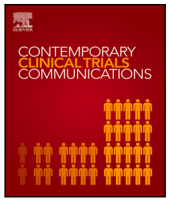
Contemporary Clinical Trials Communications 35 (2023) 101194

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

journal homepage: www.elsevier.com/locate/conctc



Research paper

Bias-corrected and doubly robust inference for the three-level longitudinal cluster-randomized trials with missing continuous outcomes and small number of clusters: Simulation study and application to a study for adults with serious mental illnesses

Chaeryon Kang [a,*], Di Zhang [a], James Schuster [b], Jane Kogan [c], Cara Nikolajski [c], Charles F. Reynolds [d]

[a] Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261, USA
[b] UPMC Health Plan, Pittsburgh, PA 15219, USA
[c] UPMC Center for High-Value Health Care, Pittsburgh, PA 15219, USA
[d] Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, PA 15213, USA

ARTICLE INFO

Keywords:
Augmented inverse probability weighted estimator
Bias-corrected variance
Longitudinal cluster-randomization
Missing at random
Small number of clusters

ABSTRACT

Longitudinal cluster-randomized designs have been popular tools for comparative effective research in clinical trials. The methodologies for the three-level hierarchical design with longitudinal outcomes need to be better understood under more pragmatic settings; that is, with a small number of clusters, heterogeneous cluster sizes, and missing outcomes. Generalized estimating equations (GEEs) have been frequently used when the distribution of data and the correlation model are unknown. Standard GEEs lead to bias and an inflated type I error rate due to the small number of available clinics and non-completely random missing data in longitudinal outcomes. We evaluate the performance of inverse probability weighted (IPW) estimating equations, with and without augmentation, for two types of missing data in continuous outcomes and individual-level treatment allocation mechanisms combined with two bias-corrected variance estimators. Our intensive simulation results suggest that the proposed augmented IPW method with bias-corrected variance estimation successfully prevents the inflation of false positive findings and improves efficiency when the number of clinics is small, with moderate to severe missing outcomes. Our findings are expected to aid researchers in choosing appropriate analysis methods for three-level longitudinal cluster-randomized designs. The proposed approaches were applied to analyze data from a longitudinal cluster-randomized clinical trial involving adults with serious mental illnesses.

1. Introduction

In 2016, an estimated 10.4 million adults in the United States (4.2% of the population) experienced serious mental illnesses (SMIs, [1]), and 6.7 million of these individuals (64.8%) received mental health treatment [2]. The lack of high-quality medical care can increase the risk of early chronic medical conditions, hospitalization, and premature deaths in adults with SMI [3]. A comparative effectiveness study ("Optimal Health (OH)") was designed to evaluate the effects of two active treatment strategies on longitudinal changes in several patient care outcomes over two years in adults with SMI [3,4]. A cluster-randomization design was used to prevent contamination between treatment groups within the same clinic and evaluate implementation in more pragmatic settings. Eleven community mental health providers (CMHs, clinics) were randomly assigned to one of the two treatment groups. Our study is motivated by the OH study.

Longitudinal cluster-randomized controlled trials (CRTs) are frequently used because they have several advantages over individually randomized controlled trials in comparative effectiveness studies [5]. However, additional challenges inherent in the study design require careful analysis. For example, a small number of clusters can lead to inflated type I error rates [6–8], and missing data in longitudinal outcomes can lead to biased estimates or loss of efficiency, if overlooked. Two common approaches for CRT analysis include the generalized linear mixed model (GLMM) [9] and generalized estimating equation (GEE) [10]. Because of the small number of clusters, Gatsonis

* Corresponding author.
E-mail address: crkang@pitt.edu (C. Kang).

https://doi.org/10.1016/j.conctc.2023.101194
Received 6 September 2022; Received in revised form 21 April 2023; Accepted 19 July 2023
Available online 28 July 2023

and Morton [11] suggested that the assumption of normality regarding random effects in GLMM might be impractical (p.143). Preisser et al. [12] suggested that population-averaged modeling involving estimating equations was more appropriate for analyzing treatment effects in cluster-randomized designs where treatment is a cluster-specific covariate. Accordingly, we focused on the GEE approach for longitudinal CRTs.

Recent statistical advances in CRTs have been observed in both classical [13–16] and Bayesian approaches [17,18]; however, the majority of statistical methods developed for CRTs have been limited to a two-level (for example, participants are nested within clinics) or three-level (participants are nested within doctors and doctors are further nested within clinics) hierarchical design without longitudinal outcomes. Several studies have focused on using advanced GEE to correct the bias due to a small number of clusters [6,12,19–23] or missing outcomes [16,24]. However, in these studies, several methods were evaluated without addressing various problematic features that occur simultaneously in longitudinal CRTs; for example, small number of clusters, heterogeneous cluster sizes, three-level hierarchical design, missing longitudinal outcomes, and change in treatment effects over time. Specifically, Lu et al. [19] compared two bias-corrected variance estimators using GEE for marginal mean models including treatment-by-time interactions, assuming equal cluster sizes and no missing outcomes. Ford and Westgate [23] provided an intensive comparison of various bias-correction methods with GEE for two-level longitudinal continuous, binary, or Poisson distributed outcomes assuming no missing data. Seaman and Copas [24] evaluated several methods, including an augmented inverse probability weighted (AIPW) estimating equation, for handling monotone missing data in two-level longitudinal studies. Prague et al. [25] further investigated the performance of various weighted GEEs in simpler design settings and found that the AIPW estimating equation, in combination with bias-corrected variance estimators, showed superior performance while inferring the main treatment effects. To the best of our knowledge, however, GEE methods for three-level longitudinal CRTs with missing outcomes and a small number of clusters and which allow for unequal cluster sizes, have not been well studied. Therefore, we investigated whether and how such estimating equation approaches using inverse probability weighted estimators could infer treatment effects in three-level longitudinal CRTs, under more pragmatic constraints with a small number of clusters, heterogeneous cluster sizes, and missing outcomes.

Two commonly assumed states of missing data ("missingness") are (1) missing completely at random (MCAR) (for example, missing data in longitudinal outcomes occur completely by chance) (2) missing at random (MAR) (for example, missing data in longitudinal outcomes depend only on observed data, such as fully observed covariates and past observed outcomes) [26]. We considered two MAR mechanisms. First, we considered the covariate-dependent MAR (hereafter "CD-MAR") mechanism, in which missingness depends on baseline covariates only. Our methods under CD-MAR allow intermittent missingness in outcomes as well as missingness due to dropout. Second, we considered the outcome and covariate dependent MAR, in which missingness depends on previously observed outcomes as well as baseline covariates ("OCD-MAR"). It is challenging to model nonmonotone missing patterns in the OCD-MAR mechanism, and it becomes even more difficult when there are many missing patterns with small subgroup sizes for each missing pattern, as observed in the OH study. Therefore, our study for the OCD-MAR was limited to monotone missingness. By "monotone missing mechanism" we mean that missingness at a particular time point implies missingness in all subsequent time points.

The rest of the paper is organized as follows: In Section 2, we describe our motivation study. In Section 3, we introduce two bias-corrected variance estimators and three IPW estimating equations in three-level longitudinal CRTs. Results of simulation studies evaluating the finite sample performance of these methods are presented in Section 4. In Section 5, we apply the proposed methods to the Optimal Health study. We conclude with a discussion of our findings and suggest future research topics in Section 6.

2. Motivating study: Optimal health for adults with SMI

The OH study is a multicenter longitudinal CRT in which the effects of two evidence-based interventions for adult SMI, Provider-Supported Care (PS) and Patient Self-Directed Care (SD), were compared in a patient population in Pennsylvania (USA) in 2013-2016 [3,4]. In the PS intervention, a full-time registered nurse provided consultation to wellness coaches as well as wellness support and education to individual participants. In the SD intervention, the participants took a more active role in managing their health by using self-management toolkits and content tailored to their needs, through a web portal. Eleven CMHs (clinics) were randomized to one of the two interventions: Five CMHs (713 participants) were allocated to the PS group and six CMHs (516 participants) to the SD group. The coefficient of variation (cv) of clinic size was 0.71 (cv = standard deviation/mean =79.24/111.73). Individual participant data were gathered every six months, over a period of two years following enrollment; thus, there were five timepoints of outcome measurements: a baseline and four follow-up measurements at months 6, 12, 18, and 24. Therefore, the data have a three-level hierarchical structure, where visits are nested within participants that are nested within CMHs; CMHs are further nested within interventions by cluster-randomization of the CMHs. Although there is general agreement that the health of adults with SMI can be effectively managed through either PS and SD cares, there was no studies in which longitudinal changes are compared between the two models of care. The researchers in the OH study primarily aimed to investigate if and how the longitudinal changes in patient care outcomes differ between the PS and SD groups.

