stats

MDPI

Article

A Simulation-Based Evaluation of the DR-GEE Approach Based on Flexible Cluster-Size Weighting Under Hybrid Informative Cluster Size Structures

Betül Dağoğlu Hark [1,*] and Zeliha Nazan Alparslan [2]

[1] Department of Biostatistics, Faculty of Medicine, Fırat University, 23119 Elazığ, Turkey
[2] Department of Biostatistics, Faculty of Medicine, Çukurova University, 01330 Adana, Turkey
* Correspondence: bdagoglu@firat.edu.tr

Abstract

This study evaluates a DR-GEE approach based on Doubly Robust Generalized Estimating Equations for marginal inference under a hybrid informative cluster size structure. Hybrid informative cluster size refers to situations where cluster size can be related to both the marginal response variable and the distribution of covariates associated with the response. The proposed approach aims to achieve a cluster-balanced marginal estimator designed to mitigate the excessive influence of cluster sizes. To this end, the method combines a flexible cluster-size weighting component, defined by the α adjustment parameter, with an augmentation term derived from the study's outcome model. Thus, both the direct cluster size–outcome relationship and the imbalance in the distribution of covariates are taken into account. A comprehensive Monte Carlo simulation evaluated performance under varying informativeness levels (γ = 0.1, 0.5, 1.0) and average cluster sizes (λ = 3, 5, 8). DR-GEE was compared with standard GEE, CWGEE, WCR, and DWGEE across different tuning parameters (α = 0.25, 0.50, 1.0). The results show that DR-GEE with α = 1 generally achieved the most favorable performance under hybrid ICS conditions. For example, when γ = 1 and λ = 5, GEE exhibited substantial bias (0.538) and high RMSE (0.566), whereas DR-GEE (α = 1) markedly reduced bias (0.076) and RMSE (0.224). Unlike CWGEE and DWGEE, which address only one dimension of informativeness, DR-GEE balances both cluster size–outcome dependence and covariate information.

Keywords: Informative Cluster Size (ICS); hybrid informativeness; doubly robust estimation; Generalized Estimating Equations (GEE); flexible cluster-size weighting; covariate-outcome dependence

Check for updates

Academic Editor: Wei Zhu

Received: 14 May 2026
Revised: 4 July 2026
Accepted: 9 July 2026
Published: 12 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The Informative Cluster Size (ICS) problem in clustered data defines situations where cluster size (the number of units in each cluster) is related to the outcome variable. In other words, if the average outcomes of a cluster change systematically as the number of observations in that cluster increases or decreases, then the cluster size is considered informative [1].

This situation can lead to significant biases in standard statistical analyses, as classical methods generally assume that cluster size is random (independent of the outcome). For example, if the average birth weight differs among offspring with higher litter sizes (cluster sizes), or if an individual with fewer female offspring has poorer oral health, then the

cluster size contains information relevant to the outcome. When ICS is ignored, cluster-internal dependence and sampling structure may be incorrectly handled, leading to bias in estimates. Therefore, special methodological approaches that account for ICS have been developed [1–4].

To accommodate this problem, Hoffman et al. (2001) proposed the within-cluster resampling (WCR) method, which is based on analyzing repeated datasets containing a randomly selected observation from each cluster and taking the mean of the estimates [5]. Although WCR yields unbiased estimation under ICS, it is computationally intensive and often suffers from substantial variance due to the repeated subsampling mechanism.

To reduce the computational burden of the WCR, Williamson et al. (2003) proposed the Cluster-Weighted Generalized Estimating Equations (CWGEE) approach, which uses inverse cluster size weights [2]. By aiming to balance the total contribution of each cluster, CWGEE reduces the dominant influence of large clusters on the estimate and offers a computationally more feasible alternative compared to WCR. However, the use of inverse cluster size weights can increase the sensitivity of estimates to weight selection, particularly when cluster sizes are highly unbalanced. In this case, while small clusters receive relatively higher contributions, the influence of large clusters is significantly reduced; this can lead to increased variance or estimation instability in certain data structures. Furthermore, CWGEE was developed primarily to address cluster size informativeness directly; however, in hybrid structures where informativeness also arises through the joint distribution of variables, additional modeling components may be required.

A more comprehensive conceptualization of ICS was proposed by Huang and Leroux [6]. The authors demonstrated that informativeness may stem not only from the direct relationship between cluster size and the outcome variable, but also from the relationship between cluster size and the distribution of covariates associated with the outcome at the subcluster level [6]. Within this framework, the proposed double-weighted GEE (DWGEE) approach offers a robust structure by combining the outcome regression model with a covariate-based weighting model. In other words, under appropriate assumptions, it can allow for consistent marginal inference provided that at least one of the outcome model or the covariate-based weighting model is correctly specified.

However, the Huang–Leroux approach primarily focuses on the information structures arising from the distribution of covariates. Therefore, in hybrid ICS structures where cluster size is directly related to the outcome variable or where informativeness arises from both the direct cluster size–outcome relationship and the covariate distribution, an additional cluster size adjustment component may be required.

The robust estimation systems provide a strong theoretical framework for solving these complex structures. The foundation of double-robust (DR) estimators is based on the literature on missing data [7]. For longitudinal data, Van der Laan and Robins (2003) proposed a DR estimator for monotonically deficient responses [8]; Bang and Robins (2005) demonstrated how to perform DR inference on longitudinal data with missing values [9]; and Seaman and Copas (2009) developed a double-robust generalized estimation equation (DR-GEE) for longitudinal data with missing values [10]. Prague et al. (2016) adapted this approach to clustered data [11]. Furthermore, Rabideau et al. (2024) recently proposed the idea of simultaneously managing multiple confounding and informative mechanisms [12]. Inspired by these theoretical frameworks, our current study integrates the robust estimation system into the context of flexible cluster-size weighting. In this regard, this study evaluates the DR-GEE approach, which addresses cluster size adjustment in such hybrid structures in conjunction with an outcome-model-based expansion component.

While the WCR, CWGEE, and DWGEE methods have made significant contributions to the ICS literature, further methodological refinements are needed for hybrid ICS structures,

where informativeness stems both from the direct relationship between cluster size and the outcome and from the relationship between the distribution of cluster size and the distribution of covariates associated with the outcome. In particular, computationally feasible approaches that can account for imbalances in the distribution of covariates while mitigating the dominant effect of cluster size on the estimate are limited.

This study evaluates the DR-GEE approach, which aims to perform marginal inference under hybrid ICS structures where there is bidirectional information flow between cluster size and the distributions of both the outcome variable and the covariates. The proposed method incorporates the direct effect of cluster size—which is typically ignored in classical GEE approaches—into the model via a flexible weighting mechanism, while also offering the advantage of double robustness through a correction term based on the outcome model. Thus, it aims to reduce systematic bias associated with cluster size and to obtain more consistent and cluster-balanced marginal parameter estimates in clustered data structures with high heterogeneity. In this regard, the study addresses the limitations of existing methods—such as ICS-sensitive WCR, CWGEE, and DWGEE—in a complementary manner, offering a computationally feasible alternative approach that can fill the methodological gap, particularly in mixed ICS scenarios. In addition, to comprehensively evaluate the performance of the proposed DR-GEE approach, we also considered simulation scenarios that included only outcome-dependent cluster size (pure) and covariate-dependent informative cluster size mechanisms.

