Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2020, Vol. 29(5) 1338–1353

DOI: 10.1177/0962280219859915
journals.sagepub.com/home/smm
SAGE

Properties and pitfalls of weighting as an alternative to multilevel multiple imputation in cluster randomized trials with missing binary outcomes under covariate-dependent missingness

Elizabeth L Turner,[1,2] Lanqiu Yao,[3] Fan Li[1] and Melanie Prague[4,5]

Abstract

The generalized estimating equation (GEE) approach can be used to analyze cluster randomized trial data to obtain population-averaged intervention effects. However, most cluster randomized trials have some missing outcome data and a GEE analysis of available data may be biased when outcome data are not missing completely at random. Although multilevel multiple imputation for GEE (MMI-GEE) has been widely used, alternative approaches such as weighted GEE are less common in practice. Using both simulations and a real data example, we evaluate the performance of inverse probability weighted GEE vs. MMI-GEE for binary outcomes. Simulated data are generated assuming a covariate-dependent missing data pattern across a range of missingness clustering (from none to high), where all covariates are measured at baseline and are fully observed (i.e. a type of missing-at-random mechanism). Two types of weights are estimated and used in the weighted GEE: (1) assuming no clustering of missingness (W-GEE) and (2) accounting for such clustering (CW-GEE). Results show that, even in settings with high missingness clustering, CW-GEE can lead to more bias and lower coverage than W-GEE, whereas W-GEE and MMI-GEE provide comparable results. W-GEE should be considered a viable strategy to account for missing outcomes in cluster randomized trials.

Keywords

Generalized estimating equations, inverse probability weights, multilevel multiple imputation, missing data, cluster randomized trial

1 Introduction

Cluster randomized trials (CRTs), also referred to as group randomized trials or community randomized trials, are commonly used to evaluate the effectiveness of interventions. CRTs are trials in which groups (i.e. clusters) of individuals are randomized to arms of the trial and outcomes are measured on individuals within those groups. As a consequence, all individuals in each cluster receive the same treatment allocation and the unit of randomization is different to the unit of measurement. The CRT design is particularly appealing when an intervention is naturally delivered at a community-level,[1] when there are concerns of treatment inequity within a community,[2] when intervention delivery at the cluster-level is logistically easier to implement[3] or, in the case of infectious diseases, to avoid the problem of identifying the indirect effect of treatment (or herd immunity) related to exposure interference or contamination.[4] The CRT design has been adopted in diverse settings and disciplines.[1,2] Guidelines on design and reporting of CRT can be found in the CONSORT statement.[5]

[1]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA
[2]Duke Global Health Institute, Duke University, Durham, NC, USA
[3]Department of Population Health, New York University, New York, NY, USA
[4]INRIA SISTM, Inserm U1219 Bordeaux Population Health, Université Bordeaux, ISPED, Bordeaux, France
[5]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

Corresponding author:
Elizabeth L Turner, Department of Biostatistics and Bioinformatics, Duke University, 11098 Hock Plaza, 2424 Erwin Road, Durham, NC 27705, USA.
Email: liz.turner@duke.edu

1.1 Missing outcome data in CRT: Current status and methods

Mixed effects models and the generalized estimating equation (GEE) approach are the most common regression modeling approaches used to analyze CRT data.[6,7] In the current article, we focus on the GEE approach for two reasons: (1) fewer distributional assumptions are required, and (2) it provides a marginal intervention effect, whose population-averaged interpretation is preferred for making public health and policy decisions rather than the conditional, cluster-specific intervention effect estimated using mixed effects models.[8,9] Two recent systematic reviews indicated that while most CRTs (72% of 132 and 93% of 86 CRT, respectively) had missing outcomes, less than half reported that they accounted for the missing data mechanism in the primary outcome analysis.[6,10] When data are missing completely at random (MCAR), it is well-known that the analysis of all available data is unbiased.[11] (See Section 1 of Supplementary Material for missing data definitions.) When data are missing at random (MAR), for example when the mechanism is covariate-dependent missingess (CDM) with fully observed baseline covariates, the MCAR assumption for GEE can be relaxed by using weighted GEE (W-GEE),[12] imputation[13,14] or a combination of both.[15] Methodological descriptions of multiple imputation approaches for a range of outcome analyses of CRT data are available for continuous outcomes[16–19] and binary outcomes.[20–24] A subset of these articles focused specifically on multiple imputation (MI) approaches for GEE analyses of CRTs, all of which focused on binary outcomes.[20–24] All concluded that the MI approach should reflect the multilevel structure of the data in a CRT through the use of a multilevel MI approach (MMI-GEE). Moreover, all of these articles considered the CDM setting with fully observed baseline covariates and all specifically considered no clustering in the missing process, except for Caille et al.[20] Furthermore, this literature fits in the broader literature on imputation for missing outcomes for correlated data in which it is also well recognized that the multiple imputation strategy should use a multilevel structure in order to reflect the multilevel data structure.[25,26] For the alternative approach of weighting to account for missing outcomes, although there are methodological descriptions of general weighting approaches,[11] of weighted GEE for longitudinal data,[25,27–29] and methods for general correlated data,[25,30,31] we have found no articles that address weighted GEE specifically for clustered outcome data that arise in a CRT. This gap in the literature could explain why few trials implemented this approach in practice.[6,10]

1.2 Motivating data example

We will use the Health and Literacy Intervention (HALI) CRT of a school-based intervention as our motivating example. The literacy intervention provided professional development (e.g. lesson plans), training (e.g. workshops) and support (e.g. weekly text messages) for teachers. One of the goals of the HALI CRT was to evaluate the impact of the literacy intervention on child educational outcomes.[3] In brief, 101 primary schools in coastal Kenya were randomized to intervention (51 schools) or control (50 schools). A cohort of 2539 children was recruited from the 101 schools (approximately 25/school) and literacy outcomes were assessed at baseline, 9- and 24-month follow-up. Approximately 12% and 20% of children were missing outcomes at each follow-up, respectively. Analyses of all available data using likelihood-based mixed effects models indicated that the intervention was effective at improving child literacy outcomes.[32] Missing data sensitivity analyses in that framework indicated no evidence of bias due to missing outcome data.[32] In the current article, we focus on the nine-month outcomes and a data set with fully observed baseline covariates, and turn our attention to a setting where the estimand of interest is a population-averaged (marginal) intervention effect, estimated using GEE.

1.3 Objectives

In the context of GEE analysis of CRTs with binary outcomes, the overall goal of the current manuscript is to compare the performance of weighted GEE to the more commonly used MMI-GEE approach in a CDM missing data setting where the covariates are collected at baseline and all are fully observed. In this context, the CDM mechanism is a special case of an MAR mechanism, and we note that if some covariate values were missing, it would be a type of MNAR mechanism. We use both real data analysis and simulation to demonstrate and evaluate the properties of weighted GEE in contrast to MMI-GEE for the analysis of CDM missing outcomes. In an additional novelty, we consider CDM mechanisms both with and without clustering and consider two forms of weighted GEE: the first, the standard approach where clustering is not accounted for in generating the weights (W-GEE) and the second where it is accounted for (CW-GEE). In Section 2, we will first provide a brief description of GEE, MMI-GEE, W-GEE and CW-GEE. In Sections 3 and 4, we provide an in-depth analysis

of our motivating HALI data set and a simulation study, respectively. Overall, the goal of the article is to provide the reader with guidance on how to use weighted GEE to analyze CRTs with missing outcomes, to demonstrate that it is a viable alternative to MMI-GEE, whilst also providing a description of drawbacks and pitfalls of the approach.

2 Theory and methods

We consider the two-arm cohort CRT design ($A_i = 1$ for intervention, $A_i = 0$ for control) with a total of M clusters, variable number of participants per cluster (n_i), a single follow-up outcome measurement, Y_{ij}, for the jth individual ($j = 1, \ldots, n_i$) in the ith cluster ($i = 1, \ldots, M$), with $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T$, a column-vector of the n_i outcomes in the ith cluster. Moreover, as noted above, our goal is to estimate the unadjusted marginal intervention effect because the population-averaged interpretation of this effect is useful for making public health and policy decisions.[8]

2.1 GEE analysis of CRT data with no missing outcomes

The semi-parametric GEE approach[33,34] estimates parameters of a generalized regression model (equation (1)) for the mean function of interest $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{in_i})^T$ with link function g expressed as

$$g(\mu_{ij}) = \theta_0 + \theta_A A_i \tag{1}$$

The (unadjusted) intervention effect (θ_A) is typically reported as a mean difference between arms for continuous outcomes (using identity-link, g) and as a prevalence (or risk) ratio or prevalence (or incidence) odds ratio for binary outcomes (using log or logit link for g, respectively). In the current article, we focus on the latter, namely binary outcomes with the logit link for which the intervention effect is naturally quantified as an odds ratio. Because of the well-known "non-collapsibility" property of the odds ratio, adjustment for a baseline covariate that is balanced between arms (i.e. which does not act as a confounding variable) will lead to a different estimated intervention effect than that estimated from Model (1).[35]

