Leveraging baseline covariates to analyze small cluster-randomized trials with a rare binary outcome

Angela Y. Zhu1 | Nandita Mitra1 | Karla Hemming2 | Michael O. Harhay1 | Fan Li3,4

1Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Perelman School of Medicine, Philadelphia, Pennsylvania, USA
2Department of Public Health, Epidemiology, and Biostatistics, University of Birmingham, Institute of Applied Health Research, Birmingham, UK
3Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut, USA
4Center for Methods in Implementation and Prevention Science, Yale School of Public Health, New Haven, Connecticut, USA

Abstract
Cluster-randomized trials (CRTs) involve randomizing entire groups of participants—called clusters—to treatment arms but are often comprised of a limited or fixed number of available clusters. While covariate adjustment can account for chance imbalances between treatment arms and increase statistical efficiency in individually randomized trials, analytical methods for individual-level covariate adjustment in small CRTs have received little attention to date. In this paper, we systematically investigate, through extensive simulations, the operating characteristics of propensity score weighting and multivariable regression as two individual-level covariate adjustment strategies for estimating the participant-average causal effect in small CRTs with a rare binary outcome and identify scenarios where each adjustment strategy has a relative efficiency advantage over the other to make practical recommendations. We also examine the finite-sample performance of the bias-corrected sandwich variance estimators associated with propensity score weighting and multivariable regression for quantifying the uncertainty in estimating the participant-average treatment effect. To illustrate the methods for individual-level covariate adjustment, we reanalyze a recent CRT testing a sedation protocol in 31 pediatric intensive care units.

KEYWORDS
covariate adjustment, finite-sample corrections, generalized estimating equations, inverse probability weights, overlap weights, sandwich variance estimator
1 INTRODUCTION

In cluster-randomized trials (CRTs), entire clusters of subjects, rather than the individuals themselves, are randomly allocated to treatment and control arms (Turner, Li, et al., 2017). Examples of clusters include hospitals, schools, and residential care homes (Murray, 1998). CRTs are essential when an intervention is delivered at the cluster level and are also often used when there are concerns over contamination or because of their logistical appeal. Due to cluster randomization and the fact that individuals within each cluster share the same physical or social factors, the intracluster correlation coefficient (ICC) that describes the similarity between outcomes within the same cluster plays a central role in the design and analysis of CRTs. This ICC must be accounted for in analysis to avoid inflated type I error rates (Turner, Prague, et al., 2017).

Methods to analyze CRTs include cluster-level analyses that use summary measures for each cluster and individual-level analyses that employ generalized linear mixed models or marginal models estimated with generalized estimating equations (GEEs) (Liang & Zeger, 1986). Marginal models are sometimes preferred because the interpretation of the intervention effect parameter does not depend on the assumed correlation structure (Preisser et al., 2003). In addition, marginal models coupled with the sandwich variance estimator have been shown to be robust to misspecification of the working covariance structure and to provide asymptotically valid confidence intervals as long as the marginal mean structure is correctly specified (Liang & Zeger, 1986). However, when the number of clusters is small (typically not exceeding 30), the sandwich variance estimator can exhibit negative bias and may lead to undercoverage of the associated confidence interval estimator. Indeed, Ivers et al. (2011) reviewed 300 CRTs published between 2000 and 2008 and found the median number of clusters to be 21. Given small CRTs are common, there remains considerable interest in improving finite-sample inference of the sandwich variance estimator for GEE analysis of such trials (Leyrat et al., 2018).

Previous simulation studies have compared the performances of several bias-corrected sandwich variance estimators under certain settings. Specifically, Lu et al. (2007) compared the bias-corrected sandwich variance estimators due to Kauermann and Carroll (2001) (KC) and Mancl and DeRouen (2001) (MD). They found that the normal confidence interval estimator with the MD variance formula often leads to coverage of 95% confidence intervals near the nominal level. P. Li and Redden (2015) have recommended the corrections proposed by KC and Fay and Graubard (2001) (FG) depending on variations of cluster size when analyzing binary outcomes. Ford and Westgate (2017) have indicated that the KC bias-corrected sandwich variance estimator may still give downwardly biased estimates of the standard error and that the FG bias-corrected sandwich variance estimator performs similarly to that of KC. Further, they note that the MD bias-corrected sandwich variance estimator tends to overcorrect resulting in conservative inference, and thus recommended the average of the MD and KC standard error estimators as the top performer in CRTs with continuous and binary outcomes. Similar observations and recommendations were discussed in F. Li and Tong (2021a) for CRTs with count outcomes subject to truncation. It is evident that the performance of these bias-corrected sandwich variance estimators can depend on the settings of interest.

One limitation of existing comparative studies of analytical methods for CRTs is that they have primarily focused on unadjusted analyses without covariates or adjusted analyses with cluster-level covariates. For example, Westgate (2013) and Ford and Westgate (2017) have investigated the small-sample corrections in the presence of cluster-level covariates. They pointed out that the performance of MD and KC bias-corrected sandwich variance estimators may deteriorate with cluster-level covariates, while the average of MD and KC standard error estimators was relatively more robust. However, the performance of small-sample corrections remains to be explored in settings where adjustment for individual-level covariates is being considered. In CRTs, individual-level covariates are frequently collected at baseline, and the need for covariate adjustment can fall into one of the following categories. First, in CRTs where cluster-level aggregates of individual-level covariates are utilized during restricted randomization, adjusting for such individual-level covariates in the analysis model can adequately control for the type I error rate (F. Li et al., 2015, 2017; Watson et al., 2021; Zhou et al., 2022). Second, adjusting for individual-level covariates can be based on precision considerations, analogous to the justifications provided in individually randomized trials (Williamson et al., 2014; Zeng et al., 2021). Third, individual-level covariate adjustment may reduce selection or recruitment bias compared to the unadjusted analysis in CRTs (Leyrat et al., 2013, 2014; F. Li et al., 2022).

In this article, we design a neutral simulation study to compare methods for analyzing small CRTs in the presence of individual-level covariate adjustment. We focus on a rare binary outcome (event rate ≤ 10% as in our motivating study in Section 5). Besides considering individual-level covariate adjustment, the unique innovations of our simulation study are threefold. First, we recognize the potential ambiguity in the treatment effect parameter under individual-level covariate adjustment for binary outcomes (due to noncollapsibility) and so employ the potential outcomes framework to define
a participant-average treatment effect on the odds ratio scale as a common target estimand. Fixing a common target estimand has been recommended for individually randomized trials (Benkeser et al., 2021) and can facilitate an objective comparison across different methods where the associated regression coefficients may have different interpretations. This work thus represents a first effort to operationalize this important consideration for CRTs. Second, we focus on a rare outcome where adjusting for multiple individual-level covariates may lead to separation issues and nonconvergence when fitting multivariable regression models. Complete separation occurs when the value of the outcome may be determined or distinguished for values of the predictors; that is, there is no observed variability in the outcome for certain combinations of covariate values. This convergence problem, however, may be ameliorated using propensity score weighting (Williamson et al., 2014; Zeng et al., 2021), which has been demonstrated to be a valid covariate adjustment approach in individually randomized trials, and we adapt this approach to estimate causal effects in CRTs. Finally, under a GEE-modeling framework, we contribute new evidence on whether the bias-corrected sandwich variance estimators can provide valid inference for the participant-average treatment effect under individual-level covariate adjustment when the number of clusters is small.