2. Materials and Methods

2.1. GEE Approaches Under the Assumption of ICS

An important topic addressed in the clustered data literature is the concept of ICS. ICS (also referred to as non-ignorable cluster size) refers to a dependence between the outcome variable and the size of the cluster [13]. In formal terms, ICS arises when the $E(Y_{ij}|n_i, \boldsymbol{X}_{ij}) = E(Y_{ij}|\boldsymbol{X}_{ij})$ is violated [2–4]. Here, Y_{ij} denotes the marginal outcome of interest for the j-th subunit $(j = 1, \ldots, n_i)$ within the i-th cluster $(i = 1, \ldots, m)$, $\boldsymbol{X}_{ij}$ represents the vector of subunit-level and cluster-level covariates, and n_i signifies the cluster size (the total number of subunits within cluster i). The expectation $E[.]$ is taken with respect to the true underlying joint distribution of the data. Under this violation, the cluster size n_i carries explicit information about the marginal probability of the outcome, rendering standard unweighted estimation techniques biased.

This phenomenon is frequently encountered in reproductive toxicology studies, where a toxic substance is administered to a maternal rodent, and the birth defects in its offspring are observed. Due to increased fetal resorptions in litters with fewer pups, a greater proportion of birth defects may be observed, thereby introducing dependence between the number of offspring and the outcome. This exemplifies the notion of ICS.

In a study on periodontal health, when examining the relationship between risk factors (such as age, oral hygiene, smoking habits, and plaque) and tooth loss, the dependence between the number of teeth and tooth loss also introduces ICS [2].

Two principal approaches have been proposed for marginal modeling in the presence of ICS. The first is the WCR approach introduced by Hoffman et al. (2001) [5]. Due to its computational intensity, however, Williamson et al. (2003) later proposed an updated methodology known as the CWGEE approach [2].

2.1.1. Within-Cluster Resampling (WCR)

The number of datasets generated via Monte Carlo resampling is denoted by Q. For each resampled dataset, $\hat{\boldsymbol{\beta}}_q$ and $\hat{\boldsymbol{\Sigma}}_q$ represent the estimated parameter vector and its corresponding covariance matrix, respectively. The WCR estimator of $\boldsymbol{\beta}$ is then defined as:

$$\hat{\boldsymbol{\beta}}_{wcr} = Q^{-1} \sum_{q=1}^{Q} \hat{\boldsymbol{\beta}}_q$$

The asymptotic variance-covariance estimator of $\sqrt{N}\left(\hat{\boldsymbol{\beta}}_{wcr} - \boldsymbol{\beta}\right)$ is given by [3,5]:

$$\hat{\boldsymbol{V}}_{wcr} = \frac{1}{Q} \sum_{q=1}^{Q} \hat{\boldsymbol{\Sigma}}_q - \frac{1}{Q} \sum_{q=1}^{Q} \left(\hat{\boldsymbol{\beta}}_q - \hat{\boldsymbol{\beta}}_{wcr}\right)\left(\hat{\boldsymbol{\beta}}_q - \hat{\boldsymbol{\beta}}_{wcr}\right)^T$$

2.1.2. Cluster-Weighted Generalized Estimating Equations (CWGEE)

The WCR estimation approach, while robust, is computationally intensive. To address this limitation, Williamson et al. proposed the CWGEE method, which is asymptotically equivalent to WCR. The CWGEE approach offers a more computationally efficient alternative by incorporating cluster-level weights directly into the generalized estimating equations framework. The CWGEE estimating equation is defined as:

$$\boldsymbol{U}_{\boldsymbol{cwgee}}(\boldsymbol{\beta}) = \sum_{i=1}^{m} \frac{1}{n_i} \sum_{j=1}^{n_i} D_{ij}^T V_{ij}^{-1} \left(Y_{ij} - \mu_{ij}(\boldsymbol{\beta})\right) = 0$$

where $D_{ij} = \partial \mu_{ij}(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}$ and V_{ij} represents the marginal variance of the outcome for the j-th subunit within the i-th cluster. The parameter estimate $\hat{\boldsymbol{\beta}}_{cwgee}$ is obtained by solving this single weighted estimating equation. The asymptotic variance-covariance matrix of $\hat{\boldsymbol{\beta}}_{cwgee}$ is estimated using the sandwich formula:

$$\hat{\boldsymbol{\mathcal{V}}} = \hat{\boldsymbol{H}}^{-1} \hat{\boldsymbol{V}} \hat{\boldsymbol{H}}^{-1}$$

Here,

$$\hat{\boldsymbol{H}} = N^{-1} \sum_{i=1}^{K} \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{\partial \boldsymbol{U}_{ij}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} |_{\beta=\hat{\beta}}$$

and

$$\hat{\boldsymbol{V}} = K^{-1} \sum_{i=1}^{m} \left\{ \frac{1}{n_i} \sum_{j=1}^{n_i} U_{ij}(\hat{\beta}) \right\} \left\{ \frac{1}{n_i} \sum_{j=1}^{n_i} U_{ij}(\hat{\beta}) \right\}^T$$

Provided that (i) the parameter space is compact, (ii) the covariates are bounded, (iii) the expected derivative of the score function is non-singular, and (iv) the cluster sizes n_i are independent and identically distributed with finite variance, it can be shown that both estimators converge in probability to the same true parameter vector as the number of clusters $K \to \infty$. Under these regularity conditions, the resampling-based estimator $\hat{\boldsymbol{\beta}}_{wcr}$ and the single-step CWGEE estimator $\hat{\boldsymbol{\beta}}_{cwgee}$ satisfy:

$$\hat{\boldsymbol{\beta}}_{wcr} - \hat{\boldsymbol{\beta}}_{cwgee} \xrightarrow{p} 0 \; as \; K \to \infty.$$

Both marginal model-based WCR and CWGEE are essential for providing valid and consistent estimates in the presence of informative cluster size [2].

Seaman et al. [14] stated that observed-cluster inference can be obtained by defining the weight differently in the CWGEE approach. In this approach, under the assumption

that losses occur with equal probability and $P(n_i \geq 1) = 1$, it is proposed to take the weight $R_j/P(R_j = 1 \,|\widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{Y}}, n_i) = R_j M/n_i$ [14].

2.1.3. Doubly Weighted GEE (DWGEE)

Huang and Leroux (2011) extended the weighted GEE framework with a double-weighted estimating equation to account for situations where cluster size may be related to the distribution of covariates at the subcluster level [6]. This approach is based on the idea that informativeness may stem not only from the direct relationship between cluster size and the outcome variable but also from systematic differences in the distribution of covariates associated with the outcome across clusters of different sizes. To address these potential sources of bias, it has been proposed to combine a cluster-level weight with a covariate-based individual-level weight within the same estimating equation.