To accommodate correlated responses, the standard, unweighted GEE approach estimates the intervention effect defined in (1) by solving the following system of equations

$$0 = \sum_{i=1}^{M} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \left(\boldsymbol{Y}_i - \boldsymbol{\mu}_i \right) \tag{2}$$

where $\boldsymbol{D}_i = d\boldsymbol{\mu}_i / d\boldsymbol{\beta}^T$ is the derivative of the marginal mean and $\boldsymbol{V}_i$ is a working covariance matrix for $\boldsymbol{Y}_i$. More specifically, $\boldsymbol{V}_i = \boldsymbol{A}_i^{1/2} \boldsymbol{R}_i \boldsymbol{A}_i^{1/2}$, where $\boldsymbol{A}_i$, a function of the marginal mean μ_{ij}, is a diagonal matrix with elements $\phi v(\mu_{ij})$, with ϕ the dispersion parameter and v the variance function, and where $\boldsymbol{R}_i = \boldsymbol{R}_i(\alpha)$ is a working correlation structure defined by the user to model correlation of outcomes on individuals within the same cluster, where α is a common correlation parameter across clusters. For CRT data, it is common to specify $\boldsymbol{V}_i$ as the identity or compound symmetric matrix, corresponding to independence or exchangeability of individuals within a cluster, respectively. The assumption of exchangeability is usually preferred over the independence structure, given that outcomes on individuals in the same cluster are expected to be correlated.[1] When there are no missing CRT outcomes, there are two key benefits of the unweighted GEE approach (equation (2)) used to estimate the intervention effect (equation (1)): consistency (i.e. the intervention effect estimate is asymptotically unbiased), and, robustness to misspecification of the working correlation structure (i.e. the intervention effect estimate is unbiased even if the working correlation matrix is different from the true correlation structure).[34] Precision is typically estimated using the sandwich variance estimator (see equation (4)[36]), to obtain the so-called "robust" standard errors (SE), given by

$$\sum_{i=1}^{M} \hat{\boldsymbol{\Omega}} \left(\boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{r}_i \boldsymbol{r}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{D}_i \right) \hat{\boldsymbol{\Omega}} \tag{3}$$

where $\hat{\boldsymbol{\Omega}} = \left(\sum_{i=1}^{M} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{D}_i \right)^{-1}$ is the model-based variance and $\boldsymbol{r}_i = \boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i$ is the residual vector for cluster i. Efficiency (i.e. smallest SE) is achieved when the true correlation structure of the outcome data is correctly specified.[36]

2.2 Complete records GEE analysis of CRT data with missing outcomes

The simplest approach to the analysis of CRT data with missing outcomes is to analyze only available data using the GEE analysis model (2). This approach is commonly referred to as a complete records analysis (CRA) when there is a single follow-up time point, the setting considered in the current article, and, as such, we refer to CRA-GEE. The CRA-GEE approach to estimate a population-averaged intervention effect is valid under MCAR mechanisms and will lead to an unbiased intervention effect. In contrast, when baseline covariates are predictive of both the probability of a missing outcome and of the outcome level itself, it is natural to consider additional adjustment for those covariates in an adjusted CRA-GEE (A-CRA-GEE) mirroring the use of such covariate adjustment in likelihood-based analytic approaches. These adjusted likelihood-based approaches provide unbiased intervention effects under a CDM missing outcome mechanism when the model is correctly specified and the effect measure is collapsible (e.g. for identity and log links).[25] Whilst an adjusted CRA-GEE analysis will provide an unbiased estimate of an adjusted population-averaged intervention effect under a CDM mechanism,[24] as mentioned earlier, it is not guaranteed to provide an unbiased estimate of an unadjusted population-averaged intervention effect with a logit link due to non-collapsibility.[35]

2.3 Multilevel multiple imputation GEE analysis of CRT data

The extensive literature on a range of MI procedures to account for missing outcomes in CRTs[16–24] and in other correlated data settings[25,27–29] indicate the importance of the imputation procedure reflecting the hierarchical outcome data structure. As such, we only consider imputation approaches that account for the clustering in outcomes in CRTs. Moreover, we adhere to the principle that the imputation model needs to be compatible with the analysis model in order to implement Rubin's rules to combine results from multiple imputed data sets.[37] Although some authors have considered fixed effects and within-cluster imputation to account for the hierarchical structure, in order to obtain good confidence coverage and nominal Type I error, a multilevel MI (MMI) procedure such as random effects MI is recommended and is the one we consider here. Given that the focus of the current article is weighted GEE, we refer the reader to Section 3.2 of Hossain et al.[24] for a more comprehensive description of MMI-GEE for the analysis of CRTs with missing binary outcomes. In brief, MMI-GEE can be considered as a three-stage approach: (1) generate K complete data sets by imputing missing outcomes using a multi-level imputation model (MMI step); (2) estimate an intervention effect (equation (1)) for each of the K complete data sets using GEE (according to equation (2)); and (3) appropriately combine the K estimated intervention effects using Rubin's rules to account for variability introduced by the imputation.[38] More specifically, MMI for binary outcomes in CRTs can be implemented by using a random-effects logistic regression model to generate outcome probabilities for each individual, from which a Bernoulli draw is implemented to impute a binary outcome in the imputation stage, i.e. stage (1). Importantly, as highlighted by Hossain et al.[24] the analysis model and the imputation model do not have to be the same but do need to be correctly specified so that, for example, interactions are specified in the models if that reflects the true functional form.

2.4 Weighted GEE analysis of CRT data with missing outcomes

Like MMI-GEE, weighted GEE for the analysis of CRT data with missing outcomes is a multi-stage approach. In contrast to MMI-GEE, which analyzes data from all individuals, weighted GEE analyzes only those with observed outcomes and provides greater weight to individuals with observed outcomes who have a low probability of being observed. Weighted GEE can be considered as a two-stage approach that involves an adaptation of (equation (2)) in the second stage. Before describing the two stages in detail, we make three simplifying assumptions which, to our knowledge, have been assumed in the previous MI literature for analyzing CRT data with missing outcomes[16–24]: (1) each cluster includes at least one observed outcome, i.e., no empty clusters; (2) the probability of missing an outcome does not depend on another participant's observed outcome (i.e., like the MMI-GEE literature, we only consider CDM); and (3) all baseline covariates are observed (i.e., there is no missing covariate data). In practice, if some baseline covariates are missing, MI could be used for these. Alternatively, if the missing covariates can reasonably be assumed to be MCAR, outcome data from those individuals can be excluded without introducing bias (although, it is important to note that MI of these covariates may be able to increase precision).[39]

In its first stage, and under our three assumptions, weighted GEE builds a propensity-score (PS) model for the probability that an individual's outcome is observed at the follow-up time-point, where we note that it could

alternatively be framed as a probability of missingness (POM) model. The PS model is typically specified as a logistic regression model such as the following

$$h(\pi_{ij}) = h\big(P\big(R_{ij} = 1|\mathbf{X_{ij}}, A_i\big)\big) = \alpha_0 + \boldsymbol{\alpha}_X^T\mathbf{X_{ij}} + \alpha_A A_i \tag{4A}$$

where $R_{ij} = 1$ indicates that the outcome of the jth individual in the ith cluster is observed, 0 that it is missing, and where $\mathbf{X_{ij}}$ is a vector of baseline covariates that are expected to be predictive of whether the outcome is observed and h is the logit function. Importantly, and as for MI, interactions must be included in the model if they are related to the probability of being observed, for example by adding terms $\boldsymbol{\alpha}_{AX}^T\mathbf{X_{ij}}A_i$ to model (4A).[19,24,40] More generally, a key assumption is that the correct functional form is used in the PS model. A weight for each individual, commonly referred to as an inverse probability weight (IPW), is then calculated as $W_{ij} = R_{ij}/\hat{\pi}_{ij}$ where $\hat{\pi}_{ij}$ can be estimated using the PS model (4A). We consider two versions of the PS model, namely (4A) and (4B), where the latter accounts for potential clustering of the probability of being observed by including a random intercept, u_i for cluster i within in the following random effects logistic regression model

$$h(\pi_{ij}) = h\big(P\big(R_{ij} = 1|\mathbf{X_{ij}}, A_i\big)\big) = \gamma_0 + \boldsymbol{\gamma}_X^T\mathbf{X_{ij}} + \gamma_A A_i + u_i \tag{4B}$$

where, again, the correct functional form should be used.