In the present study, we focused on three specific outcomes: the patient activation measure (PAM), quality of life enjoyment and satisfaction questionnaire (Q-LES-Q-18: Quality of life, QSF), and patient assessment of chronic illness care (PACIC) score. PAM is a 22-item measure that assesses the confidence and ability of patients to manage their health. The QSF score captures life satisfaction over the past week (%). PACIC measures patient satisfaction with care. The overall missing rates are 43.8% for PAM (PS: 39.7%/ SD: 49.4%), 40.1% for QSF (PS:35.2%/ SD: 46.9%), and 32.7% for PACIC (PS: 28.4%/ SD:38.7%) scores. More information on the missing data of the OH study is given in Tables 10–14 in Supplementary Materials.

3. Statistical methods

3.1. GEE with robust variance estimator

Let $\{Y_{ijk}, R_{ijk}, T_{ijk}, A_i, Z_{ijk}^T\}$ denote the data from the kth visit ($k = 1, \dots, r_{ij}$, $r_{ij} \leq K$) for the jth participant ($j = 1, \dots, m_i$) in the ith clinic ($i = 1, \dots, n$), where Y_{ijk} is a continuous longitudinal outcome, R_{ijk} indicates observation of Y_{ijk} (1 for observation; 0 for missing), T_{ijk} indicates time of visit, A_i indicates the treatment assigned to the clinic i, and Z_{ijk} is a p-dimensional vector of baseline covariates. Although our method can be extended for more general settings, we describe our method for binary treatment, with $A = 1$ or 0, and continuous outcome settings in this article.

We aim to infer the regression coefficient of treatment effects in the marginal mean of the outcomes, $\mu = (\mu_1, \dots, \mu_n)^T$, within the GEE framework. Given longitudinally measured continuous outcomes, the mean model has been expressed using a generalized linear regression model as follows:

$$\mu(X_{ijk}; \boldsymbol{\beta}) = X_{ijk}\boldsymbol{\beta} = \beta_0 + \beta_1 A_i + \beta_2 T_{ijk} + \beta_3 A_i \times T_{ijk}, \quad (1)$$

where $\boldsymbol{\beta} \in \mathbf{R}^p$, $X_{ijk} = \{1, A_i, T_{ijk}, A_i \times T_{ijk}\}$. We first test for the treatment-by-time interaction effect. In the absence of the interaction effect ($\beta_3 = 0$), we test for the main effect of treatment with $X_{ijk} = \{1, A_i, T_{ijk}\}$. The standard GEE [10] solves $\sum_{i=1}^n U_i(\beta) = \sum_{i=1}^n D_i^T V_i^{-1}\{Y_i - \mu_i(\beta)\} = 0$, where $\mu_i(\beta) = \mu(X_i; \beta)$, $D_i = \{\partial\mu_i(\beta)/\partial\beta\}$, and V_i is

the variance–covariance matrix of outcomes in the ith clinic. Here Y_i and X_i are $\tilde{m}_i = \sum_{j=1}^{m_i} r_{ij}$-dimensional vector and $\tilde{m}_i \times p$ dimensional matrix for the ith clinic, respectively. Under some regularity conditions and MCAR, the GEE estimator achieves the consistency of β when the number of clinics is sufficiently large, usually greater than 50, without the joint distribution of the outcomes being specified as long as the mean model $\mu(X_{ijk};\beta)$ is correct [10].

We consider two working correlation structures. First, we assume an exchangeable correlation structure for within-clinic and within-participant correlations. We estimate the variance–covariance matrix, V_i, by extending the variance component estimation method proposed by Kloke et al. [27] from the two-level into the three-level hierarchical structure. The variance component estimation method requires some assumptions, such as continuous random errors with zero mean, finite second moments, and an exchangeable covariance structure. However, the approach is nonparametric, and the normality assumption for the joint distribution of correlated outcomes is not needed. We describe the proposed estimation procedure for the variance–covariance matrix in Section 1.1 of Supplementary Materials. Second, we consider an exchangeable correlation among participants within the same clinic and autoregressive of order 1 (AR(1)) correlation among within-participant visits. Details regarding the AR(1) structure are provided in Section 1.2 of Supplementary Materials.

The robust sandwich variance estimator has the following form: $V^R = \Omega^{-1}\left(\sum_{i=1}^n D_i^T V_i^{-1}\epsilon_i\epsilon_i^T V_i^{-1} D_i\right)\Omega^{-1}$, where $\epsilon_i = Y_i - \mu_i(\beta)$ and $\Omega = \sum_{i=1}^n D_i^T V_i^{-1} D_i$. For a single parameter of interest, the Wald t-test rejects the null hypothesis, $H_0 : \beta_q = 0$, if $|t_R| = |\hat{\beta}_q / \sqrt{\hat{V}^R(\hat{\beta}_q)}| > t_{1-\alpha/2}(\mathrm{df})$, where $\hat{V}^R(\hat{\beta}_q)$ is the qth diagonal element of $\hat{V}^R$ corresponding to β_q and df is a degree of freedom. To test the vector of parameters $H_0 : L\beta = 0$ using a matrix of linear contrast L with a rank ℓ, F-tests with ℓ degrees of freedom can be used, where $F = (L\hat{\beta})^T\{L\hat{V}^R L^T\}^{-1} L\hat{\beta}/\ell$.

3.2. Bias-corrected sandwich variance estimators

The robust variance estimator tends to underestimate the true covariance matrix that can lead to inflated type I error rates and undercoverage of confidence intervals due to its large variation when n is small [6,28]. In practice, it can be difficult to recruit a sufficient number of clinics to conduct a longitudinal CRT (only 11 clinics were included in the OH study). We use two bias-corrected sandwich variance estimators to reduce the risk of inflated type I errors due to the small number of clinics: KC bias-corrected [6,19,29] and MD bias-corrected estimators [6,19,30].

$$V^{KC} = \Omega^{-1}\left\{\sum_{i=1}^n D_i^T V_i^{-1}(I_i - H_i)^{-1/2}\epsilon_i\epsilon_i^T(I_i - H_i)^{-1/2}V_i^{-1}D_i\right\}\Omega^{-1},$$
$$V^{MD} = \Omega^{-1}\left\{\sum_{i=1}^n D_i^T V_i^{-1}(I_i - H_i)^{-1}\epsilon_i\epsilon_i^T(I_i - H_i)^{-1}V_i^{-1}D_i\right\}\Omega^{-1}, \quad (2)$$

where $H_i = D_i\Omega^{-1}D_i^T V_i^{-1}$ and I_i is the identity matrix for the ith clinic.

3.3. The inverse probability weighted estimating equation

In longitudinal studies, the MCAR assumption is frequently violated, and the standard GEE might fail to produce consistent results [10]. Here, we first assume longitudinally collected data follow the CD-MAR mechanism, and we apply the inverse probability weighting approach [31,32]. We then present the modified IPW and AIPW estimating equations for the monotone OCD-MAR mechanism.

