Statistica Sinica **23** (2013), 791-808
doi:http://dx.doi.org/10.5705/ss.2011.252

AN EXAMINATION OF A METHOD FOR MARGINAL INFERENCE WHEN THE CLUSTER SIZE IS INFORMATIVE

Menelaos Pavlou, Shaun R. Seaman and Andrew J. Copas

University College London, MRC Biostatistics Unit, Cambridge and MRC Clinical Trials Unit, London

Abstract: Generalised Estimating Equations (GEE) are a popular method to fit marginal models to clustered data. When the total number of members in the cluster is informative, then inference may be for a typical member of a typical cluster or the population of all cluster members. Applying the GEE with independence working correlation provides inference for the population of all members, and with additional weighting by the inverse cluster size gives inference for the population of typical members. In earlier work an adaptation of GEE termed modified within-cluster resampling (MWCR) was proposed to give unbiased inference for the population of typical members with increased efficiency by recognising the correlation between measurements. We describe how bias can arise when MCWR is used, a potential that was not clear when the method was proposed. We present conditions on the data structure and on the choice of the working correlation that, if satisfied, allow consistent estimation from MWCR. We illustrate the method with an application to a dataset of AIDS-related condition events from the Delta trial of HIV therapy.

Key words and phrases: Bias, efficiency, generalised estimating equations, informative cluster size.

1. Introduction

Clustered data arise in many fields of research. The number of members in a cluster may vary. For a given outcome variable of interest and set of covariates, cluster size is informative if the relation between covariates and outcome is different in clusters of different size. Formally, if we denote the outcome for a cluster member by Y, the corresponding vector of covariates by $\boldsymbol{X}$ and the size of the cluster to which the member belongs by N, then cluster size is said to be informative if $E(Y|\boldsymbol{X}, N) \neq E(Y|\boldsymbol{X})$. For example, in studies of factors associated with periodontal disease (Williamson, Datta, and Satten (2003)), a cluster corresponds to a person's mouth, and members to the teeth. The disease status of the teeth may be associated with the number of teeth in the mouth, even conditional on covariates, because it is likely that genetic and environmental factors causing periodontal disease also lead to tooth loss.

The method of Generalised Estimating Equations (GEE - Liang and Zeger (1986)) is widely used for the marginal regression analysis of clustered data due to its robustness against the misspecification of correlation structure and its relative ease of use. Several authors have examined the performance of GEE, or proposed modifications of them, when the variation in cluster size has arisen because of missing data and the aim is to estimate parameters of the distribution of the complete data (e.g., Little and Rubin (2002); Preisser, Lohman, and Rathouz (2002); Robins, Rotnitzky, and Zhao (1995)). In the present article, we are concerned with the situation where the observed data are complete, and our interest is in parameters of the distribution of the observed data. For example, in a study of periodontal disease it is natural that our interest may be in inference about the teeth that remain and not in the disease status of teeth that have fallen out or been removed.

Williamson, Datta, and Satten (2003) suggest that there are two marginal analyses of interest: one for the population of all cluster members (population M), where larger clusters contribute more to inference than smaller ones; and one for a typical member of a typical cluster, where all clusters contribute equally. We view the latter as inference for the population of typical cluster members (population C), which is a subpopulation of population M, formed by selecting one member at random from each cluster. Therefore if $E(.)$ denotes expectation in population M (sometimes written $E^M(.)$) and $E^C(.)$ denotes expectation in population C, then $E(Y|\boldsymbol{X}) = E_{N|\boldsymbol{X}} E_{Y|\boldsymbol{X},N}(Y)$ and $E^C(Y|\boldsymbol{X}) = E_{N|\boldsymbol{X}}[1/N E_{Y|\boldsymbol{X},N}(Y)]/E_{N|\boldsymbol{X}}(1/N)$.