Here, w_i represents the cluster-level weight and aims to balance the sampling contribution based on cluster size. ω_{ij} denotes the covariate-based individual-level weight and is used to account for the potential information bias arising from the relationship between cluster size and the distribution of the covariate.

$$U(\beta) = \sum_{i=1}^{m} w_i D_i V_i^{-1} (Y_i - \mu_i) + \sum_{i=1}^{m} \sum_{i=1}^{n_i} \omega_{ij} A_{ij} = 0$$

where

- $D_i = \partial \mu_i / \partial \beta$ is the derivative matrix,
- V_i is the working covariance matrix,
- A_{ij} is an augmentation term derived from a specified outcome regression model,
- and $\mu_i = g^{-1}(X_i \beta)$ is the marginal mean vector.

The construction of w_i and ω_{ij} is such that the estimator achieves double robustness, meaning that

$$\hat{\beta}_{DWGEE} \xrightarrow{p} \beta_0 \; if \; either \begin{cases} \text{(i) the outcome regression model is correctly specified,} \\ or \\ \text{(ii) the covariate} - \text{based weighting model is correctly specified} \end{cases}$$

Under appropriate assumptions, this property offers a two-sided robust structure that allows for consistent marginal inference provided that at least one of the outcome model or the covariate-based weighting model is correctly specified. Therefore, the Huang–Leroux formulation provides an important and theoretically robust framework for addressing the informativeness arising from the distribution of covariates.

However, the Huang–Leroux approach primarily focuses on information structures derived from covariates. Consequently, in hybrid ICS structures where information arises either directly from the cluster size–outcome relationship or through both the direct cluster size–outcome relationship and the covariate distribution, an additional cluster size adjustment component may be required. This distinction forms the rationale for the present study. The DR-GEE approach evaluated in this study aims to support cluster-balanced marginal inference under hybrid ICS conditions by combining flexible cluster-size weighting with an outcome-model-based extension component [6].

2.2. The Proposed Flexible Cluster-Size-Weighted DR-GEE Approach Within the Hybrid ICS Framework

Let i = 1, …, m index the clusters and j = 1, …, n_i index the observations within cluster i. Under hybrid ICS, informativeness may arise not only from the direct relationship between cluster size and the response variable but also from the relationship between

cluster size and the distribution of covariates associated with the response. This situation can be expressed as follows:

$$E(Y_{ij}|X_{ij}, n_i) \neq E(Y_{ij}|X_{ij}) \text{ and } E(X_{ij}|n_i) \neq E(X_{ij}) \quad (1)$$

The first inequality shows that cluster size may be directly related to the distribution of results; the second inequality shows that the distribution of the covariate may vary depending on cluster size. The coexistence of these two mechanisms can amplify the effect of large clusters on the estimate in the standard GEE approach and lead to a bias from the targeted cluster-balanced marginal estimator.

The CWGEE approach aims to reduce the dominant contribution of large clusters by using inverse cluster size weights. However, in hybrid ICS frameworks where informativeness also arises through the covariate distribution, cluster size weighting alone may not be sufficient. Therefore, this study evaluates a flexible cluster-size-weighted DR-GEE approach aimed at cluster-balanced marginal inference. The proposed approach combines the cluster-size adjustment component with an outcome-model-based augmentation term, aiming to address both the direct cluster-size–outcome relationship and the informativeness arising from the covariate distribution simultaneously.

2.2.1. Flexible Cluster-Size Weighting

To mitigate the influence of cluster size on the estimate and allow varying degrees of cluster-size adjustment, a flexible sample size weighting defined by the parameter α is used.

$$w_i(\alpha) = \frac{n_i^{-\alpha}}{E(n_i^{-\alpha})}, \quad (2)$$

where $\alpha \in [0, 1]$ controls the degree of size adjustment [4]:

- $\alpha = 1$: inverse-size weighting (CWGEE)
- $\alpha = 0$: no weighting (standard GEE)
- $0 < \alpha < 1$: partial cluster-size adjustment

Normalization preserves the overall scale of the estimating equation by ensuring that $E\{w_i(\alpha)\} = 1$. Rather than completely eliminating the increase in variance, this normalization helps reduce scale-induced instability in cases where cluster sizes are highly heterogeneous. The weight in Equation (2) adjusts each cluster's contribution according to its size and aims to reduce bias that could arise directly from the cluster size–outcome relationship. This cluster size weighting component is conceptually distinct from the covariate-based weighting proposed by Huang and Leroux (2011), which primarily addresses the informativeness arising from covariate distributions at the subcluster level [6].

2.2.2. Augmentation Component for Outcome Modeling

To account for the information content arising from the joint distribution of the variables and its effect on the response variable, an augmentation term based on a parametric response model is included:

$$m_{ij}(X_{ij}; \gamma) = E(Y_{ij}|X_{ij}, n_i; \gamma) \quad (3)$$

The augmentation term is defined as:

$$A_{ij}(\gamma) = D_{ij} V_{ij}^{-1} \{Y_{ij} - m_{ij}(X_{ij}; \gamma)\}, \quad (4)$$

where

- D_{ij}, is the **j**-th row of $\boldsymbol{D_i}$,

- V_{ij} is the corresponding diagonal element of $\boldsymbol{V_i}$.

This term aims to model the information that may arise from the relationship between cluster size and the distribution of covariates associated with the outcome. Thus, it provides an additional model-based adjustment component in hybrid ICS frameworks where adjustment based solely on cluster size weighting may not be sufficient.

2.2.3. Proposed Doubly Robust Estimating Equation

The cluster-size–weighted doubly robust GEE estimator $\hat{\beta}_{DR}$ is defined as the solution to

$$U_{DR}(\beta) = \sum_{i=1}^{m} w_i(\alpha) D_i V_i^{-1} (Y_i - \mu_i) + \sum_{i=1}^{m} \{1 - w_i(\alpha)\} \sum_{j=1}^{n_i} A_{ij}(\gamma) = 0 \tag{5}$$

The estimating Equation (5) merges:

- Flexible cluster-size weighting, which aims to reduce bias that may arise from the direct relationship between cluster size and results,
- Outcome-model augmentation to correct for covariate imbalance, resulting in a structure that is doubly robust under hybrid ICS.

2.2.4. Double Robustness Under Hybrid ICS

The proposed DR-GEE approach aims to produce consistent estimates of the cluster-balanced marginal estimator under appropriate regularity conditions. In this framework, the performance of the estimator depends on two key components:

$$\hat{\beta}_{DR} \xrightarrow{p} \beta_0 \text{ if either } \begin{cases} \text{(i) the size model embedded in } w_i(\alpha) \text{ is correctly specified,} \\ \textit{or} \\ \text{(ii) the outcome regression model } m_{ij}(X_{ij}; \gamma) \text{ is correct} \end{cases} \tag{6}$$

Under appropriate conditions, this is expected when the cluster size weighting component adequately balances the cluster size–outcome relationship, or when the study's outcome model adequately captures the information derived from covariates.