2 METHODS FOR INDIVIDUAL-LEVEL COVARIATE ADJUSTMENT IN CRTS WITH A RARE BINARY OUTCOME

2.1 Notation and estimand

Suppose we have a two-arm parallel CRT with $N$ total clusters and 1:1 randomization, and let $Z$ denote the treatment indicator with $Z = 1$ corresponding to inclusion in the treatment group and $Z = 0$ corresponding to being in the control group. Let $Y_{ij}$ denote the binary outcome of the $j$th subject in cluster $i$ with $P$-dimensional covariate vector $X_{ij}$. We assume the outcomes are correlated within the same cluster but independent across clusters. Specifically, for the $i$th cluster, $Y_i = [Y_{i1}, ..., Y_{imi}]$ is the outcome vector and $X_i = [X_{i1}, ..., X_{imi}]$ is the covariate matrix. In this article, we focus on estimating an average treatment effect on the ratio scale. Briefly, as the outcome of interest is binary, we describe the target estimand as the log causal odds ratio of the treatment to the control. We proceed under the potential outcomes framework and assume $\{Y_{ij}(1), Y_{ij}(0)\} \in \{0, 1\}^\otimes 2$ as the pair of potential/counterfactual binary outcomes under the treatment and control conditions; for example, $Y_{ij}(1)$ is the binary potential outcome of the $j$th subject in cluster $i$ had cluster $i$ been randomized to treatment, and $Y_{ij}(0)$ is the binary potential outcome of the $j$th subject in cluster $i$ had cluster $i$ been randomized to control. We consider a sampling-based framework where the observed clusters are assumed to be a random sample of a population of clusters with randomly varying cluster sizes. This framework allows us to define the expectation sign over the distribution of clusters and permits us to formalize the target estimand—participant-average treatment effect on the odds ratio scale—as

$$ OR = \frac{E\left(\sum_{j=1}^{m_i} Y_{ij}(1)\right)}{E\left(\sum_{j=1}^{m_i} Y_{ij}(0)\right)} \frac{m_i - \sum_{j=1}^{m_i} Y_{ij}(0)}{m_i - \sum_{j=1}^{m_i} Y_{ij}(1)}, $$

(1)

and we further define $\Delta = \log(OR)$ as our target parameter in the ensuing simulations. While our focus is on estimand (1), we refer to Kahan et al. (2023), Wang, Turner, Li et al. (2022), and Su and Ding (2021) for alternative estimands such as the cluster-average treatment effect or other weighted cluster-average treatment effects that may also be of interest in CRTs.

2.2 Overview of generalized estimating equations analyses of CRTs

To estimate the parameter defined in (1), we primarily consider the GEE approach, under which, the relationship between the marginal mean $E[Y_{ij}|X_{ij}] = \mu_{ij}$ and the covariates $X_{ij}$ may be represented by a generalized linear model, $g(\mu_{ij}) = X_{ij}\beta$. With a binary outcome, logistic regression is often used where $g$ is taken as the canonical logit link function, which we use in this article. Let $V_i$ denote the working covariance structure for $Y_i$. The parameter estimator $\hat{\beta}$ in the marginal model is consistent and asymptotically normal, and, even if the working correlation structure is misspecified, its variance–covariance can be consistently estimated by $V = \text{Cov}(\hat{\beta}) = \Omega_s \sum_{i=1}^{N} D_i' V_i^{-1} r_i V_i^{-1} D_i \Omega_s$, where $D_i = \partial \mu_{ij}/\partial \beta'$,
\[ \Omega = (\sum_{i=1}^{N} D^T V_i^{-1} D_i)^{-1} \] and \( r_i r'_i = (Y_i - \mu_i) (Y_i - \mu_i)' \) is an estimate of the covariance of \( Y_i \). The variance \( V \) is often referred to as the robust sandwich estimator or the empirical sandwich estimator (Liang & Zeger, 1986).

Frequently, the primary analysis of CRTs proceeds with the marginal model without any baseline covariates—the so-called unadjusted analysis. In this approach, \( X_{ij} = Z_i \) in the marginal mean model and the mean model can be explicitly written as

\[
\text{logit}(\mu_{ij}) = \beta_0 + \beta_Z Z_i,
\]

and \( \beta = (\beta_0, \beta_Z)' \). To estimate \( \beta \), one may choose either the independence or exchangeable working correlation structure. When the true outcome data-generating model is indeed the unadjusted model without any further covariates \( (2) \), \( \hat{\beta}_Z \) directly corresponds to our target estimand \( \Delta \), and the choice of exchangeable working correlation structure corresponds to modeling the ICC and can often provide a more efficient average causal effect estimator when the cluster sizes are variable (F. Li & Tong, 2021a, 2021b). In this case, even when the exchangeable working structure is not the true correlation structure, GEE models are robust to such misspecification. However, when the true data-generating process involves additional individual-level covariates or when the cluster size is predictive of the treatment effect, we show in the Supplementary File Section 1 that \( \hat{\beta}_Z \) is a consistent estimator to \( \Delta \) under the independence working correlation structure but often not otherwise. This argument extends the ones provided in Wang, Turner, Li et al. (2022) and Kahan et al. (2023) to ratio effect measures. Because of this rationale, we primarily focus on the case with an independence working correlation model. Beyond the unadjusted analysis, the GEE approach can be extended to leverage baseline individual-level covariates to potentially increase the efficiency for estimating \( \Delta \). In Section 2.3, we consider using propensity score weighting for covariate adjustment. In addition, model \( (2) \) can be expanded to include additional baseline covariates, in which case a population standardization procedure is required to estimate \( \Delta \) because the regression coefficient does not directly correspond to \( \Delta \) due to noncollapsibility. We describe this multivariable regression approach in Section 2.4. As CRTs often include a limited number of clusters (typically not exceeding 30), a set of bias-corrected sandwich variance estimators to address the finite-sample bias in inference is also described below.

### 2.3 Propensity score weighting for individual-level covariate adjustment

Rosenbaum and Rubin (1983) developed the propensity score, which is defined as the probability of treatment conditional on observed covariates; that is, \( e(X) = P(Z = 1 | X) \). Approaches based on the propensity score, such as matching, weighting, and stratification, are commonly employed in the design and analysis of observational studies to control for confounding since it has been shown that conditional on the propensity score, treatment is randomized. In randomized trials, the true propensity score is known by design and there is no need to model the propensity score for unbiased estimation of treatment effect. However, propensity score weighting has been shown to provide efficiency gains by addressing chance imbalance of baseline covariates. Williamson et al. (2014) have shown in individually randomized trials that (i) inversely weighting by the estimated propensity score with prognostic covariates can reduce the variance of the unadjusted treatment effect estimator without compromising the population-level causal estimand and (ii) with a rare binary outcome, the propensity score weighting approach often circumvents nonconvergence issues that multivariable regression is vulnerable to. In addition, Zeng et al. (2021) demonstrated that in individually randomized trials, weighting by overlap weights almost always leads to smaller variance than inverse propensity score weighting. Compared to multivariable regression adjustment for covariates, propensity score weighting has the additional benefit of preserving the population-level causal estimands but has not yet been explored in CRTs as a potentially effective tool for individual-level covariate adjustment. Here, we explore the use of these two propensity score weighting approaches as individual-level covariate adjustment strategies for CRTs, without compromising the clarity of the target estimand.

Inverse probability of treatment weighting (IPW) seeks to construct a sample in which the distributions of observed baseline variables are similar between treatment and control groups. These weights are defined to be the reciprocal of the conditional probability of being assigned to the treatment group that they were observed to be in. In CRTs, for subject \( j \) in cluster \( i \), the weight is given by

\[
\omega_{ij} = \begin{cases} 
1/e(X_{ij}) & \text{if treated (} Z_i = 1 \text{)} \\
1/(1 - e(X_{ij})) & \text{if control (} Z_i = 0 \text{)}.
\end{cases}
\]
On the other hand, overlap weighting (OW) was proposed to overcome possible limitations of IPW when there is limited overlap in covariate distributions between treatment arms in observational studies (F. Li, Morgan, & Zaslavsky, 2018) and to also improve upon IPW in individually randomized trials. Specifically, the overlap weights are defined to be the probability of being in the opposite treatment group (the one the subject was not observed to be in) given baseline covariates:

$$w_{ij} = \begin{cases} 1 - e(X_{ij}) & \text{if treated (} Z_i = 1) \\ e(X_{ij}) & \text{if control (} Z_i = 0). \end{cases}$$

For individual-level randomized trials, the true propensity score is usually a constant and does not depend on individual-level covariates; therefore, IPW and OW correspond to the same population estimand (Zeng et al., 2021), but previous simulations indicate that OW provides better finite-sample performance in that it is more efficient at smaller sample sizes. Under cluster randomization, the true propensity score is determined by study design and does not depend on individual-level covariates. Therefore, using OW and IPW can both target the participant-average treatment effect (1) as a common estimand.