In the second stage, the intervention effect is estimated by adapting the unweighted GEE (equation (2)), to include the individual-level weights in a weighted GEE[12,41]

$$0 = \sum_{i=1}^{M} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{W}_i\big(\boldsymbol{Y}_i - \boldsymbol{\mu}_i\big) \tag{5}$$

where $\boldsymbol{W}_i$ a matrix with W_{ij} on the diagonal and zeroes on the off-diagonal. The weights could be those from a model such as (4A) or (4B). To distinguish between the two types of weighted GEE, we use W-GEE to refer to weights obtained from a model like (4A) with no clustering accounted for and use CW-GEE (clustered W-GEE) to refer to weights obtained from a model like (4B), which does account for clustering in generating the weights. Using weights in equation (5) aims to give more importance to individuals with low probability of having an observed outcome (i.e. for individuals with large weights). One of the major assumptions of IPW is that the probability of being observed is bounded away from zero for every individual (i.e. there are no infinite weights). Nevertheless, instability can be caused by large weights and more advanced stabilized weighting approaches can be used to address this issue.[42] The precision of the parameter estimates of the GEE marginal mean model can again be obtained through the sandwich variance estimator (equation (3)),[36] with an adaptation to account for $\boldsymbol{W_i}$. It has been demonstrated that, for some simple longitudinal settings, MMI-GEE and W-GEE are equivalent and this equivalence, therefore, should apply for simple settings of CRT data, such as the CDM missing data mechanism assumed here.[43]

3 Analysis of motivating data set

In order to provide an understanding of different GEE approaches to analyzing binary outcomes in CRTs in the presence of missing outcomes with complete baseline covariates, we first analyzed the motivating HALI data set using the following five approaches: CRA-GEE, A-CRA-GEE, W-GEE, CW-GEE and MMI-GEE. We then performed an extensive simulation study to evaluate the performance of all methods in settings with known data generation mechanisms (see Section 4). Here we describe applications to the motivating data example.

3.1 Data analysis

In this example, we focus on a single binary outcome, "*high literacy*" at 9-month follow-up and do not consider outcome data at 24-month follow-up. That variable (spelling score >10) is a dichotomized version of the primary outcome variable, spelling score, which ranges from 0 to 20 and which was also measured at baseline. In order to be able to show results from all five analyses without having to deal with missing baseline covariates, we sub-select data of the 2465 (97% of 2539) children who have no missing baseline covariates. Baseline cluster size distributions and baseline characteristics of these 2465 children are summarized by arm (Table 1, A). We restrict our attention to five baseline child-level covariates expected to be predictive of nine-month literacy outcome

Table 1. Analysis of HALI motivating data set.

A: Baseline and outcome characteristics of motivating HALI data set for n = 2465 participants with complete baseline covariates

	Intervention (n = 1230)	Control (n = 1235)
Baseline cluster characteristics		
Number	51	50
Cluster size at baseline – mean (SD)	24.1 (3.3)	24.7 (1.9)
Baseline child-level characteristics – % (n)[a]		
Female	47.9% (589)	49.5% (611)
Age – mean (SD)	7.7 (1.7)	7.9 (1.7)
Household head education		
Did not complete primary education	29.1% (358)	34.4% (425)
Primary	55.6% (684)	52.7% (651)
Secondary	11.5% (141)	10.6% (131)
College/degree	3.8% (47)	2.3% (28)
Household socioeconomic status (SES)[b]		
Poorest	19.0% (234)	26.3% (325)
Poor	19.9% (245)	21.1% (261)
Median poor	21.1% (259)	17.7% (219)
Less poor	19.9% (245)	18.4% (227)
Least poor	20.1% (247)	16.4% (203)
Baseline literacy – spelling score (0–20) – mean (SD)	8.4 (4.6)	7.8 (4.3)
Outcome at nine-month follow-up		
High literacy (spelling score >10)	52.4% (644)	39.5% (488)
Missing outcome	12.4% (152)	11.6% (143)

B: Estimated intervention effects under five different GEE approaches (with robust SE and exchangeable working correlation matrix)

	OR	(95% CI)
Complete records analysis (CRA)-GEE	1.82	(1.26, 2.65)
Adjusted-complete records analysis (A-CRA)-GEE	1.93	(1.38, 2.71)
Weighted (W)-GEE	1.80	(1.23, 2.62)
Clustered weighted (CW)-GEE	1.83	(1.26, 2.65)
Multilevel multiple imputation (MMI)-GEE	1.81	(1.25, 2.63)

[a]Unless otherwise stated.
[b]Obtained by principal components analysis of baseline household assets.

levels: gender, age, household head education, household socioeconomic status and baseline literacy (i.e. spelling score on the scale 0–20).

All models used a logit link and were unadjusted (i.e. included only an intercept and treatment arm in order to obtain a marginal intervention effect like equation (1)) except for A-CRA-GEE which also included the five baseline covariates as additive terms in the logistic model and therefore estimated an effect conditional on the covariate values. All analyses were conducted in R (version 3.5.0). For each of the five GEE analysis methods, we fitted both an exchangeable and independent working correlation structure using the geeM package (version 0.10.1)[44] and reported robust standard errors (SEs). Inference for the intervention effect was based on standard Wald Z-based confidence interval for each modeling approach, except for MMI-GEE for which t-based confidence intervals were used with appropriate degrees of freedom (see Hossain et al.,[24] p. 7, which is due to Barnard and Rubin[45]).

For MMI-GEE, like Hosssin et al.,[24] we used the jomo package (version 2.6-7),[46] specifically the jomo1rancat form of the procedure (based on personal communication with the authors). We generated 15 imputed data sets where the random effects logistic regression imputation model included the five baseline covariates as additive terms. As noted in Hossain et al. (Section 5.2[24]), the jomo package uses the probit link to perform the individual-level imputation which provides similar results to those from a logit link when the probabilities of missing outcomes are not too extreme. We used 100 burn-in iterations and a thinning rate of 25 because trace-plots of the Monte Carlo Markov Chain generated by the jomo1rancat procedure showed that the chain quickly stabilized

(converged) and had minimal auto-correlation, showing good mixing (see Figure S1). Although these diagnostics indicated that the selected burn-in and thinning rate were acceptable for analysis of the HALI data set (as did those for all combinations of parameter values used in the simulation study described below), we note that a larger burn-in or thinning rate may be needed in other applications.

3.2 Results

Children in the data set averaged 7.8 years of age and mostly resided in households where the household head had primary education or less (Table 1, A). Intervention and control arms were broadly comparable in terms of baseline characteristics, with the control arm children on average from households with lower SES. Overall, 295 (12%) out of 2465 children were missing the *high literacy* outcome, with comparable proportions missing for intervention (12.4%) and control (11.6%). Using CRA-GEE (equation (2)) with exchangeable working correlation and robust SE, the intervention was estimated to be effective at improving literacy with an odds ratio for *high literacy* of 1.82 (95% CI: 1.26, 2.65) for intervention vs. control (Table 1, B). Results from W-GEE, CW-GEE and MMI-GEE all yielded comparable results to each other (Table 1, B) which is to be expected as all predictors of missing outcomes were included in all approaches using the same additive form, e.g. for W-GEE the estimated intervention effect was slightly reduced (1.80, 95% CI: 1.24, 2.62) compared to CRA-GEE (where we note that for both the W-GEE and CW-GEE approaches, no extreme weights were estimated with all being smaller than 1000). In contrast, the A-CRA-GEE provided a slightly higher estimated intervention effect of 1.94 (95% CI: 1.38, 2.71) because it no longer estimates the unadjusted marginal intervention effect. Overall, results estimated using an independent working correlation matrix were comparable to those with exchangeable (Table S1).

In summary, in the analysis of this single data set with a reasonably high degree of outcome clustering (with an ICC on the logistic scale of 0.25, which was estimated from a generalized linear mixed model,[47] in the same way as for the simulation study below), there are minimal differences between estimated intervention effects from the commonly used MMI-GEE analysis and the W-GEE approaches and, in this specific example under an assumption of a CDM missing data pattern, there was no evidence that a complete records analysis was biased. This can be partly explained by examining the relationship between the five child-level baseline covariates and the outcome as well as their relationship with the probability of missing outcomes, as well as the relationship of treatment arm to the probability of missingness. First, three were predictive of the outcome: age, baseline literacy score and household head education (Table S1). Of these, only the latter was predictive of missing outcomes. Based on data summaries for children with non-missing outcomes (i.e. with *high literacy* observed), the proportion attaining *high literacy* at nine months was higher in intervention than control (52.4% vs. 39.5%, Table 1, A). Importantly, when we tested for interactions between intervention arm and each of the baseline covariates in the model of predictors of missing outcome, none of those interactions were significant. Relatedly, whilst intervention arm was predictive of the outcome (at least for the complete-records data), it did not appear to be predictive of the probability of missing (Table S2). As a consequence, it is to be expected that any analysis that assumes a CDM missing data mechanism would not show much difference to the CRA-GEE analysis whereas we would have expected differences in analysis approaches had intervention arm also been predictive of the outcome. However, there are situations where a CRA-GEE will provide different results to alternative methods and it is important to understand when such differences may arise.

4 Simulation study

To better understand the results of the HALI data analysis, and, more importantly, to assess the comparative performance of the five approaches (namely CRA-GEE, A-CRA-GEE, MMI-GEE, W-GEE and CW-GEE) in general settings, we conducted a simulation study. As in the real data analysis, we again focus on CRTs with missing binary outcomes in the presence of complete baseline covariates. The assumed missing data mechanism was CDM conditional on intervention arm and a single individual-level baseline covariate under two types of clustering of missingness, namely no clustering of missingness and under a range of levels of clustering of missingness.