Williamson, Datta, and Satten (2003) provide a guide to the analyst as to which population should be selected for inference according to the objective of the analysis, and Huang and Leroux (2011) also consider further possible populations for inference. In an economic assessment of how many, and which, teeth among patients seen at a dental clinic require a costly intervention, the population of all members (teeth) might be preferred, as clustering by patient may not be of direct relevance. Conversely, in a study of patient factors linked to the disease status of teeth, the population of typical cluster members (typical teeth for patients) might be of more interest.

Inference for population M can be obtained by applying the standard GEE with independence working correlation. For population C two inference methods were initially proposed: the computationally-intensive within-cluster resampling method (WCR - Hoffman, Sen, and Weinberg (2001)) and the simpler inversely-weighted-by-cluster-size GEE with independence working correlation (WIEE - Williamson, Datta, and Satten (2003); Benhin, Rao, and Scott (2005)). Williamson, Datta, and Satten (2003) proved that the two methods are asymptotically equivalent and showed through simulations that WIEE may perform better than WCR in terms of bias when the number of clusters is small. To provide inference for population C, a potentially more efficient method (MWCR) was proposed by Chiang and Lee (2008) based on the WCR method. When the minimum cluster size, m, is greater than 1, the authors propose randomly sampling m members from each cluster and then applying the GEE with a realistic working correlation to each resampled dataset. As the intracluster correlation is accounted for, efficiency may be gained.

Accounting for informative cluster size by adjusting for cluster size in the marginal regression model would not be a meaningful or useful choice because, for the scenarios considered, we assume that the scientific interest lies in the marginal effect of $\boldsymbol{X}$ on the expected outcome and not its effect conditional on N. For example, in toxicology where foetuses are clustered within litters, we might want to investigate the effect of a toxin (exposure) on the weight of a typical foetus from a typical litter. Unless cluster size is a predictor of primary scientific interest, such as in volume-outcome studies (see, for example, French et al. (2012)), there are two reasons why we do not wish to formulate a regression model involving N. First, N might lie in the causal pathway between Y and $\boldsymbol{X}$. In the toxicology application, adjusting for the cluster size may cause misleading inferences for the effect of the exposure if unobserved factors that contribute to the foetal loss induced by the toxin are also associated with the foetal weight. Second, if the effect of $\boldsymbol{X}$ on Y is different in clusters of different sizes then the effect of $\boldsymbol{X}$ conditional on N is a quantity which might not be scientifically useful.

Previous authors focused primarily on simple cases of informative cluster size, in the sense that the covariates of interest were either cluster-constant or cluster-size balanced (covariates are cluster-size balanced if their distribution is independent of the cluster size). These authors have also focused on scenarios in which the expected value of the outcome depends on cluster size and covariates, but not on interactions between the two. In this article we consider more general scenarios where the covariates involved are cluster-varying and non-size balanced. We explain why MWCR may lead to biased inference in these cases, a fact that is not clear in the original presentation of the method (Chiang and Lee (2008)). Furthermore, bias in MWCR can arise from realistic choices of the working correlation.

In the next section we introduce the standard notation used in GEE, present the WIEE and MWCR methods, and explain why bias can occur with MWCR in some cases. In Section 3, we use simulation studies to assess the performance of MWCR in terms of bias and relative efficiency compared to WIEE. In Section 4, we give an application of MWCR to a dataset of AIDS-related condition events from the Delta clinical trial of HIV therapy. Finally we discuss the results and possible future extensions of the methodology.