Therefore, the proposed approach offers a more flexible model-based adjustment framework than a correction based solely on cluster size weighting in hybrid ICS structures. However, this feature should not be interpreted as an absolute guarantee of double robustness, but rather as a DR-type correction structure where the cluster size weighting and the outcome model-based expansion component are used together.

The consistency of the DR-GEE estimator is based on classical GEE assumptions, such as an increasing number of clusters, limited cluster sizes, the correct specification of the outcome model or weighting mechanism, and the variables and outcomes having limited moments. When these conditions are met, the method is expected to produce unbiased and reliable estimates.

2.2.5. Distinction from Existing DR-Type Estimators

The proposed DR-GEE differs from previous work as follows:

1. CWGEE uses non-stabilized $1/n_i$ weights and is not doubly robust [2].
2. WCR is unbiased but computationally intensive and has high variance [5].
3. Huang & Leroux (2011) integrate outcome regression and covariate-based weights, not cluster-size weights; their method does not target direct size-outcome informativeness [6].

In contrast, our proposed method is the first DR-GEE estimator explicitly tailored for hybrid ICS, integrating:

- Flexible cluster-size weighting designed to reduce bias that may arise from the direct relationship between cluster size and results,
- An outcome-model-based expansion component designed to address the informativeness that may arise from imbalances in the distribution of covariates,
- A DR-type model-supported adjustment framework that combines cluster size weighting with outcome-model-based expansion.

2.3. Simulation Design

To evaluate the finite-sample performance of the proposed flexible cluster-size-weighted DR-GEE approach, Monte Carlo simulation experiments were conducted under pure, covariate-dependent and hybrid ICS conditions. This simulation framework is designed to represent complex clustered data structures where larger or smaller clusters may exhibit systematically different outcome distributions, and where the distributions of covariates may also vary depending on cluster size. Such structures may arise in multicenter clinical trials, healthcare research, toxicology experiments, and design-based clustered sampling studies. The simulation design is structured to allow for a comparison of the proposed DR-GEE approach with existing GEE-based methods under different levels of informativeness and average cluster sizes.

2.3.1. Data-Generating Mechanism

For each simulation run, a clustered dataset consisting of m = 200 clusters was generated under hybrid ICS. Let $i = 1, \ldots, m$ index the clusters, and let $j = 1, \ldots, n_i$ index the observations within cluster i. For each cluster, a latent cluster-level random effect was first generated as follows:

$$U_j = N\left(0, 0.5^2\right)$$

Cluster sizes were then generated according to the ICS mechanism. Under hybrid ICS scenarios, cluster size was allowed to depend on the latent cluster-level random effect:

$$M_j \sim Poisson\left(\lambda \exp\left(\gamma U_j\right)\right)$$

where λ controls the baseline mean cluster size. Three levels were considered:

- $\lambda = 3$ (small average cluster size),
- $\lambda = 5$ (moderate average cluster size),
- $\lambda = 8$ (large average cluster size).

Within each cluster, a binary covariate $\boldsymbol{X_{ij}}$ was generated independently of n_i:

$$\boldsymbol{X_{ij}} \sim Bernoulli\left(p_j\right)$$

$$p_j = \frac{\exp\left(\gamma\left(M_j - \lambda\right)\right)}{1 + \exp\left(\gamma\left(M_j - \lambda\right)\right)}$$

This formulation allows the distribution of the outcome-related covariate to vary systematically across clusters of different sizes, thereby representing the covariate-driven component of ICS.

Binary responses were generated from the following logistic model:

$$Y_{ij} \sim Bernoulli\left(\pi_{ij}\right)$$

$$logit\left(\pi_{ij}\right) = \beta_0 + \beta_1 X_{ij} + U_j$$

Here, $\beta_0 = -0.5$ denotes the intercept and $\beta_1 = 0.5$ denotes the true covariate effect. The ICS strength was controlled by γ, and three levels were examined:

- γ = 0.1 (weak ICS),
- γ = 0.5 (moderate ICS),
- γ = 1.0 (strong ICS).

The hybrid ICS scenario combines both mechanisms, with cluster size related to the latent outcome-generating effect and the covariate distribution varying systematically with cluster size. This design enables comparison of the estimators under direct, covariate-driven, and hybrid forms of informative cluster size.

Following the hybrid scenario—which is the primary hypothesis—the production steps for the pure and covariate-dependent ICS scenarios are provided below.

In the pure ICS scenario, the cluster size n_i is determined by U_i, a hidden and cluster-specific random effect, and the data generation process works as follows:

1. Latent Effect: A cluster-level random effect is generated from a normal distribution: $U_i \sim N(0, \sigma^2)$
2. Cluster Size: The size of the i-th cluster, n_i, is simulated using a conditional Poisson distribution: $n_i \sim Poisson(\lambda.\exp(\gamma U_i))$ where λ represents the baseline mean cluster size, and γ governs the strength of the informative cluster size mechanism. To avoid empty clusters, we restrict $n_i \geq 1$.
3. Covariates: The subunit-level predictor $\boldsymbol{X}_{ij}$ is independently drawn from a Bernoulli distribution: $\boldsymbol{X}_{ij} \sim \text{Bernoulli}(0.5)$, ensuring no baseline association between the covariate space and the cluster size $(\boldsymbol{X}_{ij} \perp n_i)$.
4. Outcome: The binary outcome $\boldsymbol{Y}_{ij}$ is generated via the following marginal logistic model:

$$logit\left(P\left(\boldsymbol{Y}_{ij} = 1 \middle| \boldsymbol{X}_{ij}, U_i\right)\right) = \beta_0 + \beta_1 \boldsymbol{X}_{ij} + \alpha U_i$$

where α quantifies the conditional cluster effect on the log-odds of the outcome.

In the Covariate-Dependent ICS scenario, the informativeness of the cluster size does not propagate through an unobserved latent vector, but is instead explicitly driven by the observed covariate space.

The subunit-level covariate prevalence explicitly shifts as a function of the realized cluster size. For a baseline cluster size λ, the probability of a subunit possessing the risk factor $\boldsymbol{X}_{ij}$ is governed by:

$$P(\boldsymbol{X}_{ij} = 1 | n_i) = logit^{-1}(\gamma(n_i - \lambda))$$

The outcome is subsequently simulated without a latent confounder: $logit\left(P\left(\boldsymbol{Y}_{ij} = 1\right)\right) = \beta_0 + \beta_1 \boldsymbol{X}_{ij}$

2.3.2. Competing Estimators

Each simulated dataset was analyzed using five marginal estimators:

1. Standard GEE

   Ignores ICS and serves as a baseline.

2. CWGEE (Cluster-Weighted GEE)

   Applies inverse-cluster-size weights $1/n_i$, following Williamson et al. (2003) [2].

3. WCR (Within-Cluster Resampling)

   Resamples one observation per cluster repeatedly and fits GEE to each pseudo-sample (Hoffman et al. 2001) [5].

4. Huang–Leroux Doubly Weighted GEE (DWGEE)

   A covariate-driven doubly weighted approach designed for ICS induced through covariate–size interactions.