We used a $3 \times 4 \times 3(=36)$ factorial simulation study design in which the outcome ICC (3 levels), missingness ICC (4 levels) and number of clusters (3 levels) were varied. We broadly followed the simulation framework adopted by Hossain et al.[24] and fixed other parameters at comparable levels. We considered additive outcome and missing data models (on the logit scale) and assumed $M = 2k$ clusters were randomly and evenly assigned to two arms in a parallel-arm CRT design ($k = 10, 25, 50$). Specifically, the outcome model assumed a constant

additive intervention effect under a logistic model with a constant baseline covariate effect (i.e. no heterogeneity of intervention effect, equivalently no interaction between intervention and covariate) and the missing outcome model (defined by a POM) assumed an additive intervention effect under a logistic model with a constant baseline covariate effect (i.e. no interaction between intervention and covariate) so that the overall probability of missingness was approximately 30% (as such, the coefficients were varied slightly under different scenarios in order to achieve this level). The cluster size was sampled from a Poisson distribution with mean of 50. We selected a total of 1000 simulated data sets for each of the 36 scenarios so that acceptable coverage of the 95% confidence intervals would range from 93.6% to 96.4% based on the sampling distribution of a sample proportion centered at the target value of 95% (i.e. $0.95 \pm 1.96\sqrt{0.95 \times 0.05/1000}$).

4.1 Data generation

There were three stages in the data generation process for each of the 1000 data sets under a fixed set of parameters. First, a continuous baseline covariate value was simulated for each individual in the data set according to the following model

$$X_{ij} \sim N\left(0, \sigma_X^2\right),$$

where σ_X^2 was fixed at 0.2. In the second stage, we simulated an outcome probability, $\pi_{ij} = E(Y_{ij}|A_i, X_{ij}) = P(Y_{ij} = 1|A_i, X_{ij})$, for each individual in the data set according to the following random effects logistic regression model that is additive on the logistic scale

$$\text{logit}\left(P\left(Y_{ij} = 1|A_i, X_{ij}\right)\right) = \beta_0 + \beta_A A_i + \beta_X X_{ij} + \delta_i \tag{6}$$

where we fixed $\beta_0 = \beta_X = 1$ and $\beta_A = 1.36$ (corresponding to an intervention effect of a conditional odds ratio of 3.9) and where $\delta_i \sim N\left(0, \sigma_\delta^2\right)$. We varied σ_δ^2 in order to model different outcome ICC values ρ_O, defined as $\sigma_\delta^2/\left(\sigma_\delta^2 + \pi^2/3\right)$ where $\pi \approx 3.142$ is the exponential constant term.[47] Using the simulated $P\left(Y_{ij} = 1|A_i, X_{ij}\right)$ for each individual we then generated a Bernoulli random variable (Y_{ij}) in order to obtain the simulated outcome for each individual. We note that the treatment effect β_A carries a conditional interpretation. Because we are interested in marginal effects through analysis with a GEE outcome model, we obtain the ''true'' marginal treatment effect under a given set of fixed parameter values by fitting the (unadjusted) GEE outcome model separately with exchangeable and independent working correlation matrices to the 1000 full data sets (i.e. without missing outcomes). We then separately averaged across those 1000 estimated intervention effects for those under exchangeability and those under independence, and treat each as the truth. This process is similar to that adopted by Hossain et al.[24] except that those authors fitted the adjusted GEE with both the intervention arm and covariate effect in the model i.e. equation (1) with X_{ij} in the model. Because our target estimand is the marginal effect, we selected to generate the ''true'' value under the fully marginal model (equation (1)) and refer to this as θ_A^*. This process can be thought of as a strategy to integrate out the single covariate that was used to generate the outcomes in the covariate-adjusted conditional model (equation (6)).

In the third stage, we generated a missing outcome probability for each individual in each data set under an additive logistic model, both with and without clustering of missingness, from which we simulated the binary missingness indicator. Specifically, the two forms were

$$\text{logit}\left(P\left(R_{ij} = 0|A_i, X_{ij}\right)\right) = \alpha_0^M + \alpha_A^M A_i + \alpha_X^M X_{ij} \tag{7A}$$

$$\text{logit}\left(P\left(R_{ij} = 0|A_i, X_{ij}, u_i\right)\right) = \gamma_0^M + \gamma_A^M A_i + \gamma_X^M X_{ij} + u_i \tag{7B}$$

where $u_i \sim N\left(0, \sigma_u^2\right)$ and where we varied σ_u^2 to model different missingness ICCs with $\rho_M = \sigma_u^2/\left(\sigma_u^2 + \pi^2/3\right)$. (We note too that POM (7A) and (7B) are equivalent to propensity score models (4A) and (4B) that model $P\left(R_{ij} = 1\right)$ rather than $P\left(R_{ij} = 0\right)$). Specifically, in order to model a setting with no clustering and three cases of clustering of missingness, we fixed ρ_M at the following four values: 0, 0.1, 0.3 and 0.5 (corresponding to values of σ_u^2 of 0 and approximately 0.366, 1.410 and 3.291, respectively). We fixed $\gamma_A^M = \gamma_X^M = \alpha_A^M = \alpha_X^M = 1$ and varied α_0^M or γ_0^M so that the overall missingness proportion across both arms was approximately 30%, with values of approximately 40% and 20% in intervention and control arms, respectively. In summary, we used a $3 \times 4 \times 3$ (=36) factorial simulation study design that assumed the following parameters: outcome ICC values of 0.01, 0.05 and 0.2; missingness ICC values of 0, 0.1, 0.3, 0.5; and number of clusters per arm of 10, 25 and 50 (for a total of 20, 50 and 100 clusters in the trial).

4.2 Data analysis

We then analyzed each of the 1000 simulated data sets with missing outcomes using the following five approaches: CRA-GEE, A-CRA-GEE, MMI-GEE, W-GEE and CW-GEE. All procedures were the same as those used in the analysis of the motivating data example, unless otherwise stated. In brief, all models used a logit link and were unadjusted except for A-CRA-GEE which included the covariate X_{ij} as an additive term in the logistic model. We fitted both exchangeable and independence working correlation matrices. As for the HALI CRT data analysis (see Section 3.1), inference was based on standard Wald Z-based confidence intervals except for MMI-GEE for which t-based confidence intervals were used with appropriate degrees of freedom. Because of known finite sample bias of the robust SE obtained from GEE analysis of CRTs with a small number of clusters (e.g. fewer than 40 in total), we also considered three small-sample corrections that have previously shown reasonable performance in other simulation studies. These corrections are the Mancl and DeRouen (MD),[48] Kauermann and Carroll (KC)[49] and Fay and Graubard (FG)[50] corrections, details of which are provided in Section 3 of the Supplementary Material. Some GEE models did not converge and therefore we reported on the fraction of non-convergence for each simulation scenario.

4.3 Simulation summary statistics

For each of the 36 parameter scenarios considered, we calculated the ''true'' value of the marginal intervention effect, θ_A^*, as outlined above and reported the following five statistics across the 1000 replicates: (1) mean relative bias (i.e. the mean deviation from θ_A^* of the estimated marginal intervention effect relative to θ_A^*), (2) coverage of the nominal 95% Z-based or t-based (for MMI-GEE) confidence intervals (i.e. the fraction of replicates for which the estimated 95% CI contains θ_A^*), (3) the mean standard error (SE) of the point estimate of θ_A^*, (4) the Monte Carlo standard deviation (MCSD) of the intervention effect (i.e. the sample SD of the point estimates of θ_A^*), and, (5) the fraction of replicates for which the GEE algorithm did not converge. In order to indicate the extent of extreme weights in W-GEE and CW-GEE, which could lead to instability, we additionally obtained the mean of the fraction of weights above 1000, corresponding to the mean of the fraction of estimated propensity scores below 0.001.

4.4 Simulation results

The main statistics (mean relative bias, coverage and the deviation of mean SE from the MCSD) summarizing the performance of the five GEE approaches using an exchangeable working correlation matrix and robust SE are presented in Figures 1 to 3, respectively, where the low coverage of A-CRA-GEE is not shown (Figure 2) to better compare the performance of the other four methods (see Figure S2 for the version also containing A-CRA-GEE). The corresponding numerical results for all scenarios are reported in Table S3 assuming an exchangeable working correlation matrix. Coverage for all scenarios (i.e. both ''large'' and ''small'' sample settings) under no correction to the robust SE and under the three finite-sample corrections (KC, MD and FG) are provided in Table S4 with the corresponding mean SEs reported in Table S5. Corresponding results under an independence working correlation matrix are displayed in Tables S6 to S8.