Assuming the CD-MAR mechanism, Prague et al. [25] proposed the following IPW estimating equation and robust variance estimator:

$$\sum_{i=1}^n U_{i,IPW}(\beta) = \sum_{i=1}^n D_i^T V_i^{-1} W_i(\xi)\{Y_i - \mu_i(\beta)\} = 0, \quad (3)$$

$$V^R_{IPW} = \Omega_1^{-1}\left\{\sum_{i=1}^n D_i^T V_i^{-1} W_i(\xi)\epsilon_i\epsilon_i^T W_i(\xi)V_i^{-1}D_i\right\}\Omega_1^{-1}, \quad (4)$$

where the weight matrix $W_i(\xi) = \mathrm{diag}\left\{R_{ijk}/\pi_{ijk}(X_{ijk}, Z_{ijk};\xi)\right\}$, the missingness model $\pi_{ijk} = \mathrm{P}(R_{ijk} = 1|X_{ijk}, Z_{ijk};\xi)$, and $\Omega_1 = \sum_{i=1}^n D_i^T V_i^{-1} W_i(\xi) D_i$. Assuming the CD-MAR mechanism, standard logistic regression models have been commonly used to predict π, and the inverse of π is used to weigh the observed outcomes. To improve the unstable features of IPW estimators, we used a stabilized weight approach [33] and defined the weight as $W_i(\xi) = \mathrm{diag}\left\{\mathrm{P}(R_{ijk} = 1)R_{ijk}/\pi_{ijk}(X_{ijk}, Z_{ijk};\xi)\right\}$. Under a correctly specified missingness model and MAR, the IPW estimator is consistent and normally distributed. We then modified (4) for two bias-correction methods:

$$V^{KC}_{IPW} = \Omega_1^{-1}\left\{\sum_{i=1}^n D_i^T V_i^{-1} W_i(\xi)(I_i - H_i^W)^{-1/2}\epsilon_i\epsilon_i^T(I_i - H_i^W)^{-1/2} W_i(\xi)V_i^{-1}D_i\right\}\Omega_1^{-1},$$
$$V^{MD}_{IPW} = \Omega_1^{-1}\left\{\sum_{i=1}^n D_i^T V_i^{-1} W_i(\xi)(I_i - H_i^W)^{-1}\epsilon_i\epsilon_i^T(I_i - H_i^W)^{-1} W_i(\xi)V_i^{-1}D_i\right\}\Omega_1^{-1}, \quad (5)$$

where $H_i^W = D_i\Omega_1^{-1}D_i^T W_i(\xi)V_i^{-1}$. The variance estimators are computed at fixed inverse probability weights ($\hat{\xi}$) for illustrative purposes. Failing to account for the variability of weights can lead to a more conservative inference [24,32]. More details on bias-corrected variance estimators for IPW estimators are provided in Section 1.3 of Supplementary Materials.

Compared to the IPW estimator, the AIPW estimator improves efficiency and robustness against a misspecified working missingness model via an appropriate augmentation for missingness [31,32]. We denote the outcome model as $B_i(\eta) = B(X_{ijk}, Z_{ijk};\eta)$ for a regression function B and a vector of parameters η. If the outcome regression model is correctly specified, $B(X_{ijk}, Z_{ijk};\eta) = \mathrm{E}(Y_{ijk}|X_{ijk}, Z_{ijk})$. Let $\theta = (\beta, \xi, \eta)$. We solved the following AIPW estimating equation with augmentation for missingness:

$$\sum_{i=1}^n U^M_{i,AIPW}(\theta) = \sum_{i=1}^n [D_i^T V_i^{-1} W_i(\xi)\{Y_i - B_i(\eta)\} + D_i^T V_i^{-1}\left\{B_i(\eta) - \mu_i(\beta)\right\}] = 0. \quad (6)$$

The AIPW estimator exhibits double-robustness that produces consistent estimates for β, when either the outcome model B or missingness model π is specified correctly [31,32] given the mean model is correct. Prague et al. [25] proposed a modified AIPW estimator to augment missingness under CD-MAR and the treatment allocation mechanism in a two-level CRT:

$$\sum_{i=1}^n U^{MA}_{i,AIPW}(\theta) = \sum_{i=1}^n [D_i^T V_i^{-1} W_i(\xi)\{Y_i - B_i(\eta)\} + \sum_{a=0,1} p^a(1-p)^{1-a} D_i(a)^T V_i^{-1}(a)\left\{B_i(a;\eta) - \mu_i(a;\beta)\right\}] = 0, \quad (7)$$

where $p = \mathrm{P}(A = a)$, for $a = 0, 1$. The simulation study in Prague et al. [25] showed that the AIPW estimating Eq. (7) could perform more effectively than the AIPW estimating equation augmented for missingness alone (6), with covariate interference, in case where the covariates of a participant could impact outcomes of other participants within the same clinic (that is, for any two participants j and k, $j \neq k$ from the same clinic i, $\mathrm{E}(Y_{ij}|X_{ij}) \neq \mathrm{E}(Y_{ij}|X_{ij}, X_{ik})$). Although we have not considered covariate interference in our study, Prague's AIPW estimator is expected to be beneficial due to the additional augmentation of the treatment allocation mechanism in the presence of baseline imbalances among clinics.

To compute the variance estimator of AIPW estimating equations, a vector of estimating functions involving nuisance parameters (ξ, η)

Table 1
Details of the simulation settings. Two sets of data were generated for each simulation setting(# of clinics and missing rates). First, data under the null (no treatment-by-time interaction effect) were generated to investigate the type I error rates. Second, data under the alternative (significant treatment-by-time interaction effect) hypothesis were generated to investigate the power and coverage probabilities. In all scenarios, the clinic size varies, ranging from 30 to 300 participants (an average of 165 per clinic with cv = 0.5). In all data generated under CD-MAR, within-clinic and within-participant ICCs were approximately 0.04 and 0.4, respectively. In all data generated under OCD-MAR, within-clinic and within-participant ICCs were around 0.044 and 0.48, respectively.

Missing mechanism	# of clusters	% of missing	Description
CD-MAR	10	25%	Small # of clinics, 25% missing under CD-MAR
		50%	Small # of clinics, 50% missing under CD-MAR
	20	25%	Moderate # of clinics, 25% missing under CD-MAR
		50%	Moderate # of clinics, 50% missing under CD-MAR
	50	25%	Large # of clinics, 25% missing under CD-MAR
		50%	Large # of clinics, 50% missing under CD-MAR
OCD-MAR	10	25%	Small # of clinics, 25% missing under monotone OCD-MAR
		50%	Small # of clinics, 50% missing under monotone OCD-MAR
	20	25%	Moderate# of clinics, 25% missing under monotone OCD-MAR
		50%	Moderate # of clinics, 50% missing under monotone OCD-MAR
	50	25%	Large # of clinics, 25% missing under monotone OCD-MAR
		50%	Large # of clinics, 50% missing under monotone OCD-MAR

in the missingness and outcome models is used [25]. The estimating function for θ is

$$\psi_i(\theta) = \left(U_{i,AIPW}(Y_i, R_i, X_i, Z_i; \beta, \xi, \eta),\ S_i^W(X_i, Z_i; \xi),\ S_i^B(X_i, Z_i; \eta) \right)^T, \tag{8}$$

where S^W and S^B are score equations of ξ and η in the missingness and outcome models, respectively. We estimate variance empirically using the nuisance-adjusted sandwich variance estimator proposed by Prague et al. [25]:

$$\widehat{\mathrm{Var}}(\hat{\theta}) = \left[\left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \psi_i(\hat{\theta})}{\partial \theta} \right\}^{-1} \right]^T \left[\frac{1}{n} \sum_{i=1}^{n} \psi_i(\hat{\theta}) \psi_i(\hat{\theta})^T \right] \left[\left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \psi_i(\hat{\theta})}{\partial \theta} \right\}^{-1} \right]. \tag{9}$$

We then modify (9) by using it in combination with two bias-corrected methods. Let $H_i^{AW} = D_i \Omega_2^{-1} D_i W_i(\xi) V_i^{-1}$, where $\Omega_2 = \sum_{i=1}^{n} D_i^T V_i^{-1} W_i(\xi) D_i$. For KC bias-corrected variance, we replace $U_{i,AIPW}$ in (8) with $U_{i,AIPW}^{KC} = D_i^T V_i^{-1} W_i(\xi)(I_i - H_i^{AW})^{-1/2}\{Y_i - B_i(\eta)\} + \sum_{a=0,1} p^a(1-p)^{1-a} D_i(a)^T V_i^{-1}(a)\{I_i - H_i^{AW}(a)\}^{-1/2}\{B_i(a;\eta) - \mu_i(a;\beta)\}$. For MD bias-correction, the term $(I_i - H_i^{AW})^{-1/2}$ is replaced with $(I_i - H_i^{AW})^{-1}$. The R package `CRTgeeDR` [34] implements similar IPW and AIPW methods with the bias correction of Fay and Graubard (FG) [35] for two-level CRTs under CD-MAR.

We also considered estimating equations under the OCD-MAR mechanism. We modified the IPW and AIPW estimating equations from Tsiatis [32] and Seaman and Copas [24] for longitudinal CRTs and bias-corrected variance estimations. We implemented two additional estimators: Weighted GEE (WGEE) and GEE with mean imputation (MIGEE) using Paik's imputation approach [36]. We present the modified IPW and AIPW estimating equations and bias-corrected variance estimators for OCD-MAR in Section 1.4 of Supplementary Materials.