2. Methods of Estimation

2.1. Standard GEE and notation

Suppose that K clusters are randomly sampled from a population of clusters and that the values of an outcome Y and a vector (of length q) of covariates $\boldsymbol{X}$ are recorded for each member of each of these clusters. Let N denote the number of members in a cluster. We use subscripts i and j for the cluster and the member, respectively, so N_i is the number of members in cluster i, and Y_{ij} and $\boldsymbol{X}_{ij}$ are the values of Y and $\boldsymbol{X}$ for member j of cluster i ($i = 1, \ldots, K$; $j = 1, \ldots, N_i$). Let $\boldsymbol{Y}_i^* = (Y_{i1}, \ldots, Y_{iN_i})^T$ and $\boldsymbol{X}_i^* = (\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{iN_i})$. Let $\mu(\boldsymbol{X}) = E(Y \mid \boldsymbol{X})$, $\mu_{ij} = \mu(\boldsymbol{X}_{ij})$, and $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{iN_i})^T$.

A marginal regression model $\mu(\boldsymbol{X}) = h^{-1}(\boldsymbol{X}^T\boldsymbol{\beta})$ is specified, where h is a known canonical link function and $\boldsymbol{\beta}$ is a $q-$dimensional vector of unknown parameters of interest. Let $v(\mu) = dh^{-1}(\theta)/d\theta$, evaluated at $\theta = h(\mu)$. The working variance assumption is that $\mathrm{Var}(Y \mid \boldsymbol{X}) = v(\mu)\phi$, where ϕ is a scale parameter. Examples: for a continuous outcome, $h^{-1}(\theta) = \theta$ and $v(\mu) = 1$; for a binary outcome, $h^{-1}(\theta) = e^\theta/(1 + e^\theta)$, $v(\mu) = \mu(1 - \mu)$ and $\phi = 1$; for a count outcome, $h^{-1}(\theta) = e^\theta$, $v(\mu) = \mu$ and $\phi = 1$. A working correlation structure is also chosen. In this paper, we assume that it is either independence, exchangeable, auto-regressive, or fixed (Liang and Zeger (1986)). Depending on this choice, the actual working correlation may involve unknown parameters $\boldsymbol{\rho}$ that need to be estimated. Let $\hat{\boldsymbol{\rho}}$ denote the estimate of $\boldsymbol{\rho}$. We assume that $\hat{\boldsymbol{\rho}}$ converges to a value $\boldsymbol{\rho}_0$ as $K \to \infty$. Let $\hat{\boldsymbol{R}}_i$ and $\boldsymbol{R}_i$ denote the working correlation matrix for cluster i evaluated at $\hat{\boldsymbol{\rho}}$ and at $\boldsymbol{\rho}_0$, respectively. Note that $\hat{\boldsymbol{R}}_i$ and $\boldsymbol{R}_i$ can depend on observed variables which may or may not be included in $\boldsymbol{X}^*$. For example, for an auto-regressive working correlation structure, they will depend on the 'times' of the members in the clusters, and time may or may not be included as a variable in the analysis model.

If cluster size is non-informative, $E(Y \mid \boldsymbol{X})$ and $E^C(Y \mid \boldsymbol{X})$ are the same. If, furthermore, the marginal model $\mu(\boldsymbol{X}) = h^{-1}(\boldsymbol{X}^T\boldsymbol{\beta})$ is correctly specified, then under regularity conditions the solution $\hat{\boldsymbol{\beta}}$ to the GEE

$$\sum_{i=1}^{K} \boldsymbol{U}(\boldsymbol{\beta}; \boldsymbol{Y}_i^*, \boldsymbol{X}_i^*) = \sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \hat{\boldsymbol{V}}_i^{-1}(\boldsymbol{Y}_i^* - \boldsymbol{\mu}_i) = \boldsymbol{0} \tag{2.1}$$

is a consistent asymptotically normally distributed estimator of $\boldsymbol{\beta}$, where $\hat{\boldsymbol{V}}_i = \boldsymbol{A}_i^{1/2}\hat{\boldsymbol{R}}_i\boldsymbol{A}_i^{1/2}\phi$ is the working covariance matrix for cluster i and $\boldsymbol{A}_i$ is the $N_i \times N_i$ diagonal matrix whose j^{th} diagonal element is $v(\mu_{ij})$. This is so even if the working variance and correlation assumptions are false.