4.4.1 Relative bias

W-GEE and MMI-GEE had similarly small relative bias (absolute value <1% in most cases), which was largely insensitive to the outcome ICC and missingness ICC, as well as to the number of clusters (Figure 1 and Table S3) under both under an exchangeable working correlation matrix and independent working correlation matrix (Table S6). In contrast, CW-GEE showed increasing bias as the missingness ICC, ρ_M, increased at each fixed level of outcome ICC, ρ_O, and fixed number of clusters, k attaining absolute values >1% for some settings with $\rho_M = 0.3$ and 0.5. The poorest performance was observed in the small sample $k = 10$ scenario, for which absolute relative bias was as high as 3–6% when the outcome ICC, ρ_O, was largest. Relative bias of A-CRA-GEE was approximately 3% for all scenarios. This non-trivial bias was expected because the target parameter of interest was the marginal, and not adjusted, intervention effect. In contrast, relative bias of CRA-GEE was in the opposite direction (i.e. negative) and appeared to depend only on the missingness ICC, ρ_M, and not on the two other parameters, namely ρ_O and k. Importantly, Hossain et al.[24] demonstrated that CRA-GEE can provide unbiased estimates of intervention effects under a CDM missing data mechanism when the ''true'' marginal effect is estimated from a GEE model that also adjusts for the covariate X_{ij}. In contrast to an exchangeable working correlation matrix (Table S3), slightly smaller bias was observed for W-GEE under independence (Table S6), particularly as the missingness ICC, ρ_M, increased. Similar, though less marked, improvements were observed for

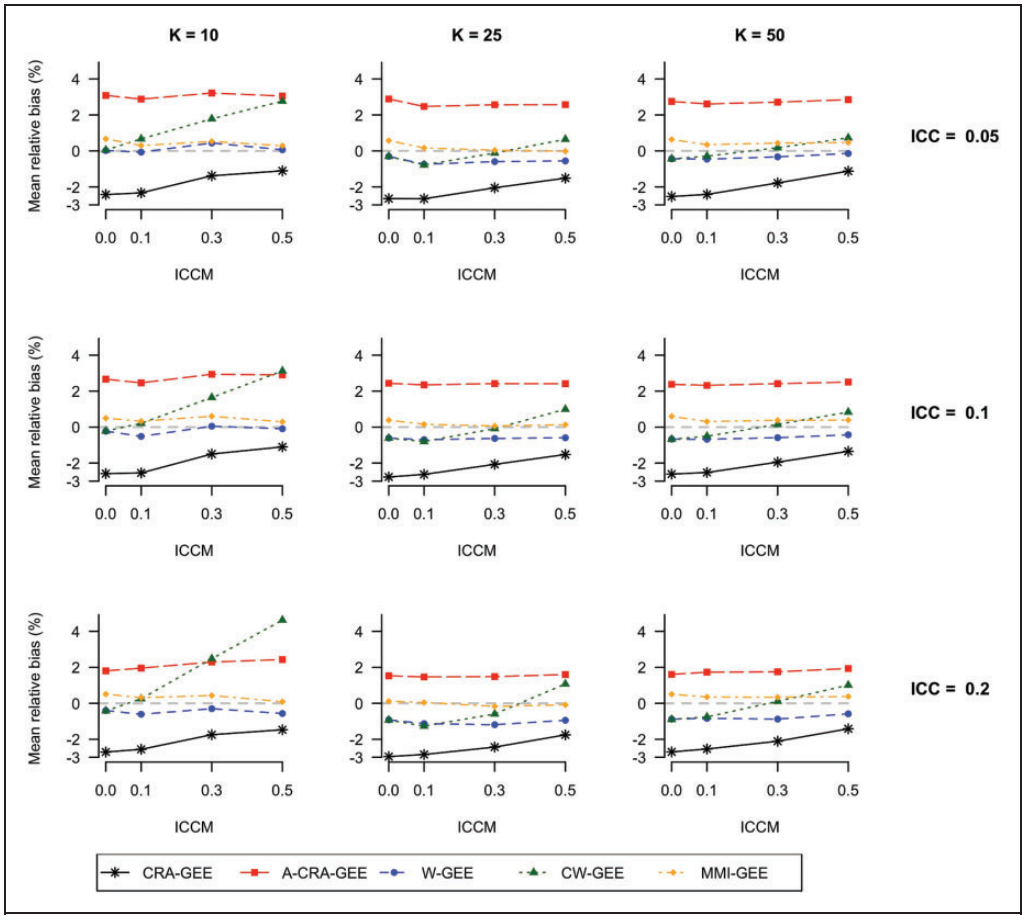

Figure 1. Mean relative bias (%) of five GEE methods to handle missing outcomes in CRTs. The results are based on 1000 simulated data sets per scenario (with 15 imputed data sets for MMI-GEE). ICC and ICCM, on the logistic scale, correspond to the logistic outcome model ICC (ρ_O) and the missing outcome ICCM (ρ_M), respectively (see equations (6) and (7A) and (7B), respectively). Robust standard errors under exchangeable working correlation matrix were used for all models. CRA-GEE: complete records GEE; A-CRA-GEE: adjusted CRA-GEE; W-GEE: weighted GEE (no adjustment for clustering when estimating the weights); CW-GEE: weighted GEE accounting for clustering when estimating the weights; MMI-GEE: multilevel multiple imputation GEE.

MMI-GEE and, to a lesser extent, for CRA-GEE. Large relative bias was observed for CW-GEE under large $\rho_M$whether an exchangeable (Figure 1, Table S3) or independent working correlation matrix (Table S6) was used.

4.4.2 *Coverage*

Performance of W-GEE, MMI-GEE and A-CRA-GEE was good for both $k = 25$ and $k = 50$ with values close to the nominal 95% level for all methods and for all scenarios with both exchangeable (Figure 2 and Table S3) and independent working correlation matrix (Table S6), with that under exchangeability slightly lower than under independence. However, as noted above, bias for CRA-GEE and A-CRA-GEE was non-trivial (see Tables S3 and Figure S1 for exchangeable working correlation matrix, and Table S4 under independence) and therefore those methods are not ones we would wish to use in practice. Under both exchangeable and independent working correlation matrices, coverage of CW-GEE decreased to levels outside of the range expected due to Monte Carlo error (i.e. <0.936 or >0.964) as ρ_M increased (e.g. $\rho_M > 0.5$). For the $k = 10$ scenario, reasonable coverage was only attained for some settings under the t-based inference of MMI-GEE (e.g., $\rho_O = 0.1, 0.2$ but not for $\rho_O = 0.05$). The low coverage under the $k = 10$ scenario for other methods using Z-based inference is expected and can be corrected using small-sample corrections to the robust SE. Specifically, each of the KC and MD corrections performed well, with the best coverage seen with the MD correction. In contrast, that from the FG approach led to over-coverage (see Tables S4 and S7 for exchangeable and independent working correlation matrix, respectively). The observations on improved coverage due to finite-sample bias-correction were generally consistent with prior simulation evidence in CRTs without missing outcome data.[51–53]