The propensity score is often estimated using parametric logistic regression. Alternative models for propensity score estimation that have been considered include neural nets, decision trees, Bayesian additive regression trees (BART), and ensemble learners (Westreich et al., 2010; Zhu et al., 2021). In observational studies, these nonparametric machine learning approaches can provide greater flexibility and more accurate estimates when there are complex relationships among the variables; however, their role in the analysis of CRTs remains underexplored. In particular, BART is a Bayesian nonparametric sum-of-trees model that involves binary splits in the predictor space and a regularization prior to avoid overfitting (Chipman et al., 2010). This model has become popular for estimating treatment effects in observational studies due to its computational efficiency and flexibility in modeling complex relationships between variables (Hill, 2011). In settings where there are nonlinear relationships or interactions among the covariates in the propensity score or outcome models, there may be difficulty in specifying those patterns through a simple parametric model. BART models can handle the inclusion of several predictors, allow for nonlinear relationships and interactions between predictors, and have been demonstrated to be effective in past simulations for observational studies (Dorie et al., 2019). In the ensuing simulations, we explore whether using BART to estimate propensity scores will impact results in the CRT context.

Next, we describe the point estimator and the bias-corrected sandwich variance estimators under propensity score weighted GEE. The propensity score weighted GEE is essentially based on the same marginal mean model as in (2) but includes propensity score weights in the estimating equations for $\beta$. Specifically, once the propensity scores are estimated for each subject, the weight matrix $W_i$ is formed for each cluster $i$ with $w_{ij}$ on the diagonal and 0 for the off-diagonal elements. The regression parameters $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_Z)'$ are estimated by solving the weighted GEE, $\sum_{i=1}^N D_i^t V_i^{-1} W_i (Y_i - \mu_i) = 0$, where $\mu_i = (\mu_{i1}, ..., \mu_{im_i})'$.

$$\mu_{ij} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_Z Z_i)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_Z Z_i)},$$

and $V_i$ is the working variance under the independence working correlation model.

Under the independence working correlation assumption, since the weighted GEE model only contains the treatment indicator as a covariate and the weights are only used to control for chance imbalance, the treatment coefficient estimator is the $\hat{\beta}_Z = \log(\hat{OR})$, which is a consistent estimator for the participant-average treatment effect $\Delta$. For convenience, we ignore the variability due to the estimation of propensity scores, and the corresponding bias-corrected sandwich variance estimators incorporating these weights are developed in Table 1 (panel a). Here, $\hat{\Omega} = (\sum_{i=1}^N D_i^t V_i^{-1} W_i D_i)^{-1}$ is the propensity score weighted “model-based” variance, and $H_i = D_i \hat{\Omega} D_i^t V_i^{-1} W_i$ is the propensity score weighted leverage matrix for cluster $i$. Although we focus on the MD and KC bias-corrected variance estimators in our simulations, prior studies have indicated that the FG bias-corrected variance estimator had very similar performance to the KC bias-corrected variance estimator (F. Li, Turner, & Preisser, 2018). In fact, without weighting, Scott et al. (2017) and Wang, Turner, Preisser et al. (2022) have shown that the KC bias-corrected variance is equivalent to a modified version of the FG bias-corrected variance. This analytical insight also holds in the presence of weighting. We, therefore, do not further consider the FG bias-correction due to its similarity to the KC bias-correction but still provide code to implement the FG bias-correction in the Supporting Information.
TABLE 1 Bias-corrected sandwich variance estimators for propensity score weighted GEE and multivariable-adjusted GEE estimators of the participant average treatment effect. Notice that under propensity score weighting, Cov(\hat{\beta}) is a 2 \times 2 matrix, and the variance estimator for \hat{\beta}_z = \log(\hat{\Omega}) is the lower-right element of Cov(\hat{\beta}). The matrix M is defined in Equation (3).

| (a) Variance estimation under propensity score weighting | Cov(\hat{\beta}) |
|--------------------------------------------------------|-----------------|
| Robust sandwich estimator                               | \Omega(\sum_{i=1}^{N} D'V_i^{-1}W_i D_i)\Omega |
| Mancl and DeRouen bias-corrected estimator              | \Omega(\sum_{i=1}^{N} D'V_i^{-1}W_i (I - H_i)^{-1} D_i)\Omega |
| Kauermann and Carroll bias-corrected estimator          | \Omega(\sum_{i=1}^{N} D'V_i^{-1}W_i (I - H_i)^{-1/2} D_i)\Omega |

| (b) Variance estimation under a multivariable model     | \text{Var}[\log(\hat{\Omega})] |
|--------------------------------------------------------|-------------------------------|
| Robust sandwich estimator                               | M'\Omega(\sum_{i=1}^{N} D'V_i^{-1} r_i (I - H_i)^{-1} D_i)\Omega M |
| Mancl and DeRouen bias-corrected estimator              | M'\Omega(\sum_{i=1}^{N} D'V_i^{-1} (I - H_i)^{-1} D_i)\Omega M |
| Kauermann and Carroll bias-corrected estimator          | M'\Omega(\sum_{i=1}^{N} D'V_i^{-1} (I - H_i)^{-1/2} D_i)\Omega M |

2.4 Direct multivariable regression for individual-level covariate adjustment

As an alternative to propensity score weighting, we also consider the direct multivariable regression approach for covariate adjustment in CRTs. Due to noncollapsibility with the logit link function, the estimator of the coefficient for the treatment variable (when there are other individual-level covariates in the model) only reflects a conditional treatment effect. To ensure we are targeting the correct estimand \Delta, we will use the model fit to obtain estimates of the probability of the potential outcomes averaged or standardized over the covariate distribution. Suppose we fit the multivariable model \text{logit}\{E(Y_{ij}|X_{ij}, Z_i)\} = \beta_0 + \sum_{p=1}^{P} \beta_p X_{ij}^{(p)} + \beta_2 Z_i. Let \hat{\beta}' = [\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_P, \hat{\beta}_2] denote the estimates of the coefficients from the GEE model fit, where \hat{\beta} represents the log odds ratio estimator conditional on all covariates. For estimating the multivariable-adjusted GEE model, one could choose the working correlation to be either the independence correlation or the exchangeable correlation, and we focus on the former in this paper for an objective comparison with the propensity score approaches. Based on the GEE model fit, the predicted risk for subject \(j\) in cluster \(i\), if he/she is given treatment is \(\hat{P}_{1,ij} = \exp(X_{1,ij}'\hat{\beta})/\{1 + \exp(X_{1,ij}'\hat{\beta})\}\), where \(X_{1,ij} = [1, X'_{ij}, 1]'\). Similarly, the predicted risk for subject \(j\) in cluster \(i\) under the control is \(\hat{P}_{0,ij} = \exp(X_{0,ij}'\hat{\beta})/\{1 + \exp(X_{0,ij}'\hat{\beta})\}\), where \(X_{0,ij} = [1, X'_{ij}, 0]'\). Then a standardization estimator (or referred to as the g-computation formula estimator) of the participant-average treatment effect is

\[
\log(\hat{\Omega}) = \log \left( \frac{\sum_{i=1}^{N} \sum_{j=1}^{m_i} \hat{P}_{1,ij}}{\sum_{i=1}^{N} \sum_{j=1}^{m_i} \hat{P}_{0,ij}} \right) \left( \frac{\sum_{i=1}^{N} \sum_{j=1}^{m_i} (m_i - \sum_{j=1}^{m_i} \hat{P}_{0,ij})}{\sum_{i=1}^{N} \sum_{j=1}^{m_i} (m_i - \sum_{j=1}^{m_i} \hat{P}_{1,ij})} \right).
\]

For variance estimation, using the delta method, we arrive at \(\text{Var}(\log(\hat{\Omega})) = M'\text{Cov}(\hat{\beta}) M\), where \text{Cov}(\hat{\beta}) may either be the robust sandwich estimator or the bias-corrected estimators in the existing literature (F. Li et al., 2015; Ford & Westgate, 2017) and \(M = \partial \log(\hat{\Omega})/\partial \beta\). We provide detailed steps of the derivation in Supplementary File Section 2 to show that \(M\) takes the form of

\[
M = \left( \frac{1}{\sum_{i=1}^{N} \sum_{j=1}^{m_i} \hat{P}_{1,ij}} + \frac{1}{\sum_{i=1}^{N} (m_i - \sum_{j=1}^{m_i} \hat{P}_{1,ij})} \right) \left( \sum_{i=1}^{N} \sum_{j=1}^{m_i} \frac{\exp(X_{1,ij}'\hat{\beta})X_{1,ij}}{\{1 + \exp(X_{1,ij}'\hat{\beta})\}^2} \right)