4. Simulation study

4.1. Simulation settings

We conducted two sets of simulation studies to evaluate the finite sample performance of the equations that estimate the treatment-by-time interaction effect in the marginal mean model (1) in longitudinal CRTs, when data follow either CD-MAR or monotone OCD-MAR. We compared the following methods: (1) standard GEE, (2) inverse probability weighted estimating equation (IPW), (3) augmented inverse probability weighted estimating equation for the missingness(AIPW^M), combined with the robust, KC, and MD bias-corrected variance estimators. For the first simulation study under CD-MAR, we also compared the AIPW estimating equation for the missingness and treatment allocation mechanism (AIPW^{MA}). As well, for the second simulation study under monotone OCD-MAR, we compared the WGEE and GEE with Paik's imputation approach. All of the estimators are shown in Table 1 of Supplementary Materials.

We considered different simulation scenarios by varying (1) the number of clinics ($n = 10$ for small, 20 for moderate, or 50 for large) and (2) the overall missing rate in outcomes (25% for moderate and 50% for severe). We generated some baseline imbalance between groups. As observed in the OH study, clinic sizes is variable, ranging from 30 to 300 participants (an average of 165 per clinic with cv=0.5), and there may be up to five visits per participant. We performed the Monte Carlo (MC) simulation 1,000 times and we rejected the null hypothesis of no treatment-by-time interaction effects, using the Wald t-statistic at a significance level of 0.05. The degree of freedom is approximated by Satterthwaite's method [37] unless specified otherwise. Table 1 summarizes all simulation scenarios. We provide details of the simulation settings in Section 2 of Supplementary Materials.

4.2. Simulation results

Results under CD-MAR: In Table 2, the bias and variance estimates are shown, simulating the results of both a small and moderate number of clinics under the null hypothesis, with the overall missing rates of 25% and 50%. Results of simulations with a large number of clinics are given in Table 4 of Supplementary Materials. The IPW and AIPW estimating equations under the correct working models improve the bias of the GEE method, which is associated with violating the MCAR assumption. The doubly robust property protects the AIPW method from a significant bias when either the missingness or outcome model is correctly specified. The results show large biases in the IPW method with the misspecified working missingness model and in the AIPW method with both misspecified working missingness and outcome models. Compared to the empirical variance estimates of $\hat{\beta}_3$, denoted by "MC Var", robust variance estimators tend to underestimate the variance, given a small number of clinics. The underestimation is particularly severe for the GEE method, at a high missingness rate. Both bias-corrected estimators improve variance estimation over robust estimators, with more conservative results using the MD bias-correction.

The type I error rates, powers, and coverage probabilities of the 95% confidence interval for treatment-by-time interaction tests are provided in Figs. 1 and 2 for $n = 10$ and 20. The results obtained when $n = 50$ are shown in Figure 1 of Supplementary Materials. Given a small number of clinics, the GEE method shows severely inflated type I error

Table 2
The results of the simulation study with the number of clinics **n = 10** and **20** under the null hypothesis of zero treatment-by-time interaction effect are shown at the overall missing rate of 25% and 50% under **CD-MAR**. Three estimating equations are compared: (standard) generalized estimating equations(GEE), inverse probability weighted estimating equations (IPW), and inverse probability weighted estimating equations augmented for missingness and the treatment allocation mechanism (AIPW^{MA}). Three variance estimators are compared for each estimating equation: Robust estimator and bias-corrected sandwich variance estimators using KC and MD. GEE.F denotes the result of standard GEE applied to the complete data set (no missing outcomes). "M", "O", and "MO" denote estimators with an incorrectly specified working missingness model, an incorrectly specified working outcome regression model, and both incorrectly specified working models. Average bias estimates (Bias), Monte Carlo variance estimates (MC variance), and average estimated variances of $\hat{\beta}_3$ using the three variance estimators (Estimated variance) over 1000 MC replications are reported. MC Var is the benchmark measurement of the variance estimation. Estimators with **zero bias and variance estimation close to the MC Var are more desirable**. Estimators with significant bias and/or underestimated variance due to missingness mechanism violation or a small number of clusters are highlighted in bold.

Method	Number of clinics = 10					Number of clinics = 20				
	Bias	MC	Estimated variance			Bias	MC	Estimated variance		
		Var (benchmark)	Robust	KC	MD		Var (benchmark)	Robust	KC	MD
25% missing										
GEE.F	−0.002	0.006	**0.005**	0.007	0.009	−0.001	0.003	0.003	0.003	0.004
GEE	**−0.099**	0.011	0.007	0.009	0.013	**−0.097**	0.006	0.004	0.005	0.005
IPW	−0.002	0.029	**0.02**	0.043	0.091	−0.007	0.013	**0.011**	0.019	0.031
IPW.M	**−0.164**	0.020	0.027	0.026	0.032	**−0.163**	0.009	0.016	0.016	0.016
AIPW^{MA}	0.002	0.016	**0.013**	0.029	0.061	−0.004	0.007	0.007	0.012	0.02
AIPW^{MA}.M	−0.003	0.010	0.014	0.014	0.015	−0.002	0.005	0.008	0.008	0.008
AIPW^{MA}.O	−0.009	0.022	**0.014**	0.036	0.077	−0.014	0.009	0.007	0.013	0.022
AIPW^{MA}.MO	**−0.155**	0.016	0.018	0.017	0.018	**−0.155**	0.007	0.01	0.01	0.01
50% missing										
GEE.F	−0.001	0.006	**0.005**	0.006	0.009	0	0.003	0.003	0.003	0.004
GEE	**−0.133**	0.022	0.007	0.01	0.013	**−0.127**	0.012	0.004	0.005	0.005
IPW	−0.011	0.107	**0.071**	0.472	1.854	−0.005	0.063	**0.051**	0.189	0.471
IPW.M	**−0.205**	0.076	0.121	0.119	0.145	**−0.19**	0.040	0.075	0.077	0.083
AIPW	0.001	0.028	**0.018**	0.111	0.431	0.001	0.016	**0.012**	0.043	0.105
AIPW^{MA}.M	0	0.015	0.022	0.024	0.03	0.003	0.008	0.013	0.014	0.016
AIPW^{MA}.O	−0.031	0.041	**0.019**	0.116	0.408	−0.028	0.024	**0.012**	0.052	0.13
AIPW^{MA}.MO	**−0.185**	0.031	0.028	0.03	0.036	**−0.181**	0.016	0.016	0.017	0.019

rates, even in combination with MD bias-correction, particularly at high missing rates. The IPW and AIPW estimating equations with the bias-corrected variance estimators successfully control the type I error rate at 0.05 under correctly specified working models. The doubly robust property ensures good performance of AIPW estimating equations when either the working missingness or outcome model is correctly specified. Overall, the AIPW method is more powerful than the IPW method at a similar type I error rate, with coverage probabilities closer to the nominal level (0.95). The gaps in performance measures among the variance estimators decrease as the number of clinics increases, which is consistent with the general theory of bias-corrected methods [6,29,30]. The results of the AIPW method augmented for missingness are similar but slightly worse than those from the AIPW method augmented for both missingness and treatment allocation mechanism.

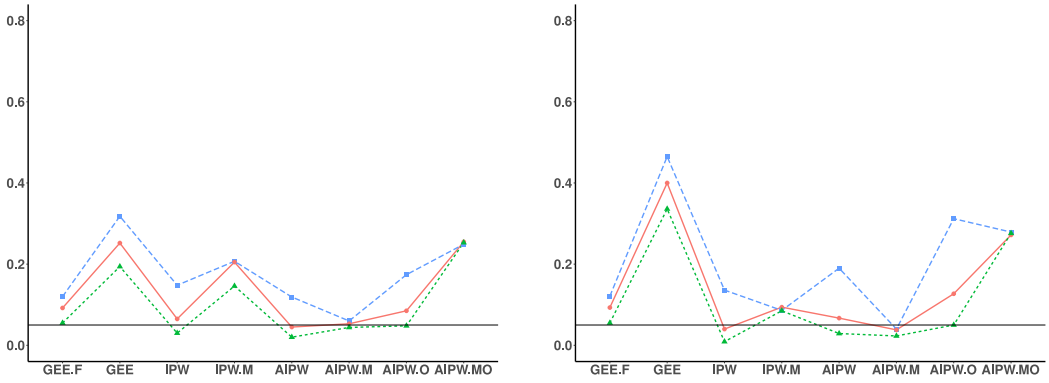
In our simulation study, the main cause of bias in estimating β depends on the estimating equations: in GEE, due to the violation of MCAR; in the IPW methods, due to the misspecified working missingness model; and in the AIPW method, when working models for both missingness and outcome models are incorrectly specified. The robust variance estimator tends to underestimate the variance of $\hat{\beta}$ given a small number of clusters. Consequently, the biases in regression parameters and variance estimation associated with missingness mechanisms, a small number of clusters, and misspecified working models produce poor type I error rate, power, and coverage probability. The IPW and AIPW methods combined with bias-corrected variance estimators could reduce such biases and improve the inference.