The variance of $\hat{\boldsymbol{\beta}}$ is consistently estimated by the sandwich estimator

$$\left(\sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \hat{\boldsymbol{V}}_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^T}\right)^{-1} \left(\sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \hat{\boldsymbol{V}}_i^{-1} (\boldsymbol{Y}_i^* - \boldsymbol{\mu}_i)(\boldsymbol{Y}_i^* - \boldsymbol{\mu}_i)^T \hat{\boldsymbol{V}}_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^T}\right) \\ \times \left(\sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \hat{\boldsymbol{V}}_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^T}\right)^{-1}.$$

The terms $\boldsymbol{\mu}_i$, $\boldsymbol{\partial \mu}_i^T/\partial \boldsymbol{\beta}$ and $\hat{\boldsymbol{V}}_i$ are evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$.

Let r_{ilj} denote the $(l,j)^{th}$ element of $\boldsymbol{R}_i^{-1}$, and let $r_{i+j} = \sum_{l=1}^{N} r_{ilj}$ and $r_{i++} = \sum_{j=1}^{N} r_{i+j}$. Let $\hat{r}_{ilj}$, $\hat{r}_{i+j}$ and $\hat{r}_{i++}$ denote the analogous quantities for $\hat{\boldsymbol{R}}_i^{-1}$. In preparation for Section 2.4, it is useful to consider the special case in which the identity link function $h^{-1}(\theta) = \theta$ is used and there are no covariates. In this case, all entries of $\boldsymbol{X}^*$ are one and β is just the population mean of Y. Equation (2.1) then becomes

$$\sum_{i=1}^{K} \boldsymbol{U}(\beta; \boldsymbol{Y}_i^*) = \sum_{i=1}^{K} \sum_{j=1}^{N_i} \frac{1}{\phi} \hat{r}_{i+j} (Y_{ij} - \beta) = 0,$$

to which the solution is $\hat{\beta} = \sum_{i=1}^{K} \sum_{j=1}^{N_i} \hat{r}_{i+j} Y_{ij} / \sum_{i=1}^{K} \hat{r}_{i++}$. Thus, $\hat{\beta}$ is a weighted average of the Y_{ij}'s in which the total weight given to the measurements in cluster i is $\hat{r}_{i++}$.

2.2. Within-cluster resampling and weighted independence estimating equations

When cluster size is informative, (2.1) will not, in general, give consistent estimation for either population M or C. In Section 2.1, we took $\mu(\boldsymbol{X}) = E(Y \mid \boldsymbol{X})$ and now, analogously, take $\mu^C(\boldsymbol{X}) = E^C(Y \mid \boldsymbol{X})$ for the population of typical cluster members.

The within-cluster resampling method (WCR) was proposed by Hoffman, Sen, and Weinberg (2001) to give inference for population C. In WCR a new dataset is created from the original dataset by sampling, at random, one member from each of the K clusters. This is done repeatedly, say Q times, so that Q datasets are created, each containing K members. A generalised linear model is used to estimate $\boldsymbol{\beta}$ for each of these Q datasets (since the K observations are independent), and then these Q estimates are averaged. As each cluster contributes one member to each estimate regardless of its size, it is apparent that the parameter estimated is that for the population of typical cluster members. Provided the marginal regression model $\mu^C(\boldsymbol{X}) = h^{-1}(\boldsymbol{X}^T\boldsymbol{\beta})$ is correctly specified, the WCR estimator is a consistent estimator of $\boldsymbol{\beta}$.