- \left( \frac{1}{\sum_{i=1}^{N} \sum_{j=1}^{m_i} \hat{P}_{0,ij}} + \frac{1}{\sum_{i=1}^{N} (m_i - \sum_{j=1}^{m_i} \hat{P}_{0,ij})} \right) \left( \sum_{i=1}^{N} \sum_{j=1}^{m_i} \frac{\exp(X_{0,ij}'\hat{\beta})X_{0,ij}}{\{1 + \exp(X_{0,ij}'\hat{\beta})\}^2} \right).
\]

The analogous robust sandwich estimator and bias-corrected sandwich estimators are also summarized in Table 1 (panel b) for quick reference.
3 | SIMULATION STUDIES

We employ the ADEMP (aims, data-generating mechanisms, estimands, methods, and performance measures) structured approach proposed by Morris et al. (2019) to describe the details of our simulation studies.

3.1 | Aims

The motivation of our simulation studies is to inform practical choices for individual-level covariate-adjusted analysis of CRTs with a small number of clusters and rare binary outcomes. Our goals are twofold. First, we aim to demonstrate the utility of individual-level covariate adjustment in small CRTs when the binary outcomes are rare, hoping to provide some justification for incorporating baseline individual-level covariates in this particularly challenging scenario. Second, we aim to assess and compare the performances of propensity score weighting and multivariable adjustment models as well as the associated bias-corrected sandwich estimators in this context. These two aims collectively address an important gap in the existing literature which has primarily focused on evaluating similar methods assuming common binary outcomes and without sufficient attention to individual-level covariate adjustment.

3.2 | Data-generating mechanism

We generate CRT data with two parallel arms. Suppose $N$ total clusters are randomized to the two arms under 1:1 randomization. The cluster size for each cluster is drawn from a $Poisson(m)$ distribution, where $m$ is the mean cluster size; the exact number of subjects in cluster $i$ is denoted by $m_i$. The outcome ICC under the latent response formulation, $\rho_{Logit}$, reflects similarity among people in the same cluster (F. Li et al., 2017). This parameter will be set at values according to the range reported in practice (Murray et al., 2004). Defining $P_1$ to be the population incidence under treatment and $P_0$ to be the population incidence under the control, we consider two levels of the outcome incidence—low incidence ($P_1 \approx 0.05, P_0 \approx 0.10$) and very low incidence ($P_1 \approx 0.025, P_0 \approx 0.05$).

The general combinations of simulation parameters that we consider are summarized as follows: (i) the number of total clusters: $N = \{6, 10, 20, 30\}$; (ii) mean cluster size: $m = \{100, 30\}$; (iii) ICC on the latent scale: $\rho_{Logit} = \{0.001, 0.01\}$; (iv) the number of covariates: $P = \{6, 15\}$; (v) incidence levels: low and very low. For each simulation setting, 1000 datasets were generated and the Monte Carlo errors will be described in Section 3.5.

For the outcome-generating mechanism, we consider four parametric models. First, we simulate covariates from a standard normal distribution, $X(\cdot) \sim N(0,1)$, $p = 1, \ldots, P$. The first two outcome models assume a constant additive treatment effect with constant covariate effects on the logistic scale—that is, there are no interactions among variables and no treatment effect heterogeneity explained by covariates. Specifically, the latent continuous outcome for the $j$th subject in the $i$th cluster is

$$Y_{ij}^c = \beta_0 + \beta_1 \frac{P}{3} \sum_{p=1}^{P/3} X_{ij}^{(p)} + \beta_2 \sum_{p=P/3+1}^{2P/3} X_{ij}^{(p)} + \beta_3 \sum_{p=2P/3+1}^{P} X_{ij}^{(p)} + \beta_Z Z_i + u_i + \epsilon_{ij}$$

where $u_i \sim N(0,\sigma^2_u)$ is the random effect and $\epsilon_{ij}$ is assumed to follow the standard logistic distribution with mean 0 and variance $\sigma^2_\varepsilon = \pi^2/3$. Then $\sigma^2_u = \rho_{Logit}/(1 - \rho_{Logit}) \times (\pi^2/3)$ according to the latent response definition of binary ICC. The binary outcome is obtained by dichotomizing $Y_{ij}^c$: $Y_{ij} \sim Bernoulli(\text{expit}(Y_{ij}^c))$. We consider two sets of fixed coefficients $\beta_1, \beta_2, \beta_3$ for the covariates $X^{(1)}, \ldots, X^{(P)}$, which correspond to varying strengths of covariate–outcome associations, and $\beta_Z$ is the main effect of the treatment.

(i) Outcome-generating model 1: $\beta_1 = 0, \beta_2 = 0.4, \beta_3 = 0.8$. In this model, some of the covariates are not related to the outcome while the others are weakly correlated with the outcome.

(ii) Outcome-generating model 2: $\beta_1 = 0.8, \beta_2 = 1.6, \beta_3 = 2.4$. With this model, all covariates are prognostic and some are strongly correlated with the outcome.

Next, we consider more complex outcome-generating models that incorporate nonlinearity and treatment effect heterogeneity (interaction between treatment and covariates). For these models, we only consider six ($P = 6$) covariates, which
TABLE 2 Specification of key parameters in the outcome-generating models.

| Generating model | Incidence level | Number of covariates (P) | \( \beta_0 \) | \( \beta_Z \) | \( (P_1, P_0) \) | \( \Delta \) |
|------------------|-----------------|--------------------------|--------|--------|-----------------|--------|
| Outcome 1        | Low             | 6                        | -3.6   | -1.2   | (0.0455, 0.0987) | -0.8317|
|                  |                 |                          | -4.7   | -1.1   | (0.0224, 0.0490) | -0.8103|
| Outcome 1        | Low             | 15                       | -4.2   | -1.2   | (0.0484, 0.0954) | -0.7292|
|                  |                 |                          | -5.4   | -1.2   | (0.0221, 0.0486) | -0.8155|
| Outcome 2        | Low             | 6                        | -6.4   | -1.8   | (0.0490, 0.0974) | -0.7392|
|                  |                 |                          | -8.1   | -1.7   | (0.0240, 0.0507) | -0.7756|
| Outcome 2        | Low             | 15                       | -9.0   | -2.4   | (0.0559, 0.1045) | -0.6785|
|                  |                 |                          | -11.8  | -2.2   | (0.0254, 0.0499) | -0.7007|
| Outcome 3        | Low             | 6                        | -4.8   | -2.8   | (0.0507, 0.0999) | -0.7315|
|                  |                 |                          | -6.6   | -4.2   | (0.0252, 0.0505) | -0.7214|
| Outcome 4        | Low             | 6                        | -4.9   | -3.0   | (0.0501, 0.1001) | -0.7462|
|                  |                 |                          | -6.6   | -3.2   | (0.0246, 0.0487) | -0.7080|

are simulated from a multivariable model with mean vector 0 and covariance matrix for which diagonal elements are 1 and off diagonal elements are 0.1, reflecting weak correlations among the covariates.

(iii) Outcome-generating model 3: 

\[
Y^{c}_{ij} = \beta_0 - \frac{3}{1+\exp(-6X^{(1)}_i + X^{(2)}_i + X^{(3)}_i + X^{(4)}_i)} + \frac{1}{2}(X^{(5)}_i + X^{(6)}_i) + 2X^{(5)}_i X^{(6)}_i + 1.8(X^{(3)}_i + X^{(4)}_i)Z_i - \frac{2}{1+\exp(-4X^{(5)}_i + X^{(6)}_i)}Z_i + \beta_Z Z_i + u_i + \epsilon_{ij}.
\]

(iv) Outcome-generating model 4: 

\[
Y^{c}_{ij} = \beta_0 - \frac{3}{2(1+\exp(-4X^{(1)}_i + X^{(2)}_i))} + 2\sin(X^{(3)}_i + X^{(4)}_i) + 1.8(X^{(1)}_i X^{(3)}_i + X^{(2)}_i X^{(4)}_i) + X^{(5)}_i + X^{(6)}_i - 1.5X^{(5)}_i X^{(6)}_i - 1.5(X^{(3)}_i + X^{(4)}_i)Z_i + \frac{2}{1+\exp(-2X^{(5)}_i + X^{(6)}_i)}Z_i + \beta_Z Z_i + u_i + \epsilon_{ij}.
\]

Further, the parameters \( \beta_0 \) and \( \beta_Z \) for each combination of generating model, incidence level, and the number of covariates are set at values that give the desired incidences. These more granular considerations on parameter specification are summarized in Table 2.