Results under OCD-MAR & monotone missingness: The overall finite sample performance of the proposed methods under monotone OCD-MAR is similar to that under CD-MAR; hence so, we report a few new observations. When the probability of missingness is highly dependent on previously observed outcomes under the monotone missing mechanism, the type I error of GEE and all estimating equations with incorrectly specified working models is extremely high, even in combination with a conservative MD bias-corrected estimator. Unlike results with CD-MAR data, the bias estimate in the IPW method is relatively smaller than that in other estimators when the working missingness model is incorrectly specified. Including participants who completed all visits might reduce the impact of the misspecified missingness model, although it leads to a slight loss of efficiency under the correctly specified models in the IPW method. We also note that the finite sample performance of the IPW approach under the monotone OCD-MAR mechanism depends not only on missing rates but also on the pattern of missing data. We observed that the missingness model using logistic regression performs poorly in terms of estimating missingness probability when the number of missing cases at time $t+1$ among those observed at time t is small. The poor performance is due to the fact that the maximum likelihood estimation using the logistic regression model produces a bias with extremely unbalanced outcomes (in our setting, significantly few missing cases are observed at time $t+1$ among those with observed outcomes at time t). We used the penalized logistic regression approach proposed by Firth [38] (see Section 1.5 of Supplementary Materials) through R packages `brglm` [39] to reduce bias, with very few missing cases between measurement times. More simulation results regarding monotone OCD-MAR can be found in Section 2.2, Tables 6–8, and Figures 2–4 in Supplementary Materials.

5. Data analysis: Application to the optimal health study

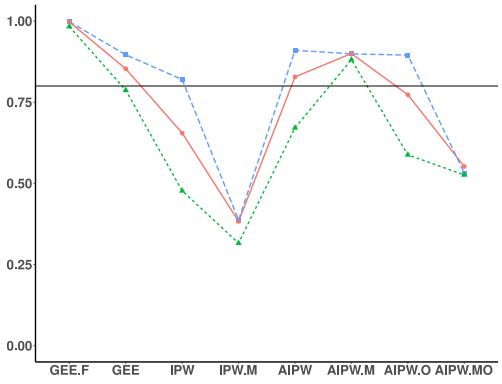
We applied the proposed methods to infer treatment effects on three outcomes: PAM, QSF, and PACIC score. PAM is the primary outcome, and the analysis results of PAM using the GLMM with the robust variance estimator can be found in Schuster et al. [3]. QSF and PACIC are secondary outcomes; the analysis results were not presented in Schuster et al. [3]. Building on the work of Schuster et al. [3], we tested the data for treatment-by-time interaction effects.

Based on our knowledge of the OH study and similar CTRs reported in the literature, we tested the data assuming the MAR mechanism. We applied the proposed methods for the CD-MAR mechanism because (1) nonmonotone missingness patterns are observed from the three

