world settings without affecting our asymptotic theory. For example, the upper bound of the source population size may be 100 for cluster-randomized experiments when the randomization units are classrooms, whereas, in healthcare settings, this upper bound may be much larger but still considered to be finite. The extension of our current asymptotic theory to allow for potentially infinite source population size along with required restrictions on $\mathcal{P}^{W|N} \times \mathcal{P}^{N}$ is an area of future research.

A common objective for covariate adjustment in cluster-randomized experiments is to address chance imbalance and improve precision (Su and Ding 2021). However, how to best select the optimal set and functional forms of covariates is an open problem that remains to be addressed in future research. For cluster-randomized experiments, this problem may be more challenging due to the unknown intracluster correlations of the outcomes and covariates within each cluster. In many cases where the investigators can only include a limited number of clusters, there will be a tradeoff between the loss of degrees of freedom by adjusting for weakly prognostic covariates and the potential asymptotic precision gain by including more baseline variables. While the proposed estimators may be a useful vehicle to incorporate variable selection techniques in the working models, we maintain the recommendation of pre-specifying prognostic covariates for adjustment in the design stage based on subject-matter knowledge for practical applications.

Throughout the article, we have defined the cluster-average treatment effect and individual-level treatment effect estimands as a function of the source population size N_i, which can differ from the observed cluster size M_i. Therefore, accurate identification of these estimands require knowledge of the source population size. Conceptually, this source population represents the set of eligible participants in each cluster that may be recruited had the investigator obtained unlimited financial and logistical resources, and is precisely the set of individuals that the intervention is designed to target. The availability of N_i typically depends on the types of clusters and the resource of the study. For example, the source population size can be readily available if schools, worksites and villages are randomized, whereas the source population size may be estimated from historical data if clinics or hospitals are randomized. In this latter case, one may set $N_i = M_i$ to perform an enrolled population analysis, which implicitly assumes equivalence between the observed population and the source population in each cluster. In the case where $N_i > M_i$ but no information of N_i is available, we stated in Remark 3 that only the cluster-average treatment effect is identifiable. In future work, it would be worthwhile to establish alternative conditions to identify the individual-level treatment effect without knowledge of the source population size.

For developing the asymptotic properties of the model-robust estimators, we have primarily focused on simple randomization and discussed the implications under stratified cluster randomization and biased-coin cluster randomization. Beyond these randomization schemes, cluster rerandomization is a useful strategy to address baseline imbalance and further improve the study power. Lu et al. (2023) recently developed the asymptotic theory for cluster rerandomization under a finite-population framework. It would be useful to further extend our results to accommodate cluster rerandomization.

Supplementary Materials

Supplementary material includes the R code, a causal graph summarizing all random variables, regularity conditions for Theorems 1–4, consistent variance estimators, proofs, additional simulation studies, and a review of other methods for cluster-randomized experiments.

Acknowledgments

The statements presented are solely the responsibility of the authors and do not necessarily represent the official views of PCORI®, its Board of Governors or Methodology Committee, or NIAID.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

Research in this article was supported by Patient-Centered Outcomes Research Institute Awards® (PCORI® Awards ME-2020C3-21072, ME-2022C2-27676) and National Institute of Allergy and Infectious Diseases (NIAID) grants R01AI148127, K99AI173395.

ORCID

Fan Li https://orcid.org/0000-0001-6183-1893

References

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, 502–517. [2960,2964]

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, 1761–1780. [2964]

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. (2021), "Comparative Methods for the Analysis of Cluster Randomized Trials," arXiv preprint arXiv:2110.09633. [2960,2964]

Bray, J. W., Kelly, E. L., Hammer, L. B., Almeida, D. M., Dearing, J. W., King, R. B., and Buxton, O. M. (2013), *An Integrative, Multilevel, and Transdisciplinary Research Approach to Challenges of Work, Family, and Health*, Methods Report, pp. 1–38, Research Triangle Park, NC: RTI Press. [2968]

Breslow, N. E., and Clayton, D. G. (1993), "Approximate Inference in Generalized Linear Mixed Models," *Journal of the American statistical Association*, 88, 9–25. [2959]

Bugni, F. A., Canay, I. A., and Shaikh, A. M. (2018), "Inference Under Covariate-Adaptive Randomization," *Journal of the American Statistical Association*, 113, 1784–1796. [2963]

Bugni, F. A., Canay, I. A., Shaikh, A. M., and Tabord-Meehan, M. (2023), "Inference for Cluster Randomized Experiments with Non-ignorable Cluster Sizes," arXiv preprint arXiv:2204.08356. [2959,2960,2961,2966]

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018), "Double/Debiased Machine Learning for Treatment and Structural Parameters," *The Econometrics Journal*, 21, C1–C68. [2964,2965]

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, 46–55. [2968]

Donner, A., and Klar, N. (2000), *Design and Analysis of Cluster Randomization Trials in Health Research*, London: Arnold. [2959,2962]

Efron, B. (1971), "Forcing a Sequential Experiment to be Balanced," *Biometrika*, 58, 403–417. [2961]

Farrell, M. H., Liang, T., and Misra, S. (2021), "Deep Neural Networks for Estimation and Inference," *Econometrica*, 89, 181–213. [2965]

Hines, O., Dukes, O., Diaz-Ordaz, K., and Vansteelandt, S. (2022), "Demystifying Statistical Learning based on Efficient Influence Functions," *The American Statistician*, 76, 292–304. [2964,2966]