3.3 | Estimands

To address the issue that regression coefficients of different modeling strategies can correspond to different parameters, we have articulated a common, nonparametric causal estimand of interest in Section 2.1—the participant-average treatment effect on the log odds ratio scale. Note that the main effect of treatment in the data-generating model has a conditional interpretation due to other covariates. The “true” participant-average treatment effect \( \Delta \) for each setting is based on the log odds ratio calculated from a simulated population with \( N = 5000, m = 100 \). Specifically, each individual has two potential outcomes, which are obtained by taking draws from the Bernoulli distribution with probability given by the outcome-generating model under \( Z = 1 \) and \( Z = 0 \), respectively. From these, the population incidences under treatment and under control are obtained, so that the true parameter value is \( \Delta \approx \log(P_1(1 - P_0)) - \log(P_0(1 - P_1)) \), where \( P_1 \) is the large sample (simulated population) incidence under treatment and \( P_0 \) is the large sample (simulated population) incidence under control.

3.4 | Methods: Analytical strategies with and without baseline individual-level covariates

We consider two modeling approaches for obtaining propensity scores indicating the estimated probability of being in the treatment group conditional on the subject’s baseline covariates. The first is the multivariable logistic model that regresses the cluster-level treatment variable on the main effects of the covariates. We next employ a more flexible BART model
with a probit link (Chipman et al., 2010) to test if a more flexible propensity score model can potentially adjust for chance imbalances on higher moments of the covariates beyond the mean. Specifically, we intend to test whether using BART for estimating propensity scores provides improvements over traditional parametric propensity scores for individual-level covariate adjustment in CRTs and to identify settings in which that is the case. Once the individual propensity score values are estimated, we construct two types of weight matrices based on IPW and OW. We consider six different models and evaluate their respective performances.

**Crude:** The crude model is our reference model in which there is no adjustment for individual-level covariates, and only the treatment indicator is included. The model is \( \logit \{ P(Y_{ij} = 1) \} = \beta_0 + \beta_1 Z_i \).

**Multi:** The multivariable model involves individual-level covariate adjustment and includes the main effects of the covariates along with the treatment indicator, given by \( \logit \{ P(Y_{ij} = 1) \} = \beta_0 + \sum_{p=1}^{P} \beta_p X_{ij}^{(p)} + \beta_1 Z_i \).

**IPW-Logit, IPW-BART, OW-Logit, OW-BART:** The individual-level propensity scores are estimated with logistic regression and BART, inverse probability or overlap weights based on the estimated propensity scores are calculated, and then the resulting weight matrix \( W_i \) is included in the GEE model for estimation and inference.

For the **Crude** approach, we consider the robust sandwich variance estimator and its bias-corrected versions as in P. Li and Redden (2015). For the **Multi** approach, we consider the sandwich variance estimators described in Table 1 (panel b), and for the propensity score weighting approaches we consider the sandwich variance estimators defined in Table 1 (panel a). We use the \( R \) statistical software to estimate propensity scores and perform regression analysis. Specifically, the dbarts package is used to implement BART and the geepack package is used to fit the GEE models. We developed our own code for computing the sandwich variance estimators in Table 1, and the \( R \) code is available in the Supporting Information.

### 3.5 Performance measures

We report five performance metrics for each combination of simulation setting and analytic approach to compare the relative performances of the methods considered. At each replication \( r \), we obtain an estimate of the replication-specific participant-average treatment effect, \( \hat{\Delta}_r \). Then over the 1000 replications, we can obtain an estimate of the true participant-average treatment effect, \( \Delta = \sum_{r=1}^{1000} \hat{\Delta}_r / 1000 \). Bias is calculated as the mean difference in each estimate and the true effect value, \( \text{Bias} = \sum_{r=1}^{1000} (\hat{\Delta}_r - \Delta) / 1000 \). To determine whether individual-level covariate adjustment provides efficiency gains over the unadjusted model, we present the relative efficiency (RE), which is the ratio of the empirical variance of the crude model to the empirical variance of the regression or propensity score weighting approaches. Further, for each replication, we construct a 95% normality-based confidence interval. Then the coverage using standard error estimates from the robust and bias-corrected estimators is obtained as the proportion of intervals across the replications that includes the true estimand value, \( \Delta \). Based on the binomial model with 1000 replications, we consider the coverage between \([93.6\%, 96.4\%]\) to be nominal, and, in general, higher coverage represents conservative performance which is typically more tolerable than lower coverage (as a reflection of the sandwich variance estimator being anticonservative). Here, we consider 1000 simulations sufficient for evaluation of coverage because the associated Monte Carlo standard error ranges from 0.7% (when the empirical coverage is 95%) to 1.3% (when the empirical coverage is as low as 80%) (Morris et al., 2019). Lastly, to get a sense of how often separation issues arise for each model, we report the nonconvergence proportion (Non-Con), which refers to the proportion of replications that resulted in an error when fitting the model; this metric is particularly relevant as we assume a low incidence binary outcome, and success in model fitting represents a common practical issue for analyzing such data.

### 4 SIMULATION RESULTS

Performance measures for each combination of the GEE model, variance estimator, and trial parameters are provided in detail in the supplementary tables. To keep the main illustration simple and concise, we focus on the settings with an average cluster size of 100 and ICC of 0.01 since the patterns and results do not vary much for the remaining combinations of simulation parameters.
FIGURE 1 Measures of relative efficiency for four different simulation settings with an average cluster size of 100 and ICC of 0.01 and assuming low incidence. BART, Bayesian additive regression trees; IPW, Inverse probability of treatment weighting; OW, overlap weighting.

4.1 Outcome-generating model 1: Additive effect model that includes weakly prognostic and nonprognostic variables

**Efficiency:** The relative efficiency of the covariate adjustment methods as compared to that of the crude model for data generated using outcome-generating model 1 is close to 1 across the number of clusters considered (Figure 1 and Supplementary Figure 1). This is also the case for the scenarios with other values of average cluster size \(m\), outcome ICC \(\rho_{\text{Logit}}\), and the total number of covariates \(P\). Thus, in this setting, there is limited efficiency gain and, in some cases, slightly less efficiency from covariate adjustment if the included variables are unrelated or weakly related to the outcome, but the number of clusters is limited.

**Coverage of confidence interval estimator:** Figure 2 and Supplementary Figure 2 present the coverage rates from 95% confidence intervals for varying the total number of clusters under outcome-generating model 1 and assumed low and very low incidences, respectively. For the crude and multivariable models, most of the estimators result in undercoverage for fewer than 30 clusters with the MD bias-corrected sandwich variance estimator giving closest to nominal. Under IPW and OW, the MD bias-corrected sandwich variance estimator generally provides nominal coverage except for the slight undercoverage observed when the number of clusters \(N = 6\) with logistic propensity score models. Overall, even with \(N = 30\) clusters, the uncorrected robust sandwich estimator may still result in undercoverage, while both bias-corrected estimators give similar coverage rates that are at the nominal level, regardless of whether individual-level covariate adjustment is considered.