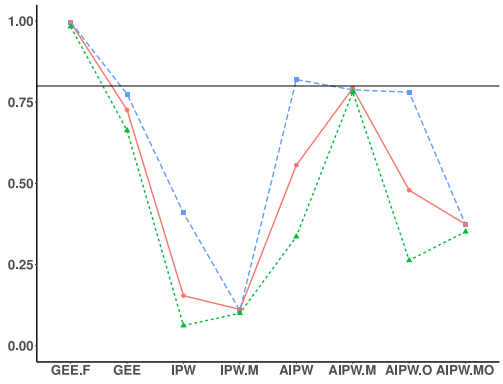


(a) Type 1 error rate: 25% missing

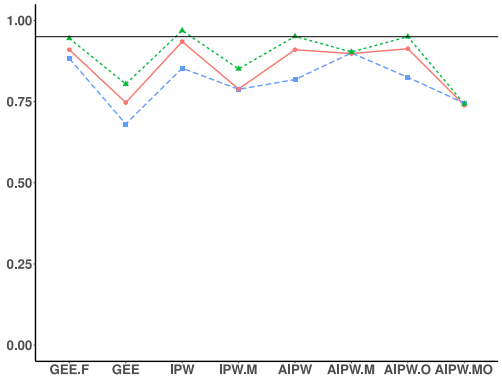
(b) Type 1 error rate: 50% missing



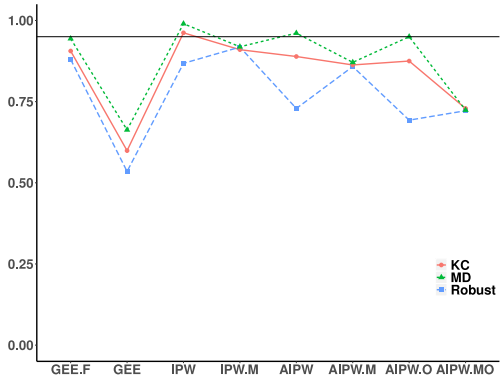
(c) Power: 25% missing



(d) Power: 50% missing



(e) Coverage probability: 25% missing

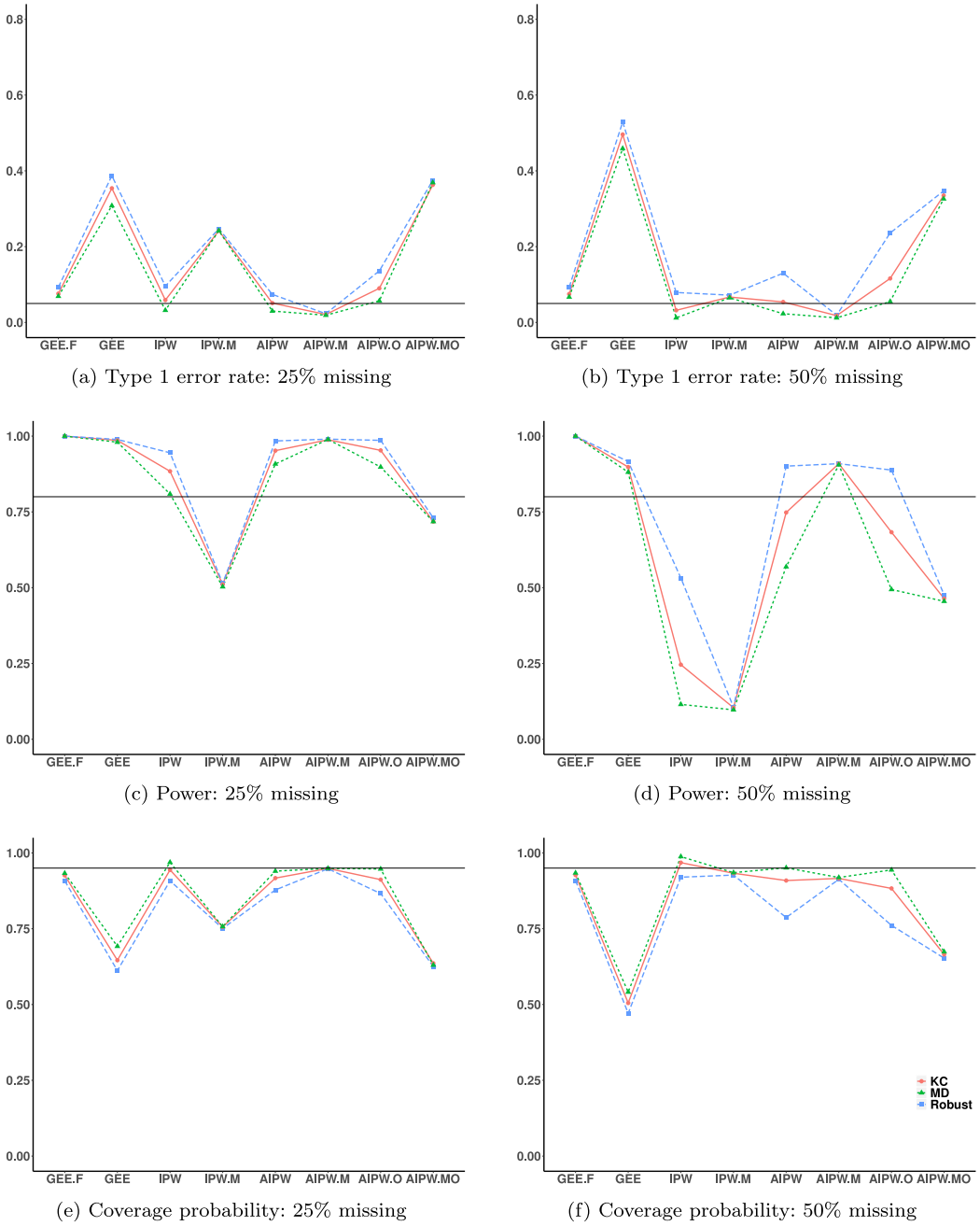


(f) Coverage probability: 50% missing

Fig. 1. The simulation results with **n = 10** at the overall missing rate of 25% and 50% under **CD-MAR**. Three estimating equations are compared: (standard) generalized estimating equation(GEE), inverse probability weighted estimating equation (IPW), and IPW augmented for missingness and treatment allocation mechanism (AIPW). For each estimating equation, three variance estimators are compared: The robust estimator (Robust: Blue) and bias-corrected sandwich estimators using KC (KC:Red) and MD (MD: Green). GEE.F denotes the result of standard GEE applied to the complete data set (no missing outcomes). "M", "O", and "MO" denote estimators with an incorrectly specified working missingness model, working outcome model, and both working models, respectively.

outcomes (Tables 11-14, Supplementary Materials) and (2) multiple baseline covariates are significantly associated with the missingness of three outcomes (Table 15, Supplementary Materials). Certain baseline outcomes are associated with the missingness of the outcomes at 6 months in the case of PAM and QSF. Nevertheless, it was challenging to model OCD-MAR for the nonmonotone missingness patterns presented. Exchangeable and AR(1) correlation structures were used for within-clinic and within-participant correlations, respectively. Exploratory analyses show non-linear change patterns in three outcomes over time. Therefore, the variable time was treated as discrete, and the four degrees of freedom F-test was used for testing treatment-by-time interaction effects at the significance level of 0.05. The mean model for three outcomes is $\mu(X;\beta) = \beta_0 + \beta_1 A + \beta_2 M_6 + \beta_3 M_{12} + \beta_4 M_{18} + \beta_5 M_{24} + A(\beta_6 M_6 + \beta_7 M_{12} + \beta_8 M_{18} + \beta_9 M_{24})$, where $A = \mathbf{1}\{\text{Patient Self-Directed Care}\}$ and M_t indicates the follow-up months: 6, 12, 18, and 24.

For the missingness and outcome working models, we considered the 14 baseline covariates that were included in the regression model described by Schuster et al. [3]: age, race, gender, the severity level of mental illness, 80% Medicaid eligibility in the 12 months from baseline, indicators of anxiety disorder and substance use, engagement level in interventions, four interpersonal support evaluation list (ISEL) scores (self-esteem, belonging, appraisal, tangible), medical stability score, and the indicator of the social security income with Medicare.



(a) Type 1 error rate: 25% missing

(b) Type 1 error rate: 50% missing

(c) Power: 25% missing

(d) Power: 50% missing

(e) Coverage probability: 25% missing

(f) Coverage probability: 50% missing

Fig. 2. The simulation results with **n = 20** at the overall missing rate of 25% and 50% under **CD-MAR**. Three estimating equations are compared: (standard) generalized estimating equation(GEE), inverse probability weighted estimating equation (IPW), and IPW augmented for missingness and treatment allocation mechanism (AIPW). For each estimating equation, three variance estimators are compared: The robust estimator (Robust: Blue) and bias-corrected sandwich estimators using KC (KC:Red) and MD (MD: Green). GEE.F denotes the result of standard GEE applied to the full data set (no missing outcomes). "M", "O", and "MO" denote estimators with an incorrectly specified working missingness model, working outcome model, and both working models, respectively. The type I error rates under $H_0 : \beta_3 = 0$, and the power and coverage probability of the 95% confidence interval under $H_1 : \beta_3 = 0.4$ are computed using 1,000 MC replications.

We present the analysis results for 1,120 participants (91.13%) with complete baseline covariates. We provide more information regarding the OH study and preliminary analysis results in Section 3.1 of Supplementary Materials.

In applying the methods, we found some discrepancies between the test results of treatment-by-time interaction effects in the three outcomes. There was strong evidence of the different effects of treatment on changes in PAM scores. In all estimating equations using the robust and KC bias-corrected variance estimators, the data indicate a significantly different effect of the treatment on the changes in PAM scores over time (Table 3). This result is consistent with the results described by Schuster et al. [3], in which the GLMM was used with the robust variance estimator. The test results remain significant for all estimators except for the IPW method in combination with the MD bias-corrected estimator. The changes in QSF score differed significantly for the two treatment groups in all estimating equations except the IPW method, even in combination with the most conservative MD bias-correction (Table 16, Supplementary Materials). The difference is not large enough when the IPW method is used with KC or MD bias-corrections. The changes in PACIC score differed significantly with treatment groups over time when GEE or either of the two AIPW methods were combined with robust or KC bias-corrected estimators (Table 17, Supplementary

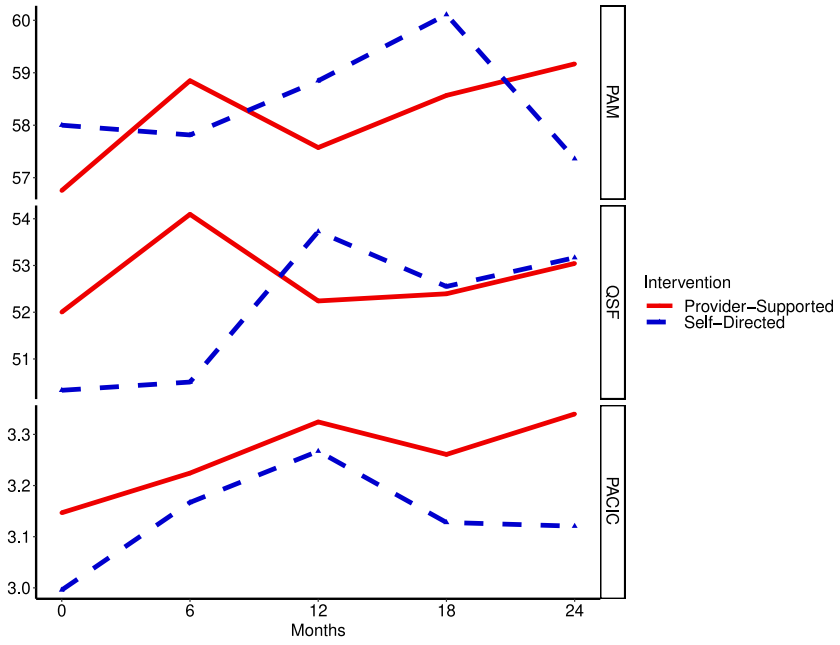


Fig. 3. Estimated mean score changes over time in the Optimal Health study by intervention groups: PAM, QSF, and PACIC scores. The mean curve is computed by plugging in the estimated regression coefficients to the mean model using the results of AIPW^{MA}.

Materials). The difference is not large enough when either the MD bias-corrected method or the IPW method (regardless of the variance estimators) is used. The curves of the predicted mean values, based on the AIPW^{MA} results, are given in Fig. 3. The curves indicate that the PS group is expected to show a more rapid increase, initially, in the PAM and QSF scores at first, and this increase is then sustained over time. The PAM and QSF scores are expected to increase slowly in the SD group.

We learned from the simulation results and previous reports that the GEE approach could lead to an inflated type I error rate, even with the MD bias-corrected estimator, when the number of clinics is small (for example, $n = 10$) or the missing rate in outcomes is high. The results of the AIPW methods, combined with bias-corrected estimators, could be conservative and reduce the risk of inflated type I errors.