The weighted independence estimating equations (WIEE) approach, proposed by Williamson, Datta, and Satten (2003), provides an estimator that is asymptotically equivalent to WCR (as $K, Q \to \infty$) but avoids the Monte Carlo element of WCR. The WIEE are

$$\sum_{i=1}^{K} \left(\frac{\partial \boldsymbol{\mu}_i^C}{\partial \boldsymbol{\beta}} \right)^T \frac{1}{N_i \phi} \boldsymbol{A}_i^{-1} (\boldsymbol{Y}_i^* - \boldsymbol{\mu}_i^C) = \mathbf{0}. \tag{2.2}$$

Note that $\boldsymbol{A}_i \phi$ is the working covariance matrix based on an independence working correlation and the inclusion of the term N^{-1} means that clusters are inversely weighted by their size. If the marginal model $\mu^C(\boldsymbol{X}) = h^{-1}(\boldsymbol{X}^T \boldsymbol{\beta})$ is correctly specified, the solution to (2.2) is a consistent estimator of $\boldsymbol{\beta}$.

If, instead, the marginal model $\mu(\boldsymbol{X}) = h^{-1}(\boldsymbol{X}^T \boldsymbol{\beta})$ is correctly specified, inference can be made for population M by deleting N_i^{-1} from (2.2) and replacing $\boldsymbol{\mu}_i^C$ by $\boldsymbol{\mu}_i$. Doing this gives the standard GEE, (2.1), with independence working correlation. Note that using a non-independence working correlation in (2.2) does not give unbiased inference for population M when cluster size is informative.

2.3. A more efficient method for the population of typical cluster members

The MWCR method, proposed by Chiang and Lee (2008), is a modification of the WCR method and, like it, provides inference for the population of typical cluster members (population C). It can be more efficient than WCR when m, the size of the smallest cluster that appears in the dataset, is greater than 1. As is evident from its asymptotic equivalence to WIEE, WCR effectively uses an independence working correlation. MWCR, on the other hand, allows a non-independence working correlation to be used. We now describe this method.

For any subcluster s composed of m elements from cluster i, let $\hat{\boldsymbol{V}}_{i(s)}$, $\boldsymbol{Y}_{i(s)}^*$ and $\boldsymbol{\mu}_{i(s)}^C$ denote, respectively, the submatrix of $\hat{\boldsymbol{V}}_i$ and the subvectors of $\boldsymbol{Y}_i^*$ and $\boldsymbol{\mu}_i^C$ corresponding to those m members. There are two versions of the MWCR method, the first of which is more intuitively understandable but also more computationally intensive. The first resembles WCR, and the second, WIEE.

In the first version of MWCR, Q datasets are created from the original dataset by each time sampling at random (and without replacement) m members from each of the K clusters. For each of these Q datasets, $\boldsymbol{\beta}$ is estimated using the standard GEE with $\hat{\boldsymbol{V}}_i$ and $\boldsymbol{Y}_i^*$ replaced by the appropriate submatrix/subvector $\hat{\boldsymbol{V}}_{i(s)}$ and $\boldsymbol{Y}_{i(s)}^*$, and $\boldsymbol{\mu}_i$ replaced by $\boldsymbol{\mu}_{i(s)}^C$. The resulting Q estimates of $\boldsymbol{\beta}$ are then averaged.

Since each cluster contributes m members to each estimate of $\boldsymbol{\beta}$ regardless of its size, the parameter estimated is that for the population of typical cluster members. Also, since intracluster correlation is accounted for, MWCR may give increased efficiency relative to WCR.

The second version of MWCR is asymptotically equivalent to the first version as $K, Q \to \infty$, but avoids the Monte Carlo element. In this second version, $\boldsymbol{\beta}$ is estimated as the solution to the weighted GEE

$$\sum_{i=1}^{K} \frac{1}{\Delta_i} \sum_{s=1}^{\Delta_i} \left(\frac{\partial \boldsymbol{\mu}_{i(s)}^C}{\partial \boldsymbol{\beta}} \right)^T \hat{\boldsymbol{V}}_{i(s)}^{-1} (\boldsymbol{Y}_{i(s)}^* - \boldsymbol{\mu}_{i(s)}^C) = \mathbf{0}, \tag{2.3}$$