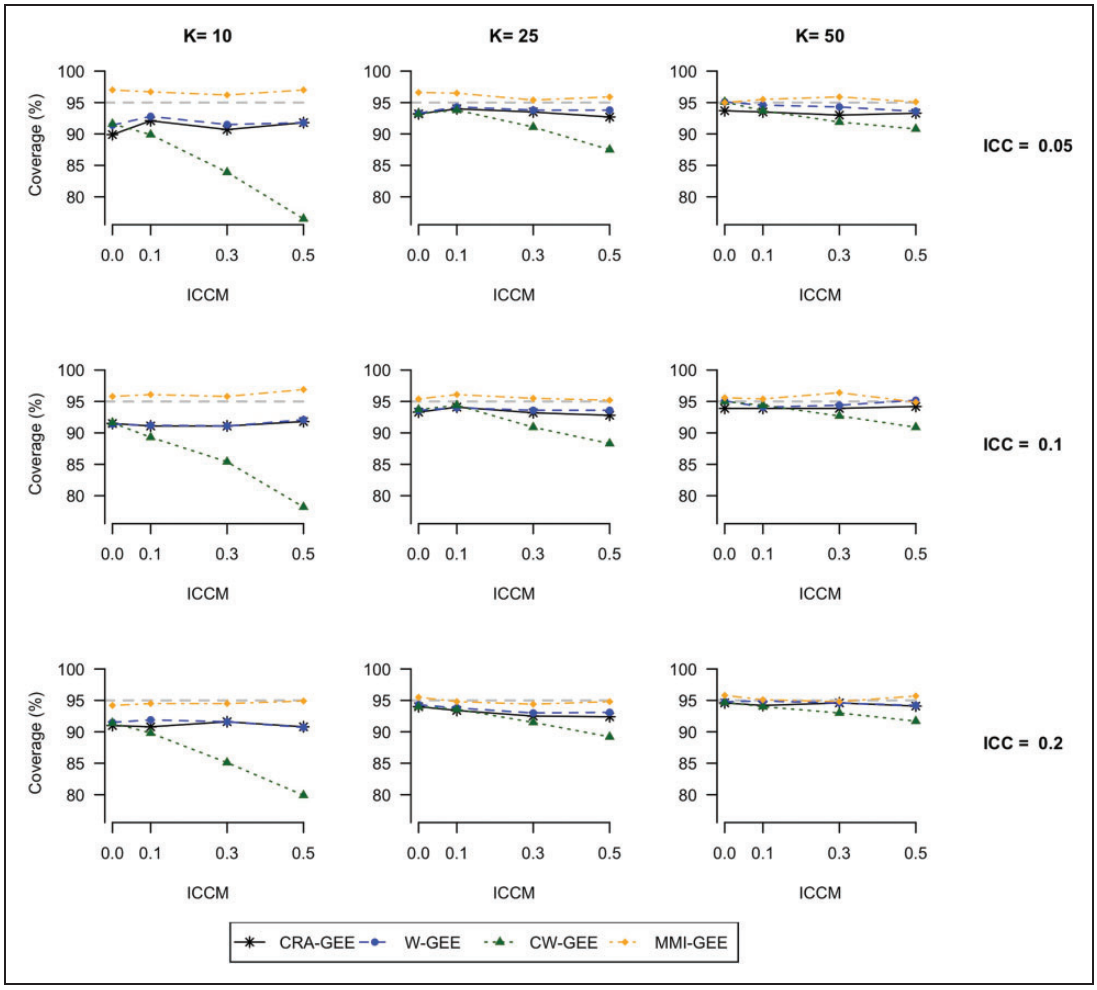

Figure 2. Coverage (%) of five GEE methods to handle missing outcomes in CRTs. The results are based on 1000 simulated data sets per scenario using standard Wald Z-based confidence intervals for each modeling approach, except for MMI-GEE for which t-based confidence intervals are used based on 15 imputations (see Section 4.2 for details, including the *t*-distribution degrees of freedom). Acceptable coverage ranges from 93.6% to 96.4%. ICC and ICCM, on the logistic scale, correspond to the logistic outcome model ICC (ρ_O) and the missing outcome ICCM (ρ_M), respectively – see equations (6) and (7A) and (7B), respectively. Due to low coverage, results from A-CRA-GEE are not presented here. Instead, refer to Figure S2. Robust standard errors under exchangeable working correlation matrix were used for all models. CRA-GEE: complete records GEE; A-CRA-GEE: adjusted CRA-GEE; W-GEE: weighted GEE (no adjustment for clustering when estimating the weights); CW-GEE: weighted GEE accounting for clustering when estimating the weights; MMI-GEE: multilevel multiple imputation GEE.

4.4.3 Mean SE and MCSD

The performance in terms of coverage can be explained by that of mean SE and bias. Comparable deviations of mean SE from the MCSD was observed for W-GEE, CRA-GEE and MMI-GEE for each fixed k and ρ_O, with no dependence on ρ_M (Figure 3 and Tables S3 and S6). This, combined with negligible bias for W-GEE and MMI-GEE led to good coverage of those two approaches. In contrast, although this led to reasonable coverage for CRA-GEE, such performance is offset by the observed bias of this approach (Figure 1). The lowest and largest mean SEs were obtained for A-CRA-GEE and CW-GEE, respectively, in all scenarios with mean SE for CW-GEE further deviating from the corresponding MCSD as $\rho_M$increases, so that the corresponding coverage of the CW-GEE procedure decreased. The low SE of A-CRA-GEE, when combined with negative bias in the point estimate, led to under-coverage of the nominal 95% CI at <90% in all cases, with lower coverage resulting from higher outcome ICC, ρ_O. MCSD was comparable to mean SE in most cases confirming that the simulation procedure provides us with good evidence as to the properties of the estimator of interest. An exception was for A-CRA-GEE, for which the MCSD was consistently larger than the mean SE, possibly because the coefficient of an additional term (i.e. covariate X) is estimated in this model.

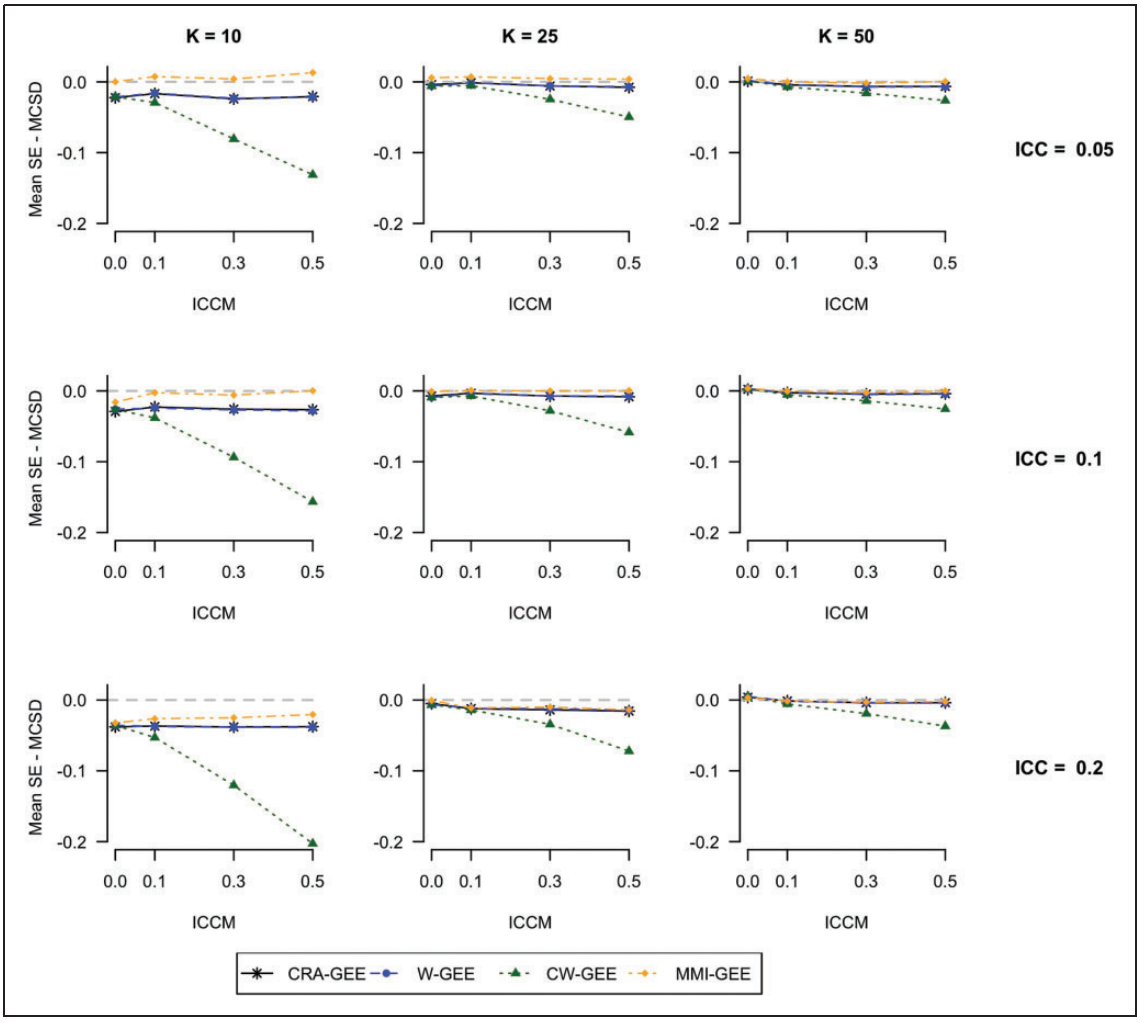

Figure 3. Deviations of mean SE from MCSD (%) of five GEE methods to handle missing outcomes in CRTs. The results are based on 1000 simulated data sets per scenario (with 15 imputed data sets for MMI-GEE) using robust SE with no small-sample correction (see Table S4 for a range of results from small-sample corrections, which is particularly relevant to the setting k=10). ICC and ICCM, on the logistic scale, correspond to the logistic outcome model ICC (ρ_O) and the missing outcome ICCM (ρ_M), respectively (see equations (6) and (7A) and (7B), respectively). Due to large deviations results from A-CRA-GEE are not presented here. Instead refer to Table S3. Robust standard errors under exchangeable working correlation matrix were used for all models. CRA-GEE: complete records GEE; A-CRA-GEE: adjusted CRA-GEE; W-GEE: weighted GEE (no adjustment for clustering when estimating the weights); CW-GEE: weighted GEE accounting for clustering when estimating the weights; MMI-GEE: multilevel multiple imputation GEE.

4.4.4 *W-GEE compared to CW-GEE*

W-GEE outperformed CW-GEE under CDM missing outcome mechanisms that involved clustering (i.e. for $\rho_M > 0$) in terms of both bias (it was smaller) and better coverage (mean SE was smaller). This was observed for both exchangeable (Figures 1 and 2, Table S3) and independent (Table S6) working correlation matrix. The estimated average proportion of the 1000 replicates with extreme weights (i.e. exceeding 1000) under both approaches (results not shown) was zero for $k = 25$ and $k = 50$ and less than 0.1% for the small sample case ($k = 10$).

4.4.5 *Summary*

As shown by our simulation study, W-GEE performs better than CW-GEE, even in the presence of clustering of missingness. Moreover, like t-based inference for MMI-GEE, Z-based inference of W-GEE (with finite-sample bias correction to the robust SE for $k = 10$) attains good coverage and therefore presents a good alternative approach to MMI-GEE to handle missing outcomes in CRTs.