6. Conclusion

The results of this study provide an in-depth comparison of the performance of the three most popular estimating equations combined with three variance estimators to analyze data from three-level longitudinal CRTs. We utilized empirical evaluations under pragmatic constraints; that is, a small number of clusters, heterogeneous cluster sizes, and missing outcomes under two different missingness mechanisms. Our simulation results show that the proposed IPW methods with bias-corrected variance estimators can improve bias due to the violation of the MCAR assumption and underestimated variance associated with a small number of clinics in GEE combined with the robust variance estimator in the considered complex settings. Given its doubly-robust property, we recommend the AIPW estimating equations combined with KC bias-correction when the number of clinics is moderately small, in the presence of moderate missing outcomes under MAR, and with a moderate variation in clinic sizes. The AIPW estimating equations combined with MD bias-correction would be more appropriate for preventing the inflation of false positive findings when the number of clinics is as small as 10, with severe missing outcomes. The MD bias-correction can compensate for the variability of the robust estimator due to its overcorrection for the bias associated with a small number of clinics and severe missing rates [19].

We also investigated whether and how the performance of the proposed methods under the three-level longitudinal CRTs with more pragmatic constraints can differ from the results of previous reports in relatively simpler settings, such as two-level CRTs. It is difficult to compare directly with the results in published studies, due to deviations in several features. However, the results in our study suggest that the IPW and AIPW estimating equation approaches with the bias-correction method can improve the inference of treatment effects in three-level longitudinal CRTs; the performances of these methods under more pragmatic constraints (with a small number of clusters and missing outcomes) are quite consistent with the results of previous reports in relatively simpler settings. We demonstrated the double robustness of the AIPW estimating equation in a more complex design than that of Prague et al. [25]; specifically, three-level longitudinal CRTs under the CD-MAR and monotone OCD-MAR mechanisms. Although the MD bias-correction method tends to overestimate the variance compared to the KC bias-correction method, it performs well with respect to type I error rate and coverage probability. This result is consistent with the results obtained in a simpler design by Lu et al. [19]. Our findings with the KC and MD bias-correction methods can differ from those of certain earlier reports due to several different features, such as heterogeneous clinic sizes, three-level hierarchical designs, missing longitudinal outcomes under MAR, and the use of AIPW methods. For example, in the absence of missing outcomes, Li et al. [6] recommended that the standard GEE with KC bias-correction would be preferred when using data with a small to moderate variation of cluster sizes (cv < 0.6), even for 10 clusters; otherwise, more conservative methods, such as FG bias-correction [35], should be considered. Our results show that the underestimated variance created by a small number of clusters is still significant for 10 clusters and moderate variation in cluster sizes when the KC-bias correction is used to analyze data from the three-level longitudinal CRTs with a moderate to substantial number of missing outcomes under the two MAR mechanisms.

Our study has several notable limitations that should be considered in future research. We considered a relatively simple missingness model through logistic regression. Modified estimating equations and missingness models should be considered to handle more complex missing mechanisms that may depend on time-varying covariates, OCD-MAR for nonmonotone missingness, or non-MAR mechanisms. Second, we

Table 3
Test results of the treatment-by-time interaction effect on the **Patient Activation Measure (PAM)** in the Optimal Health study. Four estimating equations are compared: (standard) generalized estimating equation (GEE), inverse probability weighted estimating equation (IPW), inverse probability weighted estimating equation augmented for missingness alone (AIPW^{M}) and missingness and treatment allocation mechanism (AIPW^{MA}). Three variance estimators are compared for each estimating equation: Robust sandwich estimator and bias-corrected sandwich variance estimators using KC and MD. CS and AR(1) correlation models are used for within-clinic and within-participant correlations, respectively. The regression model includes treatment, discrete-time variable, and their interaction term, considering the baseline as a reference level. Coefficient and standard error estimates of all regression parameters in the mean model, F-statistics, and 4 degrees of freedom Wald test result of **treatment-by-time interaction effect** are reported.

Estimator	Term	Robust			KC		MD	
		Coefficient	SE	F-value	SE	F-value	SE	F-value
GEE	Intercept	56.016	1.025		1.234		1.501	
	Treatment (SD)	1.703	1.308		1.524		1.795	
	Month of 6 (M_6)	1.949	0.531		0.674		0.870	
	Month of 12 (M_{12})	0.538	0.893		1.147		1.487	
	Month of 18 (M_{18})	1.016	1.101		1.401		1.799	
	Month of 24 (M_{24})	1.731	0.952		1.209		1.552	
	Trt ×Time: M_6	−2.344	0.660	33.136***	0.804	27.659***	0.997	22.956***
	Trt ×Time: M_{12}	0.573	1.041		1.289		1.621	
	Trt ×Time: M_{18}	0.718	1.343		1.637		2.027	
	Trt ×Time: M_{24}	−1.988	1.170		1.423		1.760	
IPW	Intercept	56.121	1.206		10.107		21.613	
	Treatment (SD)	1.967	1.746		10.733		22.560	
	Month of 6 (M_6)	2.357	0.398		0.466		1.603	
	Month of 12 (M_{12})	0.792	1.069		1.432		1.648	
	Month of 18 (M_{18})	2.298	1.265		1.550		4.641	
	Month of 24 (M_{24})	3.572	1.490		2.033		5.918	
	Trt ×Time: M_6	−2.586	0.798	6.647***	0.812	2.927*	1.839	1.482
	Trt ×Time: M_{12}	0.366	1.460		1.992		2.547	
	Trt ×Time: M_{18}	0.241	1.938		2.446		5.221	
	Trt ×Time: M_{24}	−3.835	2.157		3.152		6.810	
AIPW^{M}	Intercept	56.601	0.784		3.837		6.440	
	Treatment (SD)	1.518	0.966		4.170		7.009	
	Month of 6 (M_6)	2.095	0.259		0.184		0.565	
	Month of 12 (M_{12})	0.820	0.633		0.771		0.794	
	Month of 18 (M_{18})	1.814	0.736		0.670		1.677	
	Month of 24 (M_{24})	2.414	0.747		0.859		2.400	
	Trt ×Time: M_6	−2.281	0.432	14.808***	0.432	8.690***	0.764	4.556**
	Trt ×Time: M_{12}	0.028	0.677		0.889		1.063	
	Trt ×Time: M_{18}	0.294	0.929		0.939		1.851	
	Trt ×Time: M_{24}	−3.054	0.941		1.260		2.744	
AIPW^{MA}	Intercept	56.754	0.632		2.908		4.791	
	Treatment (SD)	1.246	0.757		3.409		5.732	
	Month of 6 (M_6)	2.095	0.282		0.171		0.486	
	Month of 12 (M_{12})	0.820	0.693		0.842		0.697	
	Month of 18 (M_{18})	1.814	0.720		0.664		1.678	
	Month of 24 (M_{24})	2.414	0.733		0.879		2.400	
	Trt ×Time: M_6	−2.281	0.493	13.025***	0.412	10.927***	0.581	6.745***
	Trt ×Time: M_{12}	0.028	0.692		0.921		0.949	
	Trt ×Time: M_{18}	0.294	0.918		0.944		1.788	
	Trt ×Time: M_{24}	−3.054	1.006		1.168		2.479	

Using 4 degrees of freedom F-test for the treatment-by-time interaction effect.
***Denote significance at the 0.001 levels, respectively.
**Denote significance at the 0.01 levels, respectively.
*Denote significance at the 0.05 levels, respectively.

assume that participants within the same clinic have an exchangeable correlation in the analysis of the OH study. The data of care managers for individual participants and visits are unavailable in the OH study. If data on care managers are available, one can extend our methods for a four-level hierarchical design to account for within-care manager correlations. Finally, no attempt was made to extend the methods to study more complex covariance structures due to the difficulty in estimating the variance components without a normality assumption. It is important to model the correlation structure correctly to achieve the efficiency of the robust variance estimators in longitudinal CRTs. Further research is warranted to handle these problems and will be left for future study.

Despite these limitations, our results can be useful for choosing the appropriate methods for preventing potential biases and type I error inflation due to the MAR missingness and the small number of clusters for the longitudinal CRT analysis. Another area deserving further work is the sample size calculation for three-level longitudinal CRTs. For example, the sample size calculation has been commonly done based on the asymptotic normality assumption of the Z-test statistics using model-based variance estimators. The AIPW methods with bias-corrected variance estimators can be used to analyze data to prevent the inflation of false positive rates. This may lead to a lower actual statistical power than the target power estimated using the Z-statistics and model-based variance estimators, given the missing outcomes and the small number of clinics for the large variation in clinic sizes. In such a case, a carefully designed MC simulation using the proposed method can help to estimate the statistical power more accurately.