where $\Delta_i = N_i!/[m!(N_i - m)!]$ denotes the number of subclusters of size m that can be formed from cluster i (no subcluster can contain the same member more than once) and these Δ_i subclusters are indexed as $s = 1, \ldots, \Delta_i$. The correlation parameters $\boldsymbol{\rho}$ are estimated using the method outlined in Williamson, Datta, and Satten (2003). The weighted GEE (2.3) can then be seen to be the sum of the contributions to standard GEE from each of the subclusters, with each subcluster inversely weighted by the number of subclusters that can be formed from its cluster. This weighting ensures that each of the K clusters contribute equally to the GEE, regardless of its size.

An easily computed variance estimator for $\hat{\boldsymbol{\beta}}$ is given by $\boldsymbol{H}^{-1}\boldsymbol{B}\boldsymbol{H}^{-1}$, where

$$\boldsymbol{H} = \sum_{i=1}^{K} \frac{1}{\Delta_i} \sum_{s=1}^{\Delta_i} \left(\frac{\partial \boldsymbol{\mu}_{i(s)}^C}{\partial \boldsymbol{\beta}} \right)^T \hat{\boldsymbol{V}}_{i(s)}^{-1} \frac{\partial \boldsymbol{\mu}_{i(s)}^C}{\partial \boldsymbol{\beta}^T} \quad \text{and}$$

$$\boldsymbol{B} = \sum_{i=1}^{K} \frac{1}{\Delta_i} \sum_{s=1}^{\Delta_i} \left(\frac{\partial \boldsymbol{\mu}_{i(s)}^C}{\partial \boldsymbol{\beta}} \right)^T \hat{\boldsymbol{V}}_{i(s)}^{-1} (\boldsymbol{Y}_{i(s)}^* - \boldsymbol{\mu}_{i(s)}^C)(\boldsymbol{Y}_{i(s)}^* - \boldsymbol{\mu}_{i(s)}^C)^T \hat{\boldsymbol{V}}_{i(s)}^{-1} \frac{\partial \boldsymbol{\mu}_{i(s)}^C}{\partial \boldsymbol{\beta}^T}$$

are evaluated at $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\rho}}$. This variance estimator was not clearly described by Chiang and Lee (2008).

Finally, we note that when the MWCR method is applied with independence working correlation matrices, it provides estimates with the same asymptotic distribution as the WCR and WIEE methods. This is shown by Chiang and Lee (2008).

2.4. Bias in MWCR for general covariate patterns

Chiang and Lee (2008) focus on scenarios in which the covariates $\boldsymbol{X}$ are either cluster-constant or cluster-size balanced. In these special cases MWCR give consistent estimation, but only with certain choices of working correlation. In general MWCR is biased, a fact that is not evident in their paper. Here we state conditions under which (2.3) are consistent estimating equations for $\boldsymbol{\beta}$.

We assume a correctly specified marginal model $\mu^C(\boldsymbol{X}) = h^{-1}(\boldsymbol{X}^T\boldsymbol{\beta})$ for population C. Consider the following sampling mechanism. A cluster is chosen at random from the population of clusters, a subcluster of size m is sampled at random (without replacement) from this cluster and, finally, a member is chosen at random from this subcluster. Let N denote the size of the chosen cluster and $\boldsymbol{X}^*$ denote the covariate values for all members of the cluster. Let Y and $\boldsymbol{X}$ denote the outcome and covariate values for the chosen member of the chosen subcluster, and $\tilde{\boldsymbol{X}}$ denote the covariate values for all the other $m-1$ members of the chosen subcluster. Let $\tilde{\boldsymbol{R}}$ denote the working correlation for the chosen subcluster when $\boldsymbol{\rho} = \boldsymbol{\rho}_0$.