**Convergence:** Nonconvergence tends not to be much of an issue when outcome variables are simulated under outcome-generating model 1, where there are six covariates except at very low incidences and \(N = 6\) and \(m = 30\). Here, the nonconvergence proportion is 14.7% and 16.0% for the multivariable model and 13.6% and 15.1% for the crude and propensity score weighted models at ICC of 0.001 and 0.01, respectively (Supplementary Table 2). If 15 covariates are adjusted for, approximately half of the replications do not converge for the multivariable model at six clusters with 30 subjects per cluster on average (Supplementary Table 4). However, weighting-based analysis encounters much less nonconvergence (Supplementary Tables 3 and 4) and is a practical solution for individual-level covariate adjustment when multivariable regression fails to provide a point estimate.
FIGURE 2 Measures of coverage for simulations based on the outcome-generating model 1, low outcome incidences, six covariates, average cluster size of 100, and latent ICC of 0.01. Values that are between the dashed lines (at 93.6% and 96.4%) are considered nominal. BART, Bayesian additive regression trees; IPW, Inverse probability of treatment weighting; KC, Kauermann and Carroll; MD, Mancl and DeRouen; OW, overlap weighting.

4.2 Outcome-generating model 2: Additive effect model that includes variables that are strongly correlated with the outcome

**Efficiency:** Adjustment for covariates that have a large effect on the outcome may substantially increase the efficiency of the model over the crude model (Figure 1 and Supplementary Figure 1). Further, the multivariable model, when it converges, tends to provide the largest efficiency gain. This is expected because the multivariable model is approximately correctly specified. For the same propensity score estimates (either with logistic regression or BART), IPW and OW provide almost identical estimates of empirical variance. Results are similar for other values of average cluster size ($m$) and outcome ICC ($\rho_{Logit}$). This result is in contrast to those under small individually randomized trials, where OW dominated IPW in terms of performance (Zeng et al., 2021).

**Coverage of confidence interval estimator:** In Supplementary Figures 3 and 4, the coverage based on 95% confidence intervals by the number of clusters under outcome-generating model 2 and for low and very low incidences, respectively, are given. When the crude and multivariable models were used, all variance estimators result in under-coverage for $N = 6$, 10 clusters, although the MD bias-corrected sandwich variance estimator is closest to nominal. Under IPW and OW, the MD bias-corrected sandwich variance estimator consistently leads to nominal coverage. However, once $N = 20$ clusters are reached, coverage becomes close to the nominal 0.95 for all estimators based on propensity score weighting.

**Convergence:** Even at low incidences, the multivariable analysis tends to have larger nonconvergence proportions than the crude and propensity score weighted models. Specifically, with six covariates under $N = 6$ and $m = 30$ at very low incidences, the multivariable model does not converge for 33.3% and 30.9% of the replications when the ICC is 0.001 and 0.01, respectively—almost tripling that of the other models (Supplementary Table 6). The problem becomes quite serious when 15 covariates are adjusted for. The analogous nonconvergence proportions are 77.4% and 76.8% when incidences are low (Supplementary Table 7) and 97.6% and 98.6% when incidences are very low (Supplementary Table 8). When $N = 10$ and $m = 30$, the nonconvergence proportions are 79.3% and 82.6% when the ICC is 0.001 and 0.01, respectively (Supplementary Table 8). Further, the multivariable model with 15 covariates is observed to have at least one nonconvergence replication even when we have as large as $N = 30$ clusters. These
findings show that, although the multivariable adjustment GEE often provides the largest efficiency gain, it could exhibit serious nonconvergence issues when covariates are strongly prognostic and when the number of clusters is limited. In those scenarios, propensity score weighting, both IPW and OW, becomes a practical solution that offers a moderate precision gain over the unadjusted analysis. When the outcome-generating model is relatively simple and additive, machine learning propensity score models do not exhibit any apparent advantage over logistic propensity scores.

4.3 Outcome-generating models 3 and 4: Nonlinear covariate–outcome associations with treatment effect heterogeneity explained by individual-level covariates

**Efficiency:** Under outcome-generating model 3, the multivariable analysis gives the largest RE. Further, weighting using BART-estimated propensity scores provides a larger RE than weighting based on logistic-estimated propensity scores, demonstrating that machine learning estimated propensity scores lead to an efficiency advantage over their parametric counterparts when the true outcome model includes complex covariate–outcome associations. Under outcome-generating model 4, weighting by BART-estimated propensity scores results in higher efficiency compared to the multivariable analysis. This is somewhat expected because the multivariate analysis does not include correctly specified functional forms of the individual-level covariates. However, this result represents an important piece of evidence that machine learning propensity scores can confer an efficiency advantage in analyzing small CRTs. Further, weighting using BART-estimated propensity scores gives higher RE as the number of clusters increases.

*Coverage of confidence interval estimator:* Under outcome-generating model 3 (Supplementary Figures 5 and 6), with propensity score weighting, all the sandwich variance estimators, including the robust estimator, gives coverage that is at least 0.90 starting with \( N = 10 \) clusters. IPW with BART-estimated propensity scores and OW for both logistic and BART-estimated propensity scores provide coverage close to 0.95 regardless of the sandwich variance estimator as long as the number of clusters is at least 10. This is in contrast with the multivariable model, which still results in undercoverage at \( N = 10 \) for the KC bias-corrected sandwich variance estimator. This set of results highlights the important benefit of propensity score weighting as a simple and effective covariate adjustment strategy when the true data-generating model is complex. Under outcome-generating model 4 (Supplementary Figures 7 and 8), patterns in coverage are generally similar to those observed under outcome-generating model 3.

**Convergence:** For these more complex outcome-generating models, the nonconvergence proportions from the multivariable analyses remain worse than the weighting approaches at a very small number of clusters (\( N = 6 \)). The results are summarized in Supplementary Tables 9–12.

5 APPLICATION TO THE RESTORE CLUSTER RANDOMIZED TRIAL

To illustrate the methods for analyzing CRTs with a low incidence outcome, we apply the propensity score weighting and multivariable regression methods to the Randomized Evaluation of Sedation Titration for Respiratory Failure (RESTORE) trial, which was a CRT that took place in \( N = 31 \) U.S. pediatric intensive care units (PICUs). RESTORE compared a nurse-implemented, goal-directed sedation protocol against usual care. The intervention was introduced to 17 PICUs from the Pediatric Acute Lung Injury and Sepsis Investigators (PALISI) Network, while 14 others in this network comprised the control clusters and were given usual care. The total sample size was 2449 children. Additional details about this study may be found in Curley et al. (2015).

We consider a secondary binary outcome that appeared to have clinically relevant effects—the postextubation stridor outcome (\( \hat{P}_1 = 0.072, \hat{P}_0 = 0.045 \)). For each of the individual-level covariate adjustment methods, we consider two sets of individual-level covariates corresponding to a small and a large number of variables, all of which are expected to be prognostic, though to varying degrees. The two sets of covariates are: (1) three individual-level covariates: age, PRISM III-12 score, and whether baseline Pediatric Overall Performance Category (POPC) score is equal to 1; and (2) 11 individual-level covariates: age, PRISM III-12 score, whether baseline POPC score is equal to 1, pneumonia as primary diagnosis, bronchiolitis as primary diagnosis, acute respiratory failure related to sepsis as primary diagnosis, oxygenation index, prematurity, asthma, cancer (current or previous diagnosis), and intubation at other hospital and transferred to participating PICU.

The baseline characteristics by treatment arm are summarized in Table 3. The absolute standardized difference (ASD)
TABLE 3  Baseline demographic and clinical characteristics of children who were mechanically ventilated for acute respiratory failure for control and intervention (use of RESTORE protocol) groups.

| Characteristics                  | Control sample size: 1224 | Intervention sample size: 1225 | ASD |
|----------------------------------|---------------------------|-------------------------------|-----|
| Age (mean (SD))                  | 5.21 (5.49)               | 4.22 (5.38)                   | 0.183 |
| PRISM score (mean (SD))          | 9.91 (7.50)               | 7.93 (7.32)                   | 0.266 |
| Baseline POPC score = 1 (%)      | 862 (70.4)                | 885 (72.2)                    | 0.040 |
| Pneumonia (%)                    | 433 (35.4)                | 394 (32.2)                    | 0.068 |
| Bronchiolitis (%)                | 228 (18.6)                | 428 (34.9)                    | 0.375 |
| Acute respiratory failure (%)    | 212 (17.3)                | 145 (11.8)                    | 0.156 |
| Oxygenation index (mean (SD))    | 8.27 (7.34)               | 8.16 (6.85)                   | 0.015 |
| Prematurity (%)                  | 175 (14.3)                | 194 (15.8)                    | 0.043 |
| Asthma (%)                       | 210 (17.2)                | 146 (11.9)                    | 0.149 |
| Cancer (%)                       | 109 (8.9)                 | 88 (7.2)                      | 0.063 |
| Transferred (%)                  | 306 (25.0)                | 334 (27.3)                    | 0.052 |

Abbreviations: ASD, absolute standardized difference; SD, standard deviation.

for each covariate is reported as a measure of imbalance between intervention and control arms. The values for age, PRISM score, bronchiolitis, acute respiratory failure, and asthma are greater than 0.10, which is a common threshold for assessing balance (Austin & Stuart, 2015), suggesting residual imbalance between the treatment arms for these variables. In Supplementary Table 13, we present the propensity score weighted covariate distributions based on the two adjustment sets of interest. With either IPW or OW, covariates that are included in the propensity score model had reduced baseline imbalance based on the weighted ASD, with all of them less than 0.10. With logistic-estimated propensity scores, OW completely removes the baseline chance imbalance for covariates that were included, due to its exact mean balance property (F. Li, Morgan, & Zaslavsky, 2018; Zeng et al., 2021).