5 Discussion

In this article, we have examined the performance of two weighted-GEE approaches (W-GEE and CW-GEE) to account for informative missing outcome data in cohort CRTs with a single follow-up time point. Importantly, we assumed that the missing outcome data process was dependent on fully observed baseline covariates, was across a range of degrees of clustering of the missing outcomes (from no clustering to extreme clustering of missingness), and that the clustering in the missingness model was independent of the clustering in the outcome model. We demonstrated that clustering should not be accounted for when estimating the weights in this setting, even when there is clustering in the missingness mechanism. This finding corroborates that of Skinner and D'Arrigo,[31] who considered weighted analyses of survey data for settings with clustering in the missingness process. (See Section 5.1 of the Supplementary Material for a brief summary of other references that consider weighted analyses of clustered outcome data in different contexts.) In their first set of simulations, Skinner and D'Arrigo[31] assumed that the random effects that drive the missingness mechanism (e.g., see our Model (7B)) are independent of those that drive the outcome mechanism (e.g., see our Model (6)), and found that the weights estimated without clustering provided less biased point estimates. In this scenario, where the random intercepts in the missingness model do not "confound" the treatment-outcome relationship, there may be "over-adjustment" for the cluster-level effect. Such over-adjustment may distort covariate balance between the weighted complete data (i.e. that with non-missing outcome data) and the unweighted full data, leading to bias in the intervention effect. Indeed, we observed similar findings as those of Skinner and D'Arrigo[31] in the survey setting with clustered outcomes (Section 5.2 of the Supplementary Material). Specifically, we observed that the weights estimated with clustering (equation (7B)) often led to worse covariate balance than the weights estimated without clustering (equation (7A)) in a CRT setting with clustered missing outcomes.

In other findings, our simulations also showed that the performance of W-GEE is comparable to the more commonly used MMI-GEE approach. Given that W-GEE can be computationally faster than MMI-GEE, W-GEE offers a good alternative to MMI-GEE. On the other hand, both CRA-GEE and A-CRA-GEE provided biased estimates for the marginal treatment effect estimate, whereas results due to Hossain et al.,[24] showed that A-CRA-GEE provides unbiased estimates for a marginal effect that was adjusted for the baseline covariate of interest. More specifically, the covariate-adjusted GEE estimates a conditional odds ratio that is different from the marginal odds ratio of interest, and hence is biased for the marginal odds ratio. Although a marginal odds ratio from the covariate-adjusted GEE model could be obtained by standardisation,[54] accounting for clustering when estimating its variance and confidence intervals involves more complicated approaches than those implemented in standard software, and therefore we have not examined its performance in our simulations. In contrast, the weighted GEE approach adjusted for covariates through the propensity score model naturally preserves the marginal estimand. In fact, propensity score weighting has also been recommended as a valid covariate-adjustment strategy in individually randomized trials in order to increase precision, whilst preserving the marginal estimand as the target of inference.[55]

Although extensive in nature, our simulation study has limitations. Three limitations of note include: (1) the nature of inference (t-based vs. Z-based), (2) the range of outcome ICC (ρ_O) values considered, and, (3) the assumed missing data mechanism. First, regarding statistical inference, that for MMI-GEE was based on the t-distribution using theory developed in the missing data literature, whereas inference for W-GEE was Z-based. As such, in the small-sample case (i.e. for $k = 10$), although MMI-GEE mostly had reasonable coverage using the robust SE, Z-based inference for W-GEE was only acceptable when small sample corrections were applied to the robust SE. Second, regarding the assumed outcome ICC values (ρ_O) of 0.05, 0.1 and 0.2, some CRTs have smaller ICC values and therefore we explored this setting through additional simulations at an ICC of 0.01 with $k = 25$. These results indicated over-coverage of MMI-GEE (i.e. exceeding 96.4%) for all values of the missingness clustering (ρ_M). In contrast, performance of W-GEE was good in terms of bias and coverage except for the case of no clustering in the missingness (ρ_M=0) for which there was undercoverage (Section 5.3 of the Supplementary Material). Third, related to the missing data mechanism, we assumed differential missingness between arms that arose due to additive effects of treatment arm and the baseline binary covariate on the logistic scale. In practice, such differential missingness may arise through a more complex model that also includes an interaction between intervention arm and the covariate. Such a mechanism has been explored by Hossain et al.[24] with analysis by MMI-GEE under a correctly specified imputation model, for which treatment effects were unbiased and confidence intervals attained nominal coverage (e.g. see scenarios S2 and S4 of Table 1 of Hossain et al.[24]). We plan future work to explore each of the three features identified here in relation to W-GEE, namely t-based inference, a broader range of clustering in missingness and outcome models, and more complex covariate dependent missingness mechanisms.

Importantly, our results confirm what has been shown elsewhere in the literature that, in order to provide unbiased estimates of the intervention effect using W-GEE and MI-GEE, it is assumed that the PS model (equation (3)) and imputation model are correctly specified, respectively, which is unlikely to be true in practice. For W-GEE, there are extensions to doubly robust approaches which can provide unbiased estimates even if the PS model is not correctly specified.[56] While such methods imply that the performance of W-GEE can be superior to MI-GEE, there is limited literature available that makes such methods available to the practitioner. Currently, an implementation of doubly robust W-GEE for missing outcomes in CRT is only available in R.[41,57] An additional benefit of that implementation is that it can increase the efficiency of W-GEE through an additional outcome regression model.[58] Doubly robustness could also be achieved for MI-GEE but theoretical and practical work is still to come.[59]

Although our results have shown that W-GEE can provide comparable results to MMI-GEE to address missing outcomes in CRTs, there are five potential limitations or pitfalls of the W-GEE approach. First, the current implementation does not account for uncertainty in the estimated weights. Second, depending on the nature of the missing data process, the weights may be extreme in magnitude. Particularly, when weights are very large (i.e. the probability of being observed is very small), this can lead to instability and strategies such as trimming or truncation may need to be used.[12,60] Third, in the small sample case (k <20), there may be bias in estimating the weights for CW-GEE because they are estimated using a random effects logistic regression model which may have poor performance in such small sample settings. Fourth, in contrast to MMI-GEE, weighted GEE does not naturally extend to account for multivariate missingness such as a situation where there are missing baseline covariates as well as missing outcomes. Fifth, when in the context of more general correlated data settings such as CRTs with longitudinal outcomes rather than the single follow-up time point considered in the current paper, weighted GEE does not easily extend to non-monotone patterns (e.g. those in longitudinal data where missingness is intermittent and participants may have outcomes measured at later times than ones that are missing). Recent work by Sun and Tchetgen Tchetgen has addressed such settings.[61]

In summary, in this article we have shown that using W-GEE to account for missing outcomes in CRT data is relatively easy to implement and that it performs similarly to MI-GEE. Although we explored only a single outcome, theory suggests that W-GEE should be preferred over MI-GEE when multiple outcomes are jointly missing.[27] This could arise when there are multiple follow-up time points with loss to follow up or where a multivariate outcome is of interest at a single time point. In this case, W-GEE may be preferred as it is often easier to specify a PS model for the probability of missing data than to model the joint distribution of all of the outcomes, which would be required in order to implement MI-GEE. It is also possible to combine MI-GEE and W-GEE to achieve additional efficiency in this setting.[14] Overall, W-GEE shows great promise as a viable alternative to MI-GEE to account for missing outcome data in CRTs.

Authors' note

Fan Li is now affiliated with the Department of Biostatistics, Yale University, New Haven, CT, USA.

Acknowledgements

We thank the principal investigators (Simon Brooker and Matthew Jukes) of HALI study for allowing the data to be used as a motivating example and thank all study participants for their involvement in the study.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: We acknowledge funding from two grants from the National Institutes of Health (R37 AI 51164 and R01 HD075875).

ORCID iDs

Elizabeth L Turner https://orcid.org/0000-0002-7638-5942
Fan Li https://orcid.org/0000-0001-6183-1893

Supplemental material

Supplemental material for this article is available online.