Theorem. *The solution to the MWCR estimating equations* (2.3) *is a consistent estimator of $\boldsymbol{\beta}$ for population C if:*

1. either (a) $Y \perp\!\!\!\perp \tilde{\boldsymbol{X}} \mid \boldsymbol{X}$ or (b) the independence working correlation is used;
2. $N \perp\!\!\!\perp \tilde{\boldsymbol{R}} \mid \boldsymbol{X}, \tilde{\boldsymbol{X}}$;
3. $Y \perp\!\!\!\perp \tilde{\boldsymbol{R}} \mid N, \boldsymbol{X}, \tilde{\boldsymbol{X}}$.

A proof of the Theorem is provided in Appendix A.

Condition 1a is closely related to the assumption that $Y \perp\!\!\!\perp \boldsymbol{X}^* \mid \boldsymbol{X}$, a condition identified by Pepe and Anderson (1994) as necessary for consistent estimation when cluster size is constant and GEE are used with a non-independence working correlation. Note that when the cluster size is informative and $\boldsymbol{X}$ involves cluster-varying and non-size balanced covariates, this condition is violated.

Whether Conditions 2 and 3 are satisfied depends on the choice of working correlation structure. Condition 2 is not satisfied, for example, if an auto-regressive structure is used and the time intervals between members of larger clusters tend to be longer (or shorter) on average than the intervals between members of smaller clusters, even after taking into account the values of $\boldsymbol{X}$ in the subcluster. Condition 3 is the requirement that the working correlation for a randomly chosen subcluster be conditionally independent of the outcome, Y, of a randomly chosen member from that subcluster given the size of the cluster to which that member belongs and the covariate values of the members of the subcluster. It is not satisfied, for example, if an auto-regressive structure is used and a member's Y value tends to be higher (or lower) in subclusters with longer time intervals between members than in subclusters with shorter intervals, even after taking into account the $\boldsymbol{X}$ values in the subcluster and the size of the cluster from which the subcluster came. However, if an auto-regressive structure is used and time is one of the covariates $\boldsymbol{X}$ in the analysis model, then Conditions 2 and 3 are satisfied. The independence and exchangeable working correlation structures are guaranteed to satisfy Conditions 2 and 3.

The necessity of Conditions 2 and 3 can be appreciated by considering the special case introduced at the end of Section 2.1: $h^{-1}(\theta) = \theta$ and there are no covariates. In this case, it can be shown, analogously to the result at the end of Section 2.1 (see proof of the Theorem for full details), that the population mean β is estimated by a weighted average of the Y_{ij}'s in which the total weight given to cluster i is the average, over each of the Δ_i $(m \times m)$ submatrices of $\hat{\boldsymbol{R}}_i$, of the sum of the elements of its inverse matrix. Therefore, if Condition 2 is not satisfied, clusters of different sizes may, on average, be given different total weights. Likewise, if Condition 3 is not satisfied, two clusters of the same size but with different expectations for Y receive different total weights.

If Conditions 1–3 are not satisfied, MWCR may not give consistent estimation of $\boldsymbol{\beta}$. For this reason, one is restricted in the choice of possible working correlations, a restriction that limits the potential for improving efficiency by using MWCR rather than WIEE.

Note that Conditions 1a and 3 are required for unbiased estimation when using a non-independence matrix even if cluster size is non-informative. Condition 1a is more likely to be violated in datasets with informative cluster size than in datasets with constant or non-informative cluster size. This is because informative cluster size indicates that there is important heterogeneity between clusters. If underlying cluster characteristics affect both the cluster size and the distribution of cluster-varying covariates then the condition is violated. Conversely, informative cluster size does not convey any additional likelihood of Condition 3 being violated. For example, in longitudinal datasets where an auto-regressive working correlation is selected for the application of MWCR or standard GEE, then Condition 3 is violated if the timing of measurements is associated with the expected outcome, after adjusting for covariates, irrespective of the total number of measurements.