Figure 3 presents the data analysis results, under both sets of adjustment variables, and on both the log OR and OR scales. An overall pattern is that the unadjusted analysis appears to provide a larger participant-average treatment effect compared to any covariate-adjusted analysis. This is possibly due to the moderate chance imbalance in baseline covariates with a limited number of clusters, which may exaggerate the treatment benefit. In our simulation results, the MD bias-corrected sandwich variance estimator frequently leads to nominal coverage for the participant-average treatment effect across methods for individual-level covariate adjustment, and, therefore, we only present the confidence intervals based on the MD bias-corrected sandwich variance estimator and the original robust sandwich variance (as a benchmark). On the log OR scale (which corresponds to the estimand in our simulations), the RE values (defined based on the MD bias-corrected sandwich variance estimates) between the covariate-adjusted estimators and the crude estimator are 0.875 (multivariable), 0.981 (IPW-Logit), 1.048 (IPW-BART), 0.970 (OW-Logit), and 1.049 (OW-BART) using covariate adjustment set 1. The analogous RE values using covariate adjustment set 2 are 0.833 (multivariable), 1.064 (IPW-Logit), 1.110 (IPW-BART), 1.081 (OW-Logit), and 1.135 (OW-BART). This indicates a slight efficiency advantage of the BART propensity score weighted estimators over the alternatives and is indicative of possible nonlinear covariate–outcome associations resembling the outcome-generating model 4 in our simulations. On the OR scale, individual-level covariate adjustment appears to bring more substantial efficiency gain as shown by the narrower confidence intervals as compared to those obtained based on the crude model. Furthermore, while the new sedation protocol is shown to significantly increase the odds of postextubation stridor, this effect is no longer statistically significant after variables are adjusted for.

6 | DISCUSSION

In this article, we have offered a balanced discussion on the benefits and limitations of propensity score weighting and multivariable adjustment methods, and we identified scenarios under which one of these methods may provide relatively higher statistical efficiency for treatment effect estimation as well as better convergence properties. Furthermore, we have investigated the small-sample inference properties of these estimators with bias-corrected sandwich variance estimators. In what follows, we organize the discussion around three main properties of the competing estimators: efficiency, coverage, and convergence.
First, regarding efficiency, our results suggest that individual-level covariate adjustment tends to provide better performance than an unadjusted model, especially when the covariates are at least moderately prognostic. This finding parallels those made for individually randomized trials and reinforces the need for individual-level covariate adjustment in CRTs. However, the optimal individual-level covariate adjustment method can differ based on the underlying data-generating process (Figure 1). When it is successfully fit, the multivariable model provides the largest efficiency gain when covariates are strongly prognostic of outcome and the model is approximately correctly specified. On the other hand, propensity score weighting may be preferred when data complexities, such as interactions and/or a nonlinear response surface, are expected. In some cases, the increase in efficiency gain from propensity score weighting may be greater from multivariable regression with linear-term adjustment. For weighting estimators, the choice of weights (IPW vs. OW) did not have much effect on the performance. Rather, efficiency gains from propensity score weighting varied with the type of model used to estimate the propensity score in some settings. This is in contrast to Zeng et al. (2021), who found OW often dominates IPW in small individually randomized trials. The likely reason is that when the number of clusters is small, the total sample size remains at least moderate in CRTs. For choices of the propensity score model, BART-estimated propensity scores may provide more efficiency gain as compared to parametric propensity scores, such as logistic regression, when the true outcome data-generating model includes possible nonlinear interactions between covariates and treatment. Even though machine learning propensity score models have been shown to provide benefit in analyzing observational studies (Lee et al., 2010), to the best of our knowledge, this is the first study that demonstrates the utility of a machine learning propensity score model for covariate adjustment in CRTs. Of note, even though BART is inherently Bayesian, we have integrated its posterior mean estimates for the propensity scores into the GEE estimators and have pursued a final frequentist estimator. The consideration of this approach is based on practical utility, and a more formal Bayesian approach is left for future work.

Second, regarding coverage, we found that the MD bias-corrected sandwich variance estimator frequently provides nominal coverage for individual-level covariate-adjusted estimation of the participant-average treatment effect in CRTs.
Although the MD bias-corrected sandwich variance estimator has been shown to be overly conservative in previous comparative studies with no or few covariates, it gave approximately 95% coverage in our case, especially when there are very few clusters. We do acknowledge that, however, our findings are based on the z-confidence interval rather than the t-confidence interval, whereas previous studies, for example, have shown the t-test coupled with the KC bias-corrected sandwich variance estimator preserved nominal test size (P. Li & Redden, 2015) when the cluster size is not extremely variable. In our work, the basis for choosing the z-confidence interval is twofold. On the one hand, we realize that it may not be straightforward to determine the appropriate degrees of freedom for the t-confidence interval with propensity score weighting as technically the marginal mean model only includes a treatment indicator, but potentially many covariates can be adjusted for in the propensity score weights. On the other hand, introducing the t-confidence interval with a conservative choice of the degrees of freedom (such as the number of clusters minus the number of covariates adjusted for as in Mancl and DeRouen (2001)) may lead to lower power compared to the z-confidence interval, which is unfavorable in small CRTs. The optimal degrees of freedom under the t-distribution for different individual-level covariate-adjusted estimators and the extent to which the t-confidence interval would change our recommendations require further research.

Finally, regarding convergence, we find that if the outcome incidences are under 0.05, propensity score weighting may be preferred over multivariable regression, which is very likely to have separation and convergence issues. This is especially the case when a large number of individual-level covariates are adjusted for, but the number of clusters is limited. Despite the fact that multivariable regression may lead to the largest efficiency gain when the model is close to correctly specified, the choice between multivariable regression and propensity score weighting for individual-level covariate adjustment in small CRTs should also take into account convergence as a practical challenge. Overall, propensity score weighting represents a useful and feasible approach for individual-level covariate adjustment in CRTs with a rare binary outcome and when the number of clusters is small.

There are three limitations that we plan to address in future work. First, we have only considered the causal odds ratio as a target estimand. However, the methods are readily applicable to estimate the causal risk difference or relative risk, for example, by choosing the log link (F. Li & Tong, 2021b) for unadjusted and weighting-based estimators. It would be valuable to compare individual-level covariate adjustment strategies for estimating these two alternative effect measures. Second, we have restricted to rare binary outcomes with an incidence rate ranging from around 2–10%, which is similar to that in our motivating study; in this case, we have already observed nontrivial nonconvergence issues from fitting multivariable regression models with individual-level covariates. There are other cases with extremely rare outcomes, such as those with an incidence rate < 1% (Westgate et al., 2022). In that case, while we suspect the nonconvergence issue for multivariable regression models can only be more frequent, whether individual-level covariate adjustment can lead to efficiency gain is an open question. Third, we have considered the independence working correlation model, and, therefore, the recommendations are limited to independence GEE. As we show in Supplementary File Section 1, the choice of independence correlation structure allows us to interpret the treatment effect coefficient as our target causal estimand by giving each individual equal weight (Kahan et al., 2023). Since the exchangeable working correlation model is another popular choice in analyzing CRTs, it would be interesting to investigate to what extent the current findings can be generalizable to the exchangeable working correlation model.