Data availability statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Acknowledgments

The authors are immensely grateful to Dr. Sally C. Morton, Executive Vice President of Knowledge Enterprise at Arizona State University,

for her helpful comments. The authors gratefully acknowledge the support from Tracy Carney, CPS, CPRP, a Senior Recovery/ Resiliency Specialist-Community Care Behavioral Health, and Jong H. Jeong, Ph.D., a Professor and Vice Chair of the Department of Biostatistics at the University of Pittsburgh. We gratefully acknowledge the support from the University of Pittsburgh Center for Research Computing through the resources provided.

Funding

This work was supported in part by the grant from the University of Pittsburgh Central Research Development Fund and by the University of Pittsburgh Center for Research Computing through the resources provided (NIH S10OD028483), and Patient-Centered Outcomes Research Institute (PCORI) Award 271. The statements in this publication are solely the responsibility of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors, or the Methodology Committee.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.conctc.2023.101194. We have provided additional information to the supplementary online materials.

References

[1] The National Institute of Mental Health, Past year prevalence of serious mental illness (SMI) among U.S. adults (2016) [Internet], 2017, https://www.nimh.nih.gov/health/statistics/mental-illness.shtml#part_154788/. (Online; Accessed 9-October-2018).

[2] The National Institute of Mental Health, Mental health treatment received in past year among U.S. adults with serious mental illness (2016)[Internet], 2017, https://www.nimh.nih.gov/health/statistics/mental-illness.shtml#part_154788/. (Online; Accessed 9-October-2018).

[3] J. Schuster, C. Nikolajski, J. Kogan, C. Kang, P. Schake, T. Carney, S.C. Morton, C.F. Reynolds III, A payer-guided approach to widespread diffusion of behavioral health homes in real-world settings, Health Aff. 37 (2) (2018) 248–256.

[4] J. Kogan, J. Schuster, C. Nikolajski, P. Schake, T. Carney, S.C. Morton, C. Kang, C.F. Reynolds III, Challenges encountered in the conduct of optimal health: A patient-centered comparative effectiveness study of interventions for adults with serious mental illness, Clin. Trials 14 (1) (2017) 5–16.

[5] R. Platt, S.U. Takvorian, E. Septimus, J. Hickok, J. Moody, J. Perlin, J.A. Jernigan, K. Kleinman, S.S. Huang, Cluster randomized trials in comparative effectiveness research: randomizing hospitals to test methods for prevention of healthcare-associated infections, Med. Care 48 (6) (2010) S52–S57.

[6] P. Li, D.T. Redden, Small sample performance of bias-corrected sandwich estimators for cluster-randomized trials with binary outcomes, Stat. Med. 34 (2) (2015) 281–296.

[7] B.C. Kahan, G. Forbes, Y. Ali, V. Jairath, S. Bremner, M.O. Harhay, R. Hooper, N. Wright, S.M. Eldridge, C. Leyrat, Increased risk of type I errors in cluster randomised trials with small or medium numbers of clusters: A review, reanalysis, and simulation study, Trials 17 (1) (2016) 438.

[8] S. Borhan, J. Ma, A. Papaioannou, J. Adachi, L. Thabane, Performance of methods for analyzing continuous data from stratified cluster randomized trials–a simulation study, Contemp. Clin. Trials Commun. (2023) 101115.

[9] P. Diggle, P.J. Diggle, P. Heagerty, P.J. Heagerty, K.-Y. Liang, S. Zeger, et al., Analysis of Longitudinal Data, Chapman and Hall/CRC, New York, 2002.

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

[11] C. Gatsonis, S.C. Morton, Methods in Comparative Effectiveness Research, CRC Press, 2017.

[12] J.S. Preisser, B. Lu, B.F. Qaqish, Finite sample adjustments in estimating equations and covariance estimators for intracluster correlations, Stat. Med. 27 (27) (2008) 5764–5785.

[13] K. Hemming, M. Taljaard, Sample size calculations for stepped wedge and cluster randomised trials: a unified approach, J. Clin. Epidemiol. 69 (2016) 137–146.

[14] R. Hooper, S. Teerenstra, E. de Hoop, S. Eldridge, Sample size calculation for stepped wedge and other longitudinal cluster randomised trials, Stat. Med. 35 (26) (2016) 4718–4728.

[15] A.J. Girling, Relative efficiency of unequal cluster sizes in stepped wedge and other trial designs under longitudinal or cross-sectional sampling, Stat. Med. 37 (30) (2018) 4652–4664.

[16] E.L. Turner, L. Yao, F. Li, M. Prague, Properties and pitfalls of weighting as an alternative to multilevel multiple imputation in cluster randomized trials with missing binary outcomes under covariate-dependent missingness, Stat. Methods Med. Res. 29 (5) (2020) 1338–1353.

[17] R.M. Turner, R.Z. Omar, S.G. Thompson, Bayesian methods of analysis for cluster randomized trials with binary outcome data, Stat. Med. 20 (3) (2001) 453–472.

[18] A.B. Clark, M.O. Bachmann, Bayesian methods of analysis for cluster randomized trials with count outcome data, Stat. Med. 29 (2) (2010) 199–209.

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

[20] S. Teerenstra, B. Lu, J.S. Preisser, T. Van Achterberg, G.F. Borm, Sample size considerations for GEE analyses of three-level cluster randomized trials, Biometrics 66 (4) (2010) 1230–1237.

[21] A.J. Stephens, E.J.T. Tchetgen, V. De Gruttola, Augmented GEE for improving efficiency and validity of estimation in cluster randomized trials by leveraging cluster-and individual-level covariates, Stat. Med. 31 (10) (2012) 915.

[22] M. Wang, L. Kong, Z. Li, L. Zhang, Covariance estimators for generalized estimating equations (GEE) in longitudinal analysis with small samples, Stat. Med. 35 (10) (2015) 1706–1721.

[23] W.P. Ford, P.M. Westgate, A comparison of bias-corrected empirical covariance estimators with generalized estimating equations in small-sample longitudinal study settings, Stat. Med. 37 (28) (2018) 4318–4329.

[24] S. Seaman, A. Copas, Doubly robust generalized estimating equations for longitudinal data, Stat. Med. 28 (6) (2009) 937–955.

[25] M. Prague, R. Wang, A. Stephens, E. Tchetgen Tchetgen, V. DeGruttola, Accounting for interactions and complex inter-subject dependency in estimating treatment effect in cluster-randomized trials with missing outcomes, Biometrics 72 (4) (2016) 1066–1077.

[26] D.B. Rubin, Inference and missing data, Biometrika 63 (3) (1976) 581–592.

[27] J.D. Kloke, J.W. McKean, M.M. Rashid, Rank-based estimation and associated inferences for linear models with cluster correlated errors, J. Amer. Statist. Assoc. 104 (485) (2009) 384–390.

[28] W. Pan, M.M. Wall, Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations, Stat. Med. 21 (10) (2002) 1429–1441.

[29] G. Kauermann, R.J. Carroll, A note on the efficiency of sandwich covariance matrix estimation, J. Amer. Statist. Assoc. 96 (456) (2001) 1387–1396.

[30] L.A. Mancl, T.A. DeRouen, A covariance estimator for GEE with improved small-sample properties, Biometrics 57 (1) (2001) 126–134.

[31] J.M. Robins, A. Rotnitzky, L.P. Zhao, Estimation of regression coefficients when some regressors are not always observed, J. Amer. Statist. Assoc. 89 (427) (1994) 846–866.

[32] A.A. Tsiatis, Semiparametric Theory and Missing Data, Springer Science & Business Media, New York, 2007.

[33] M.Á. Hernán, B. Brumback, J.M. Robins, Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men, Epidemiology 11 (5) (2000) 561–570.

[34] M. Prague, CRTgeeDR: Doubly robust inverse probability weighted augmented GEE estimator, 2017, URL https://CRAN.R-project.org/package=CRTgeeDR, R package version 2.0.

[35] M.P. Fay, B.I. Graubard, Small-sample adjustments for wald-type tests using sandwich estimators, Biometrics 57 (4) (2001) 1198–1206.

[36] M.C. Paik, The generalized estimating equation approach when data are not missing completely at random, J. Amer. Statist. Assoc. 92 (440) (1997) 1320–1329.

[37] F.E. Satterthwaite, An approximate distribution of estimates of variance components, Biom. Bull. 2 (6) (1946) 110–114.

[38] D. Firth, Bias reduction of maximum likelihood estimates, Biometrika 80 (1) (1993) 27–38.

[39] I. Kosmidis, brglm: Bias reduction in binary-response generalized linear models, 2019, URL https://cran.r-project.org/package=brglm, R package version 0.6.2.