In pure ICS settings, this model is intentionally misspecified, enabling evaluation of its robustness.

5. Proposed DR-GEE

The proposed DR-GEE approach combines an outcome-model-based expansion component with flexible cluster-size weighting. This framework aims both to reduce bias that may arise from the direct cluster-size–outcome relationship and to account for the information derived from the joint distribution of variables in a model-based manner.

The cluster-level weight is defined as a flexible function of cluster size controlled by the α parameter $w_i(\alpha) = n_i^{-\alpha}$. To preserve the overall scale of the estimating equation, this weight has been normalized: $\widetilde{w}_i(\alpha) = n_i^{-\alpha} / E\left(n_i^{-\alpha}\right)$

Here, $\alpha \in [0, 1]$ indicates the degree of cluster size adjustment. When α = 0, no cluster size weighting is applied, which corresponds to the standard GEE. When α = 1, full cluster size weighting is applied, which corresponds to the CWGEE type of full cluster size adjustment. Intermediate values of α = 0.25 and α = 0.50 represent partial cluster size adjustment. In this study, the DR-GEE approach was evaluated under three different levels of cluster size adjustment: α = 0.25, α = 0.50, and α = 1.00.

2.3.3. Simulation Scenarios

The data-generating design was defined by a full factorial combination of two simulation factors:

- ICS strength: $\gamma \in \{0.1, 0.5, 1.0\}$
- Average cluster size: $\lambda \in \{3, 5, 8\}$

This resulted in 9 distinct data-generating scenarios. For each data-generating scenario, the proposed DR-GEE approach was additionally evaluated under three values of the cluster-size adjustment parameter:

- $\alpha \in \{0.25, 0.50, 1.00\}$

The α parameter was not part of the data-generating mechanism; rather, it defined alternative DR-GEE estimator variants corresponding to different degrees of cluster-size adjustment. Thus, the simulation study consisted of 9 data-generating scenarios, with three DR-GEE variants evaluated within each scenario. For each scenario, R = 1000 independent datasets were generated. To reduce Monte Carlo variability in the WCR estimator, B = 100 resamples were used per replication.

2.3.4. Performance Metrics

The following metrics were calculated based on the $\hat{\beta}_1$ estimates obtained for each method:

- $Bias = \mathbb{E}\left[\hat{\beta}_1\right] - \beta_1$
- Root Mean Square Error (RMSE): $\text{RMSE} = \sqrt{\mathbb{E}\left[\left(\hat{\beta}_1 - \beta_1\right)^2\right]}$
- Empirical Standard Error (Emp SE): The standard deviation of the estimated coefficients across simulation runs, capturing the observed variability.
- Model-based Standard Error (Model SE): The average of the robust standard errors reported by each method's estimator, indicating the internal uncertainty quantification.
- Confidence Interval Coverage Rate: The rate at which the true value β_1 falls within the 95% confidence interval
- Confidence Interval Width (CI Width): Difference between lower and upper bound

2.3.5. Implementation

All simulations were implemented in R 4.3 using the geepack, nleqslv, and dplyr packages. The DR-GEE estimating equations were solved using a Broyden-type root-finding algorithm. To ensure reproducibility, each scenario was initialized with a fixed random seed.

3. Results

Table 1 presents the relationships between cluster size and two diagnostic indicators according to γ levels: the correlation between the latent cluster effect and cluster size, $\mathrm{cor}(U, N)$, and the correlation between the mean value of the common variable at the cluster level and cluster size, $\mathrm{cor}(\overline{X}, N)$. The results indicate that as γ increases, the simulation mechanism strengthens in the expected direction. At a low level of informativeness (γ = 0.1), the correlation between cluster size and the latent cluster effect was found to be quite low ($\mathrm{cor}(U, N) = 0.055$). In contrast, a moderate relationship was observed at the γ = 0.5 level ($\mathrm{cor}(U, N) = 0.517$), and this relationship strengthened significantly at the γ = 1.0 level ($\mathrm{cor}(U, N) = 0.740$). A similar trend was observed in the correlation between cluster size and the mean value of the covariate at the cluster level. While the $\mathrm{cor}(\overline{X}, N)$ value was 0.235 at γ = 0.1, it rose to 0.784 at γ = 1.0.

Table 1. Strength of Informative Cluster Size Mechanisms: Correlation Between Cluster Size and Latent Outcome/Covariate Structures Across Varying γ Levels.

Gamma	cor(U,N)	cor($\overline{X}$,N)
Gamma = 0.1	0.055 ± 0.018	0.235 ± 0.166
Gamma = 0.5	0.517 ± 0.096	0.694 ± 0.140
Gamma = 1.0	0.740 ± 0.055	0.784 ± 0.058

Note: $\mathrm{cor}(U, N)$ denotes the correlation between cluster size and latent cluster-level effects, indicating direct outcome-driven informativeness. $\mathrm{cor}(\overline{X}, N)$ represents the correlation between cluster size and cluster-level covariate means, reflecting covariate-driven informativeness. Values are reported as the mean ± standard deviation across simulation replicates.

Table 2 presents the performance metrics obtained from simulations conducted under different levels of informativeness (γ = 0.1, 0.5, 1.0) and average cluster size conditions (λ = 3, 5, 8). In the λ = 3 scenario, the GEE approach exhibits systematic bias as the gamma value increases. When the gamma value is 0.1, the GEE method yields stable results with low bias (Bias = -0.012), low RMSE (0.159), and a very good coverage rate (Coverage = 0.96). However, when γ rises to 0.5, the bias increases to 0.158, creating a significant deviation, and the coverage rate drops to 89%. At γ = 1.0, the bias reaches 0.443, and the coverage rate is only 36%. Particularly in the λ = 8 scenario, the bias is very high at 0.586. The RMSE is 0.600, making it one of the largest errors. The coverage ratio has dropped to 1%, and the CI width is 0.47. In the λ = 3 scenario, under weak ICS conditions (γ = 0.1), the CWGEE approach performed relatively well with low bias (-0.020), acceptable RMSE (0.181), and high coverage (0.99). At γ = 0.5, the bias increased to 0.155, and the coverage rate dropped to 0.93. Under strong ICS conditions (γ = 1.0), this deterioration became more pronounced; the bias was calculated as 0.380, the RMSE as 0.439, and the coverage rate dropped to 0.57. Specifically, at γ = 1.0 and λ = 8, the bias increased to 0.562, while the coverage rate dropped to 1%. In the WCR approach under the λ = 3 scenario, under weak ICS conditions (γ = 0.1), the WCR bias was found to be close to zero (0.000), the RMSE value was 0.169, and the coverage rate was calculated as 1.00. However, this high coverage rate was accompanied by a wide confidence interval of 0.97. As the level of information increased, an increase in WCR bias and RMSE valueswas observed. At γ = 0.5, although the coverage rate remained at 0.99, the bias increased to