ACKNOWLEDGMENTS

Research in this article was partially supported by the Patient-Centered Outcomes Research Institute® (PCORI® Awards ME-2020C1-19220 to M.O.H. and ME-2020C3-21072 to F.L.). M.O.H. is also funded by the U.S. National Institutes of Health (NIH), National Heart, Lung, and Blood Institute (grant number R00-HL141678). All statements in this report, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of the NIH or PCORI® or its Board of Governors or Methodology Committee. The authors thank the special issue editor and two anonymous reviewers for their helpful suggestions that have substantially improved this work. The authors also thank Can Meng for providing computational assistance during the revision of this work.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Data from the RESTORE trial was obtained from the U.S. National Heart, Lung and Blood Institute, Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC) (https://biolincc.nhlbi.nih.gov/home/). The authors
do not have the permission to share this data directly; however, the data are available to researchers upon request and approval from BioLINCC.

**OPEN RESEARCH BADGES**

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge “Reproducible Research” for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to confidentiality issues.

**ORCID**

Angela Y. Zhu https://orcid.org/0000-0001-5187-2667

Nandita Mitra https://orcid.org/0000-0002-7714-3910

Michael O. Harhay https://orcid.org/0000-0002-0553-674X

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

**REFERENCES**

Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, 34(28), 3661–3679.

Benkeser, D., Diaz, I., Luedtke, A., Segal, J., Scharfstein, D., & Rosenblum, M. (2021). Improving precision and power in randomized trials for covid-19 treatments using covariate adjustment, for binary, ordinal, and time-to-event outcomes. *Biometrics*, 77(4), 1467–1481.

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1), 266–298.

Curley, M. A., Wypij, D., Watson, R. S., Grant, M. J. C., Asaro, L. A., Cheifetz, I. M., Dodson, B. L., Franck, L. S., Gedeit, R. G., Angus, D. C., Matthay, M. A., & for the RESTORE Study Investigators and the Pediatric Acute Lung Injury and Sepsis Investigators (PALISI) Network. (2015). Protocolized sedation vs usual care in pediatric patients mechanically ventilated for acute respiratory failure: A randomized clinical trial. *JAMA*, 313(4), 379–389.

Dorie, V., Hill, J., Shalit, U., Scott, M., & Cervone, D. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1), 43–68.

Fay, M. P., & Graubard, B. I. (2001). Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics*, 57(4), 1198–1206.

Ford, W. P., & Westgate, P. M. (2017). Improved standard error estimator for maintaining the validity of inference in cluster randomized trials with a small number of clusters. *Biometrical Journal*, 59(3), 478–495.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1), 217–240.

Ivers, N., Taljaard, M., Dixon, S., Bennett, C., McAteer, A., Tsertsvadze, L., Agarwal, N.,刻osaurdas, L., O’Brien, K., & Li, F., & Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521), 390–400.

Kahan, B., Li, F., Tong, G., & Zaslavsky, A. M. (2018). Sample size estimation for modified Poisson analysis of cluster randomized trials with a binary outcome. *Statistical Methods in Medical Research*, 30(5), 1288–1305.
Li, F., Turner, E. L., Heagerty, P. J., Murray, D. M., Vollmer, W. M., & Delong, E. R. (2017). An evaluation of constrained randomization for the design and analysis of group-randomized trials with binary outcomes. *Statistics in Medicine*, 36, 3791–3806.

Li, F., Turner, E. L., & Preisser, J. S. (2018). Sample size determination for GEE analyses of stepped wedge cluster randomized trials. *Biometrics*, 74(4), 1450–1458.

Li, P., & Redden, D. T. (2015). Small sample performance of bias-corrected sandwich estimators for cluster-randomized trials with binary outcomes. *Statistics in Medicine*, 34(2), 281–296.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13–22.

Luo, B., Preisser, J. S., Qaqish, B. F., Suchindran, C., Bangdiwala, S. I., & Wolfson, M. (2007). A comparison of two bias-corrected covariance estimators for generalized estimating equations. *Biometrics*, 63(3), 935–941.

Manci, L. A., & De Rouen, T. A. (2001). A covariance estimator for GEE with improved small-sample properties. *Biometrics*, 57(1), 126–134.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11), 2074–2102.

Murray, D. M. (1998). *Design and analysis of group-randomized trials*. Oxford University Press.

Murray, D. M., Varnell, S. P., & Blitstein, J. L. (2004). Design and analysis of group-randomized trials: A review of recent methodological developments. *American Journal of Public Health*, 94(3), 423–432.

Preisser, J. S., Young, M. L., Zaccaro, D. J., & Wolfson, M. (2003). An integrated population-averaged approach to the design, analysis and sample size determination of cluster-unit trials. *Statistics in Medicine*, 22(8), 1235–1254.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55.

Scott, J. M., deCamp, A., Juraska, M., Fay, M. P., & Gilbert, P. B. (2017). Finite-sample corrected generalized estimating equation of population average treatment effects in stepped wedge cluster randomized trials. *Statistical Methods in Medical Research*, 26(2), 583–597.

Su, F., & Ding, P. (2021). Model-assisted analyses of cluster-randomized experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(5), 994–1015.

Turner, E. L., Li, F., Gallis, J. A., Prague, M., & Murray, D. M. (2017). Review of recent methodological developments in group-randomized trials: Part I- design. *American Journal of Public Health*, 107(6), 907–915.

Turner, E. L., Prague, M., Gallis, J. A., Li, F., & Murray, D. M. (2017). Review of recent methodological developments in group-randomized trials: Part 2- analysis. *American Journal of Public Health*, 107(7), 1078–1086.

Wang, X., Turner, E. L., Li, F., Wang, R., Moyer, J., Cook, A. J., Murray, D. M., & 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.

Wang, X., Turner, E. L., Preisser, J. S., & Li, F. (2022). Power considerations for generalized estimating equations analyses of four-level cluster randomized trials. *Biometrical Journal*, 64(4), 663–680.

Watson, S. I., Girling, A., & Hemming, K. (2021). Design and analysis of three-arm parallel cluster randomized trials with small numbers of clusters. *Statistics in Medicine*, 40(5), 1133–1146.

Westgate, P. M. (2013). On small-sample inference in group randomized trials with binary outcomes and cluster-level covariates. *Biometrical Journal*, 55(5), 789–806.

Westgate, P. M., Cheng, D. M., Feaster, D. J., Fernández, S., Shoben, A. B., & Vandergrift, N. (2022). Marginal modeling in community randomized trials with rare events: Utilization of the negative binomial regression model. *Clinical Trials*, 19(2), 162–171.

Westreich, D., Lessler, J., & Funk, M. J. (2010). Propensity score estimation: Neural networks, support vector machines, decision trees (cart), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology*, 63(8), 826–833.

Williamson, E. J., Forbes, A., & White, I. R. (2014). Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Statistics in Medicine*, 33(5), 721–737.

Zeng, S., Li, F., Wang, R., & Li, F. (2021). Propensity score weighting for covariate adjustment in randomized clinical trials. *Statistics in Medicine*, 40(4), 842–858.

Zhou, Y., Turner, E. L., Simmons, R. A., & Li, F. (2022). Constrained randomization and statistical inference for multi-arm parallel cluster randomized controlled trials. *Statistics in Medicine*, 41(10), 1862–1883.

Zhu, Y., Hubbard, R. A., Chubak, J., Roy, J., & Mitra, N. (2021). Core concepts in pharmacoepidemiology: Violations of the positivity assumption in the causal analysis of observational data: Consequences and statistical approaches. *Pharmacoepidemiology and Drug Safety*, 30(11), 1471–1485.

**SUPPORTING INFORMATION**

Additional supporting information can be found online in the Supporting Information section at the end of this article.

---

**How to cite this article:** Zhu, A. Y., Mitra, N., Hemming, K., Harhay, M. O., & Li, F. (2024). Leveraging baseline covariates to analyze small cluster-randomized trials with a rare binary outcome. *Biometrical Journal*, 66, 2200135. https://doi.org/10.1002/bimj.202200135