Imai, K., King, G., and Nall, C. (2009), "The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation," *Statistical Science*, 24, 29–53. [2960]

Jiang, L., Linton, O. B., Tang, H., and Zhang, Y. (2022), "Improving Estimation Efficiency via Regression-Adjustment in Covariate-Adaptive Randomizations with Imperfect Compliance," arXiv preprint arXiv:2201.13004. [2963]

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, 107–118. [2960,2961]

Li, F., Tian, Z., Bobb, J., Papadogeorgou, G., and Li, F. (2022), "Clarifying Selection Bias in Cluster Randomized Trials," *Clinical Trials*, 19, 33–41. [2961]

Liang, K.-Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22. [2959,2962]

Lu, X., Liu, T., Liu, H., and Ding, P. (2023), "Design-Based Theory for Cluster Rerandomization," *Biometrika*, 110, 467–483. [2970]

Luo, Y., and Spindler, M. (2016), "High-Dimensional l_2-Boosting: Rate of Convergence," arXiv preprint arXiv:1602.08927. [2965]

MacKinnon, J. G., and White, H. (1985), "Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties," *Journal of Econometrics*, 29, 305–325. [2967]

Middleton, J. A., and Aronow, P. M. (2015), "Unbiased Estimation of the Average Treatment Effect in Cluster-Randomized Experiments," *Statistics, Politics and Policy*, 6, 39–75. [2960]

Murray, D. M. (1998), *Design and Analysis of Group-Randomized Trials* (Vol. 29), New York: Oxford University Press. [2959,2962]

Neyman, J. S., Dabrowska, D. M., and Speed, T. (1990), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9," *Statistical Science*, 5, 465–472. [2960]

Prudhomme O'Meara, W., Menya, D., Laktabai, J., Platt, A., Saran, I., Maffioli, E., Kipkoech, J., Mohanan, M., and Turner, E. L. (2018), "Improving Rational Use of Acts through Diagnosis-Dependent Subsidies: Evidence from a Cluster-Randomized Controlled Trial in Western Kenya," *PloS Medicine*, 15, e1002607. [2968]

Rafi, A. (2023), "Efficient Semiparametric Estimation of Average Treatment Effects Under Covariate Adaptive Randomization," arXiv preprint arXiv:2305.08340. [2963,2966]

Rosenbaum, P. R. (1987), "Model-based Direct Adjustment," *Journal of the American Statistical Association*, 82, 387–394. [2962]

Schochet, P. Z., Pashley, N. E., Miratrix, L. W., and Kautz, T. (2022), "Design-based Ratio Estimators and Central Limit Theorems for Clustered, Blocked RCTs," *Journal of the American Statistical Association*, 117, 2135–2146. [2960]

Shi, Y., and Lee, J.-H. (2018), "Sample Size Calculations for Group Randomized Trials with Unequal Group Sizes through Monte Carlo Simulations," *Statistical Methods in Medical Research*, 27, 2569–2580. [2962]

Stephens, A. J., Tchetgen Tchetgen, E. J., and Gruttola, V. D. (2012), "Augmented Generalized Estimating Equations for Improving Efficiency and Validity of Estimation in Cluster Randomized Trials by Leveraging Cluster-Level and Individual-Level Covariates," *Statistics in Medicine*, 31, 915–930. [2964]

Su, F., and Ding, P. (2021), "Model-Assisted Analyses of Cluster-Randomized Experiments," *Journal of the Royal Statistical Society*, Series B, 83, 994–1015. [2960,2961,2962,2964,2969]

Turner, E. L., Prague, M., Gallis, J. A., Li, F., and Murray, D. M. (2017), "Review of Recent Methodological Developments in Group-Randomized Trials: Part 2-Analysis," *American Journal of Public Health*, 107, 1078–1086. [2960]

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007), "Super Learner," *Statistical Applications in Genetics and Molecular Biology*, 6, 1–21. [2967]

van der Laan, M. J.,and Rose, S. (2011), *Targeted Learning: Causal Inference for Observational and Experimental Data* (Vol. 10), New York: Springer. [2964]

van der Vaart, A. (1998), *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press. [2962]

Wager, S., and Walther, G. (2015), "Adaptive Concentration of Regression Trees, with Application to Random Forests," arXiv preprint arXiv:1503.06388. [2965]

Wang, B., Harhay, M. O., Small, D. S., Morris, T. P., and Li, F. (2021), "On the Mixed-Model Analysis of Covariance in Cluster-Randomized Trials," arXiv preprint arXiv:2112.00832. [2961,2963]

Wang, B., Susukida, R., Mojtabai, R., Amin-Esmaeili, M., and Rosenblum, M. (2023), "Model-Robust Inference for Clinical Trials that Improve Precision by Stratified Randomization and Covariate Adjustment," *Journal of the American Statistical Association*, 118m 1152–1163. [2963]

Wang, X., Turner, E. L., Li, F., Wang, R., Moyer, J., Cook, A. J., Murray, D. M., and Heagerty, P. J. (2022), "Two Weights Make a Wrong: Cluster Randomized Trials with Variable Cluster Sizes and Heterogeneous Treatment Effects," *Contemporary Clinical Trials*, 114, 106702. [2959,2962]

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. [2959]

Work, Family, and Health Study (WFHS). (2018), "Work, Family and Health Network," *Inter-university Consortium for Political and Social Research [distributor]*. [2968]

Zelen, M. (1974), "The Randomization and Stratification of Patients to Clinical Trials," *Journal of Chronic Diseases*, 27, 365–375. [2961]