0.115. Under strong ICS conditions (γ = 1.0), the bias was found to be 0.347, the RMSE 0.403, and the confidence interval width 1.02. Similarly, in the λ = 5 and λ = 8 scenarios, the coverage rates at γ = 1.0 decreased to 0.56 and 0.44, respectively. The DWGEE approach exhibits acceptable performance, particularly at low gamma levels (γ = 0.1). For example, in the λ = 3 scenario, a γ = 0.1 value resulted in a bias of -0.043, an RMSE of 0.240, and a coverage rate of 97%. However, as the γ value increases, the method weakens in terms of statistical accuracy. For γ = 0.5, the bias increased to 0.089, while the confidence interval width reached 1.00; at γ = 1.0, the bias was calculated as 0.221 and the confidence interval width as 1.05. A similar situation applies to the λ = 5 and λ = 8 scenarios. Simulation findings demonstrate that the proposed DR-GEE (α = 1.0) approach exhibits the most stable and reliable performance under hybrid ICS scenarios. Particularly at high informativeness levels such as γ = 0.5 and γ = 1.0, the DR-GEE (α = 1.0) method stands out with low bias and low RMSE values; it also supports its statistical reliability by maintaining coverage rates close to 95%. In contrast, flexible variants such as α = 0.25 and α = 0.50 provided variance control by reducing the weighting degree, but failed to adequately correct the effect of informative clusters, leading to bias and coverage issues. This clearly highlights the need for strong weight correction under hybrid ICS and the advantage of the α = 1 level.

As shown in Figures 1 and 2, a comprehensive comparison of the performance of GEE-based estimators under hybrid ICS conditions is presented by summarizing their bias and RMSE trajectories across increasing levels of informativeness (γ = 0.1, 0.5, 1.0) and cluster sizes (λ = 3, 5, 8). Specifically, as γ increases, these methods fail to maintain statistical reliability, and the bias and RMSE values increase rapidly. In contrast, the proposed DR-GEE method with flexible cluster size weights (particularly with α = 1.00) demonstrated a more favorable bias and RMSE performance, maintaining low bias and consistent RMSE values across all settings.

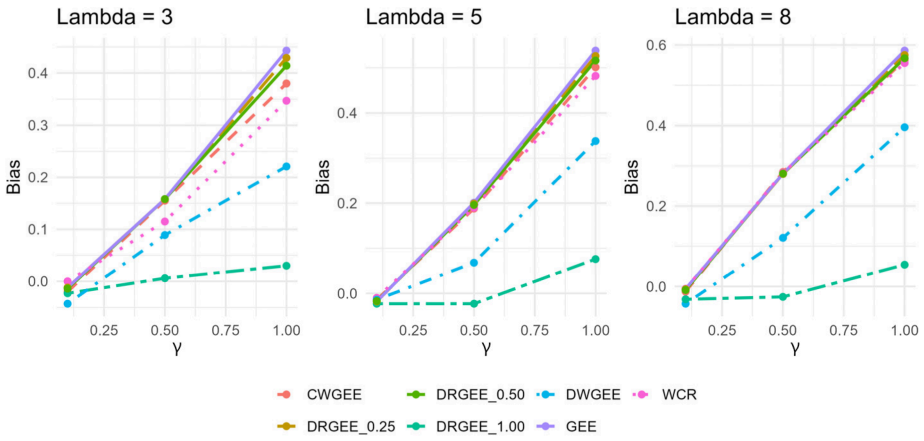


Figure 1. Comparison of Bias Across Gamma Levels for GEE-Based Estimators Under Varying Average Cluster Sizes.

Under pure ICS conditions (Supplementary Table S1), it was found that the GEE approach suffers from a loss of performance at high γ values (e.g., 92% coverage for γ = 1.0 and λ = 3). CWGEE exhibits instability, with wide confidence intervals (CI Width $\approx$ 0.76–0.84) and a relatively higher RMSE. WCR is a highly inefficient method, characterized by very wide confidence intervals (CI Width > 1.0) and a high model standard error (Model SE > 0.25). DWGEE is the method that produces the widest CI width (ranging from 0.98 to 1.03) and a high standard error (Model SE $\approx$ 0.263 for λ = 8 and γ = 1.0). Among all the methods evaluated, DR-GEE with full cluster-size weighting (α = 1.00) produced relatively narrow confidence intervals while maintaining coverage probabilities close to the nominal 95% level, particularly with strong informative cluster-size adjustments.

Table 2. Simulation-Based Performance Metrics for Marginal Estimators Under Hybrid Informative Cluster Size Across Varying γ and λ Conditions.

	λ	Gamma = 0.1						Gamma = 0.5						Gamma = 1.0					
		Bias	**RMSE**	**Emp SE**	**Model SE**	**Coverage**	**CI Width**	**Bias**	**RMSE**	**Emp SE**	**Model SE**	**Coverage**	**CI Width**	**Bias**	**RMSE**	**Emp SE**	**Model SE**	**Coverage**	**CI Width**
GEE	3	−0.012	0.159	0.160	**0.164**	**0.96**	0.64	0.158	0.216	**0.149**	**0.169**	0.89	0.66	0.443	0.480	**0.188**	**0.183**	0.36	0.72
CWGEE		−0.020	0.181	0.181	0.194	0.99	0.76	0.155	0.242	0.187	0.198	0.93	0.78	0.380	0.439	0.221	0.206	0.57	0.81
WCR		**0.000**	0.169	0.170	0.249	1.00	0.97	0.115	0.207	0.173	0.256	0.99	1.00	0.347	0.403	0.206	0.261	0.83	1.02
DWGEE		−0.043	0.240	0.237	0.249	0.97	0.97	0.089	0.259	0.245	0.255	0.94	1.00	0.221	0.371	0.299	0.268	0.86	1.05
DR-GEE (α = 1.00)		−0.023	0.180	0.179	0.195	0.99	0.76	**0.006**	**0.200**	0.201	0.212	**0.95**	0.83	**0.030**	**0.256**	0.256	0.240	**0.94**	0.94
DR-GEE (α = 0.25)		−0.012	**0.158**	**0.158**	0.165	**0.96**	0.65	0.158	0.218	0.151	0.170	0.89	0.67	0.429	0.469	0.190	**0.183**	0.38	0.72
DR-GEE (α = 0.50)		−0.014	0.160	0.160	0.170	0.97	0.67	0.158	0.223	0.159	0.175	0.90	0.69	0.414	0.458	0.196	0.187	0.43	0.73
GEE	5	−0.018	**0.120**	**0.119**	**0.129**	**0.95**	0.50	0.201	0.250	**0.149**	**0.136**	0.72	0.53	0.538	0.566	**0.178**	**0.148**	0.10	0.58
CWGEE		−0.014	0.143	0.143	0.147	0.96	0.58	0.188	0.242	0.153	0.154	0.77	0.60	0.501	0.534	0.188	0.166	0.20	0.65
WCR		**−0.010**	0.140	0.140	0.258	1.00	1.01	0.190	0.244	0.153	0.254	0.97	1.00	0.482	0.519	0.193	0.258	0.56	1.01
DWGEE		−0.014	0.215	0.216	0.209	0.93	0.82	0.068	0.233	0.224	0.221	**0.93**	0.86	0.338	0.425	0.259	0.247	0.69	0.97
DR-GEE (α = 1.00)		−0.023	0.148	0.147	0.147	0.96	0.58	**−0.023**	**0.186**	0.185	0.164	**0.93**	0.64	**0.076**	**0.224**	0.212	0.188	**0.93**	0.74
DR-GEE (α = 0.25)		−0.018	0.122	0.121	**0.129**	0.97	0.51	0.199	0.246	0.146	0.136	0.74	0.53	0.526	0.555	0.177	**0.148**	0.14	0.58
DR-GEE (α = 0.50)		−0.017	0.126	0.125	0.132	0.97	0.52	0.196	0.243	0.145	0.138	0.75	0.54	0.516	0.546	0.178	0.150	0.16	0.59
GEE	8	**−0.006**	**0.098**	**0.099**	**0.102**	0.96	0.40	0.281	0.305	**0.119**	**0.113**	0.33	0.44	0.586	0.600	**0.130**	0.121	0.01	0.47
CWGEE		−0.013	0.114	0.114	0.111	0.94	0.43	0.285	0.312	0.128	0.120	0.29	0.47	0.562	0.580	0.142	0.131	0.01	0.51
WCR		**−0.006**	0.116	0.117	0.268	1.00	1.05	0.283	0.311	0.130	0.268	**0.97**	1.05	0.555	0.573	0.142	0.268	0.44	1.05
DWGEE		−0.043	0.184	0.180	0.156	0.91	0.61	0.121	0.226	0.192	0.174	0.86	0.68	0.396	0.450	0.215	0.218	0.55	0.85
DR-GEE (α = 1.00)		−0.032	0.118	0.114	0.111	**0.95**	0.44	**−0.026**	**0.151**	0.150	0.132	**0.93**	0.52	**0.054**	**0.171**	0.163	0.149	**0.93**	0.59
DR-GEE (α = 0.25)		−0.007	0.099	**0.099**	**0.102**	0.96	0.40	0.280	0.304	**0.119**	**0.113**	0.31	0.44	0.575	0.589	0.130	**0.120**	0.01	0.47
DR-GEE (α = 0.50)		−0.009	0.101	0.102	0.104	0.96	0.41	0.281	0.305	0.120	0.114	0.31	0.45	0.567	0.582	0.132	0.121	0.01	0.47