References

1. Murray DM. *Design and analysis of group-randomized trials*. New York: Oxford University Press, 1998.
2. Hayes RJ and Moulton LH. *Cluster randomised trials*. Boca Raton: CRC Press, 2009.
3. Brooker S, Okello G, Njagi K, et al. Improving educational achievement and anaemia of school children: design of a cluster randomised trial of school-based malaria prevention and enhanced literacy instruction in Kenya. *Trials* 2010; **11**: 93.
4. Hudgens MG and Halloran ME. Toward causal inference with interference. *J Am Stat Assoc* 2008; **103**: 832–842.
5. Campbell MK, Piaggio G, Elbourne DR, et al. Consort 2010 statement: extension to cluster randomised trials. *BMJ* 2012; **345**: e5661.
6. Fiero MH, Huang S, Oren E, et al. Statistical analysis and handling of missing data in cluster randomized trials: a systematic review. *Trials* 2016; **17**: 72.
7. Turner EL, Prague MP, Gallis JA, et al. Review of recent methodological developments in group-randomized trials: part 2-analysis. *Am J Public Health* 2017; **107**: 1078–1086.
8. Preisser JS, Young ML, Zaccaro DJ, et al. An integrated population-averaged approach to the design, analysis and sample size determination of cluster-unit trials. *Stat Med* 2003; **22**: 1235–1254.
9. Turner EL, Li F, Gallis JA, et al. Review of recent methodological developments in group-randomized trials: part 1-design. *Am J Public Health* 2017; **107**: 907–915.
10. Diaz-Ordaz K, Kenward MG, Cohen A, et al. Are missing data adequately handled in cluster randomised trials? A systematic review and guidelines. *Clin Trials* 2014; **11**: 590–600.
11. Little RJA and Rubin DB. *Statistical analysis with missing data*. 2nd ed. Hoboken: Wiley, 2002.
12. Robins JM, Rotnitzky A and Zhao LP. Analysis of semiparametric regression-models for repeated outcomes in the presence of missing data. *J Am Stat Assoc* 1995; **90**: 106–121.
13. Paik MC. The generalized estimating equation approach when data are not missing completely at random. *J Am Stat Assoc* 1997; **92**: 1320–1329.
14. Schafer JL and Yucel RM. Computational strategies for multivariate linear mixed-effects models with missing values. *J Comput Graph Stat* 2002; **11**: 437–457.
15. Seaman SR, White IR, Copas AJ, et al. Combining multiple imputation and inverse-probability weighting. *Biometrics* 2012; **68**: 129–137.
16. Hunsberger S, Murray D, Davis CE, et al. Imputation strategies for missing data in a school-based multi-centre study: the Pathways study. *Stat Med* 2001; **20**: 305–316.
17. Taljaard M, Donner A and Klar N. Imputation strategies for missing continuous outcomes in cluster randomized trials. *Biometrical J* 2008; **50**: 329–345.
18. Andridge RR. Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biometrical J* 2011; **53**: 57–74.
19. Hossain A, Diaz-Ordaz K and Bartlett JW. Missing continuous outcomes under covariate dependent missingness in cluster randomised trials. *Stat Methods Med Res* 2016; **26**: 1543–1562.
20. Caille A, Leyrat C and Giraudeau B. A comparison of imputation strategies in cluster randomized trials with missing binary outcomes. *Stat Methods Med Res* 2016; **25**: 2650–2669.
21. Ma J, Akhtar-Danesh N, Dolovich L, et al. Imputation strategies for missing binary outcomes in cluster randomized trials. *BMC Med Res Methodol* 2011; **11**: 18.
22. Ma J, Raina P, Beyene J, et al. Comparison of population-averaged and cluster-specific models for the analysis of cluster randomized trials with missing binary outcomes: a simulation study. *BMC Med Res Methodol* 2013; **13**: 9.
23. Ma J, Raina P, Beyene J, et al. Comparing the performance of different multiple imputation strategies for missing binary outcomes in cluster randomized trials: a simulation study. *J Open Access Med Stat* 2012; **2**: 93–103.
24. Hossain A, DiazOrdaz K and Bartlett JW. Missing binary outcomes under covariate-dependent missingness in cluster randomised trials. *Stat Med* 2017; **36**: 3092–3109.
25. Molenberghs G and Kenward M. *Missing data in clinical studies*. Chichester: John Wiley & Sons, 2007.
26. Van Buuren S. *Flexible imputation of missing data*. Boca Raton: Chapman and Hall/CRC, 2018.
27. Vansteelandt S, Carpenter J and Kenward MG. Analysis of incomplete data using inverse probability weighting and doubly robust estimators. *Methodology* 2010; **6**: 37–48.
28. Salazar A, Ojeda B, Duenas M, et al. Simple generalized estimating equations (GEEs) and weighted generalized estimating equations (WGEEs) in longitudinal studies with dropouts: guidelines and implementation in R. *Stat Med* 2016; **35**: 3424–3448.
29. Preisser JS, Lohman KK and Rathouz PJ. Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. *Stat Med* 2002; **21**: 3035–3054.
30. Inan G and Yucel R. Joint GEEs for multivariate correlated data with incomplete binary outcomes. *J Appl Stat* 2017; **44**: 1920–1937.

31. Skinner CJ and D'Arrigo J. Inverse probability weighting for clustered nonresponse. *Biometrika* 2011; **98**: 953–966.
32. Jukes MCH, Turner EL, Dubeck MM, et al. Improving literacy instruction in Kenya through teacher professional development and text messages support: a cluster randomized trial. *J Res Educ Eff* 2016; **10**: 449–481.
33. Hanley JA, Negassa A, Edwardes MD, et al. Statistical analysis of correlated data using generalized estimating equations: an orientation. *Am J Epidemiol* 2003; **157**: 364–375.
34. Zeger SL and Liang KY. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 1986; **42**: 121–130.
35. Neuhaus JM, Kalbfleisch JD and Hauck WW. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *Int Stat Rev* 1991; **59**: 25–35.
36. Liang KY and Zeger SL. Longitudinal data-analysis using generalized linear-models. *Biometrika* 1986; **73**: 13–22.
37. Meng X-L. Multiple-imputation inferences with uncongenial sources of input. *Stat Sci* 1994; **9**: 538–558.
38. Rubin DB. *Multiple imputation for nonresponse in surveys*. Hoboken: Wiley-Interscience, 2004.
39. Molenberghs G, Fitzmaurice G, Kenward MG, et al. *Handbook of missing data methodology*. Hoboken: Chapman and Hall/CRC, 2014.
40. Belitser SV, Martens EP, Pestman WR, et al. Measuring balance and model selection in propensity score methods. *Pharmacoepidemiol Drug Saf* 2011; **20**: 1115–1129.
41. Prague M, Wang R, Stephens A, et al. Accounting for interactions and complex inter-subject dependency in estimating treatment effect in cluster-randomized trials with missing outcomes. *Biometrics* 2016; **72**: 1066–1077.
42. Wang CL and Paik MC. A weighting approach for GEE analysis with missing data. *Commun Stat Theory Methods* 2011; **40**: 2397–23411.
43. Seaman SR and White IR. Review of inverse probability weighting for dealing with missing data. *Stat Methods Med Res* 2013; **22**: 278–295.
44. McDaniel LS, Henderson NC and Rathouz PJ. Fast pure R implementation of GEE: application of the Matrix package. *R J* 2013; **5**: 181–187.
45. Barnard J and Rubin DB. Miscellanea. Small-sample degrees of freedom with multiple imputation. *Biometrika* 1999; **86**: 948–955.
46. Quartagno M and Carpenter J. jomo: a package for multilevel joint modelling multiple imputation. R package version 2.6-7, http://CRAN.R-project.org/package=jomo, 2017.
47. Eldridge SM, Ukoumunne OC and Carlin JB. The intra-cluster correlation coefficient in cluster randomized trials: a review of definitions. *Int Stat Rev* 2009; **77**: 378–394.
48. Mancl LA and DeRouen TA. A covariance estimator for GEE with improved small-sample properties. *Biometrics* 2001; **57**: 126–134.
49. Kauermann G and Carroll RJ. A note on the efficiency of sandwich covariance matrix estimation. *J Am Stat Assoc* 2001; **96**: 1387–1396.
50. Fay MP and Graubard BI. Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics* 2001; **57**: 1198–1206.
51. Li F, Forbes AB, Turner EL, et al. Power and sample size requirements for GEE analyses of cluster randomized crossover trials. *Stat Med* 2018; **38**: 636–649.
52. Li F, Turner EL, Heagerty PJ, et al. An evaluation of constrained randomization for the design and analysis of group-randomized trials with binary outcomes. *Stat Med* 2017; **36**: 3791–3806.
53. Li F, Turner EL and Preisser JS. Sample size determination for GEE analyses of stepped wedge cluster randomized trials. *Biometrics* 2018; **74**: 1450–1458.
54. Santos CAS, Fiaccone RL, Oliveira NF, et al. Estimating adjusted prevalence ratio in clustered cross-sectional epidemiological data. *BMC Med Res Methodol* 2008; **8**: 80.
55. Williamson EJ, Forbes A and White IR. Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Stat Med* 2014; **33**: 721–737.
56. Carpenter JR, Kenward MG and Vansteelandt S. A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *J Roy Stat Soc Sta* 2006; **169**: 571–584.
57. Prague M, Wang R and De Gruttola V. *CRTgeeDR: an R package for doubly robust generalized estimating equations estimations in cluster randomized trials with missing data*. Cambridge: Harvard University Working Paper, 2016.
58. Stephens AJ, Tchetgen EJT and De Gruttola V. Augmented generalized estimating equations for improving efficiency and validity of estimation in cluster randomized trials by leveraging cluster-level and individual-level covariates. *Stat Med* 2012; **31**: 915–930.
59. Birhanu T, Molenberghs G, Sotto C, et al. Doubly robust and multiple-imputation-based generalized estimating equations. *J Biopharm Stat* 2011; **21**: 202–225.
60. Li F, Thomas LE and Li F. Addressing extreme propensity scores via the overlap weights. *Am J Epidemiol* 2018; **188**: 250–27.
61. Sun B and Tchetgen Tchetgen EJ. On inverse probability weighting for nonmonotone missing at random data. *J Am Stat Assoc* 2018; **113**: 369–379.