Note: Performance metrics include Bias, Root Mean Square Error (RMSE), empirical and model-based standard errors (Emp SE, Model SE), coverage probability of 95% confidence intervals, and mean CI width. γ denotes the strength of ICS, and λ represents the mean cluster size. Results are presented for each method under increasing levels of ICS informativeness (γ = 0.1, 0.5, 1.0) and average cluster size (λ = 3, 5, 8). Bold values indicate the most favorable result for each performance metric within the corresponding simulation scenario. For Bias, RMSE, Emp SE, and Model SE, lower values were considered preferable, whereas for Coverage, values closest to the nominal level of 0.95 were considered preferable.

https://doi.org/10.3390/stats9040074

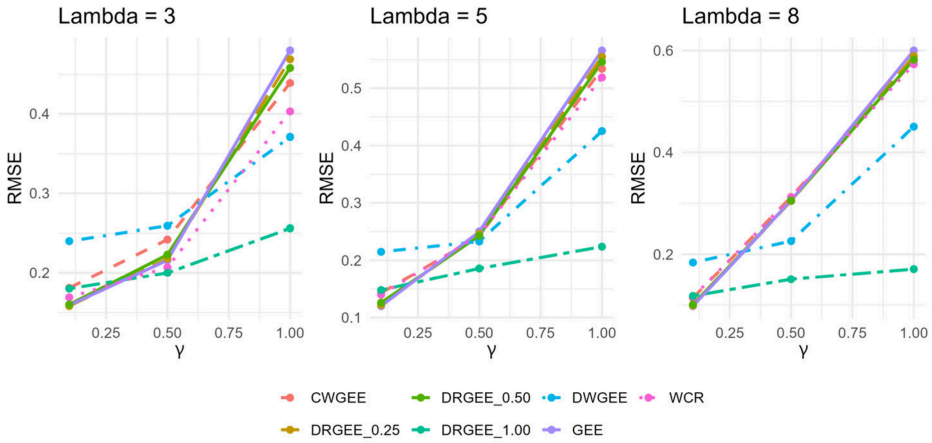


Figure 2. Comparison of RMSE Across Gamma Levels for GEE-Based Estimators Under Varying Average Cluster Sizes.

Under covariate-dependent ICS conditions (Supplementary Table S2), DR-GEE methods—particularly the $\alpha = 0.25$ and $\alpha = 0.50$ variants—achieved the most favorable trade-off between bias and variance, offering high coverage with relatively narrow confidence intervals. While DR-GEE ($\alpha = 1.00$) showed slightly higher RMSE in small cluster sizes, it remained competitive across most settings. In contrast, WCR offered inflated coverage with overly conservative intervals, and DWGEE suffered from high variance.

4. Discussion

Due to the relationship between the response variable and cluster size, the inadequacy of the classical GEE method in ICS situations has led to the development of two alternative approaches in the literature [15]. The first of these, the WCR approach, aims to eliminate the direct cluster effect on the response by selecting a single observation from each cluster; however, this method is computationally intensive. Therefore, its use has been criticized in the literature [16]. However, it is argued that the time required for resampling will decrease in parallel with the advancement of technology [5]. Nevertheless, Williamson et al. [2] proposed the CWGEE approach due to the criticisms of the WCR approach. In the CWGEE approach, $1/n_i$ is used instead of the inverse probability weight and thus CWGEE is expressed as a different case of WGEE [9,15]. Building on this literature, Huang and Leroux (2011) introduced a doubly weighted GEE method that integrates both outcome regression and covariate-based inverse probability weighting [6]. While effective for covariate-driven ICS, their approach does not target the direct outcome–cluster size association and is therefore inadequate under pure or hybrid ICS mechanisms. The proposed DR-GEE approach addresses these limitations by combining a flexible cluster size weighting component with an augmentation term based on the resulting model. In this approach, cluster size adjustment is treated not merely as a variance correction, but as a weighting strategy associated with the targeted cluster-balanced marginal estimator. This aligns with robust estimation methods such as those developed by Zhao et al. (2025), which focus on making reliable inferences in challenging data structures like heavy-tailed distributions [17]. This connection supports DR-GEE's potential to generate robust inferences. While the cluster size weighting component aims to reduce bias that may arise directly from cluster size–outcome informativeness, the expansion component provides an additional model-based adjustment for covariate-induced informativeness. Extensive simulations under varying gamma levels and cluster size conditions show that DR-GEE consistently achieves lower bias and RMSE than GEE and CWGEE, particularly when the correlation

between cluster size and both the outcome and covariate is strong (as is the case when $\gamma \geq 0.5$).

In conclusion, the DR-GEE framework offers a flexible and methodologically sound alternative for modeling marginal effects under different mechanisms of ICS. Our extended simulation findings show that in pure ICS scenarios—where cluster size is related only to the outcome variable—the proposed method can limit the loss of statistical efficiency (SE) compared to standard weighted methods; in covariate-dependent ICS scenarios, where cluster size affects the distribution of covariates, the proposed method can reduce the bias observed in standard unweighted GEE models. The DR-GEE framework, especially with full weighting ($\alpha = 1$), offers a promising and methodologically sound approach for modeling marginal effects in settings where both the covariate and outcome are correlated with cluster size. This is common in educational data (e.g., class size and achievement), multicenter clinical trials (e.g., center size and treatment response), or survey studies with design-based clustering.

Limitations

Although the DR-GEE method proposed in this study is evaluated under a comprehensive simulation framework, some limitations should be considered. First, only artificially generated simulation data are used in this study. Although real-world data applications are not within the scope of this study, the additional simulation scenarios (Supplementary Tables S1 and S2) are intended to represent the complex ICS structures that may be encountered in reproductive toxicology or dental health research. These scenarios aim to demonstrate the robustness of the DR-GEE approach under different ICS mechanisms. Another limitation is that the simulation study was conducted only on binary outcomes. The performance of DR-GEE for continuous or count-type (Poisson) data requires a separate evaluation.

5. Conclusions

In this study, the proposed flexible cluster-size-weighted DR-GEE approach was evaluated via simulation under pure, covariate-dependent, and hybrid ICS conditions. The simulation results revealed that, particularly in scenarios where the level of informativeness increased, the DR-GEE approach exhibited better performance in terms of bias and RMSE compared to the standard GEE and CWGEE. As the γ parameter, representing ICS strength, increased, the relationship between cluster size and the outcome-generation mechanism strengthened, leading to a significant loss of performance in classical GEE-based estimators. In particular, the DR-GEE variant incorporating full cluster size adjustment ($\alpha = 1.00$) in the hybrid ICS scenario provided lower bias, lower RMSE, and confidence interval coverage rates close to the nominal level, especially under conditions of strong informativeness. Overall, the proposed DR-GEE framework offers an interpretable and computationally feasible approach to reduce cluster-size-dependent bias and enable cluster-balanced marginal inferences in hybrid ICS settings where the informative cluster size is related to both the outcome variable and covariates.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/stats9040074/s1, Table S1: Simulation-Based Performance Metrics for Marginal Estimators Under Pure Informative Cluster Size Across Varying γ and λ Conditions; Table S2: Simulation-Based Performance Metrics for Marginal Estimators Under Covariate-Dependent Informative Cluster Size Across Varying γ and λ Conditions.

Author Contributions: Conceptualization: B.D.H. and Z.N.A. Data curation: B.D.H. and Z.N.A. Formal analysis: B.D.H. Funding acquisition: None. Methodology: B.D.H. and Z.N.A. Visualization:

B.D.H. and Z.N.A. Writing—original draft: B.D.H. and Z.N.A. Writing—review & editing: B.D.H. and Z.N.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. This study was based exclusively on simulated data and did not involve human participants, animals, or identifiable personal data.

Informed Consent Statement: Not applicable.

Data Availability Statement: No empirical data were collected or analyzed in this study. The simulation design, data-generating mechanisms, parameter settings, and estimation procedures are described in the manuscript. The R code used to generate the simulated datasets and reproduce the analyses is available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Senevirathne, H.K.W.; Dutta, S. Testing Informativeness of Covariate-Induced Group Sizes in Clustered Data. *Mathematics* **2024**, *12*, 1623. [CrossRef]
2. Williamson, J.M.; Datta, S.; Satten, G.A. Marginal analyses of clustered data when cluster size is informative. *Biometrics* **2003**, *59*, 36–42. [CrossRef] [PubMed]
3. Benhin, E.; Rao, J.N.K.; Scott, A.J. Mean estimating equation approach to analyzing cluster-correlated data with nonignorable cluster sizes. *Biometrika* **2005**, *92*, 435–450. [CrossRef]
4. Seaman, S.; Pavlou, M.; Copas, A. Review of methods for handling confounding by cluster and informative cluster size in clustered data. *Stat. Med.* **2014**, *33*, 5371–5387. [CrossRef] [PubMed]
5. Hoffman, E.B.; Sen, P.K.; Weinberg, C.R. Within-cluster resampling. *Biometrika* **2001**, *88*, 1121–1134. [CrossRef]
6. Huang, Y.; Leroux, B. Informative cluster sizes for subcluster-level covariates and weighted generalized estimating equations. *Biometrics* **2011**, *67*, 843–851. [CrossRef] [PubMed]
7. Scharfstein, D.O.; Rotnitzky, A.; Robins, J.M. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Am. Stat. Assoc.* **1999**, *94*, 1096–1120. [CrossRef]
8. Laan, M.J.; Robins, J.M. *Unified Methods for Censored Longitudinal Data and Causality*; Springer: New York, NY, USA, 2003.
9. Bang, H.; Robins, J.M. Doubly robust estimation in missing data and causal inference models. *Biometrics* **2005**, *61*, 962–973. [CrossRef] [PubMed]
10. Seaman, S.; Copas, A. Doubly robust generalized estimating equations for longitudinal data. *Stat. Med.* **2009**, *28*, 937–955. [CrossRef] [PubMed]
11. Prague, M.; Wang, R.; De Gruttola, V. CRTgeeDR: An R package for doubly robust generalized estimating equations estimations in cluster randomized trials with missing data. *R J.* **2017**, *9*, 105–115. [CrossRef]
12. Rabideau, D.J.; Li, F.; Wang, R. Multiply robust generalized estimating equations for cluster randomized trials with missing outcomes. *Stat. Med.* **2024**, *43*, 1458–1474. [CrossRef] [PubMed]
13. Pavlou, M.; Seaman, S.R.; Copas, A.J. An examination of a method for marginal inference when the cluster size is informative. *Stat. Sin.* **2013**, *23*, 791–808. [CrossRef]
14. Seaman, S.R.; Pavlou, M.; Copas, A.J. Methods for observed-cluster inference when cluster size is informative: A review and clarifications. *Biometrics* **2014**, *70*, 449–456. [CrossRef] [PubMed]
15. Kahan, B.C.; Li, F.; Blette, B.; Jairath, V.; Copas, A.; Harhay, M. Informative cluster size in cluster-randomised trials: A case study from the TRIGGER trial. *Clin. Trials* **2023**, *20*, 661–669. [CrossRef] [PubMed]
16. Wang, M.; Kong, M.; Datta, S. Inference for marginal linear models for clustered longitudinal data with potentially informative cluster sizes. *Stat. Methods Med. Res.* **2011**, *20*, 347–367. [CrossRef] [PubMed]
17. Zhao, P.; Wu, J.; Liu, Z.; Wang, C.; Fan, R.; Li, Q. Differential Private Stochastic Optimization with Heavy-tailed Data: Towards Optimal Rates. *Proc. AAAI Conf. Artif. Intell.* **2025**, *39*, 22795–22803. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.