3. Simulation Study

In this section we assess the performance of MWCR in terms of bias and efficiency in scenarios where cluster size is informative. We simulated clustered normal responses Y and a binary cluster-varying covariate X. We induced informative cluster size through an underlying 'susceptibility' that did not vary within the cluster. Each simulated dataset contained 100 clusters. Data were generated independently for cluster i as follows.

1. Generate $B_i \sim N(0, 0.5^2)$ as the underlying susceptibility.
2. Generate $N_i \sim Poisson\{\exp(\alpha_0 + \alpha_1 B_i)\} + m$, where m is the minimum cluster size.
3. Generate $X_{ij} \sim Bernoulli\{\lambda_0 + \lambda_1 \text{logit}^{-1}(B_i)\}$ independently for $j = 1, \ldots, N_i$. Note that if $\lambda_1 = 0$ then X is size balanced, while if $0 < \lambda_1 \leq 1$ it is non-size balanced.

4. Calculate the linear predictor $\eta_{ij} = \gamma_0 + \gamma_1 X_{ij} + \gamma_2 B_i + \gamma_3 B_i X_{ij}$, and write $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{iN_i})^T$. Parameter γ_2 determines the association between the underlying susceptibility (and consequently cluster size) and the outcome, while γ_3 determines how this association changes with X.
5. Generate $\boldsymbol{Y}_i^* \sim MVN(\boldsymbol{\eta}_i, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is an exchangeable correlation matrix with parameter (pairwise correlation) ρ.

We selected $\gamma_0 = \gamma_1 = \gamma_2 = 1$ and either $\gamma_3 = 0$ or $\gamma_3 = 1$. We selected ρ=0.2, 0.5, or 0.8, which correspond to small, medium, or high correlation/autocorrelation. For each scenario we generated 2,000 simulated datasets. When $\alpha_1 \neq 0$ and either $\gamma_2 \neq 0$ or $\gamma_3 \neq 0$, the cluster size is informative. For size balanced X we selected $\lambda_0 = 0.4$ and $\lambda_1 = 0$ and for non-size balanced, $\lambda_0 = 0$ and $\lambda_1 = 1$. For population C the correctly specified analysis model was of the form $E^C(Y_{ij}) = \beta_0 + \beta_1 X_{ij}$. The true values of β_0 and β_1 were calculated using numerical integration.

We applied the WIEE and MWCR methods. For scenarios in which MWCR is unbiased we calculated its efficiency relative to WIEE. For each scenario we present the mean estimated values of the parameters over the 2,000 simulated datasets and their empirical standard errors (ese). We also present coverage probabilities. The working correlation selected was the same as that used to generate the data at Step 5 above, though we note this is not generally the correct correlation because, at Step 5, the term $\boldsymbol{\eta}_i$ gives $E(\boldsymbol{Y}_i^*|B_i, \boldsymbol{X}_i^*)$ while our regression models condition only on X.

The correlation structure at Step 5 was exchangeable and the minimum cluster size was $m = 2$. The parameters for the cluster size model were selected to be $\alpha_0 = \alpha_1 = 1$ and these resulted in a mean cluster size of approximately 4. As shown in Table 1, MWCR led to unbiased inference with increased efficiency relative to WIEE when X was size balanced, but resulted in bias when X was non-size balanced, due to violation of Condition 1a. For scenarios in which MWCR is unbiased, we would expect slightly greater efficiency gains had the working correlations been correctly specified. The variance estimator we have presented for the MWCR method is seen to have good coverage when the method is unbiased. Similar simulation studies were carried out for binary responses with a cluster-varying size balanced or non-size balanced covariate. The results were generally consistent with those (reported above) for the corresponding scenarios with Normal responses.

Additional simulation studies demonstrating bias from the violation of Conditions 2 and 3 are presented in Appendix B. These scenarios concern longitudinal data and the correlation between measurements depends on the times at which they are obtained. Bias is seen when we apply the MWCR approach using an

Table 1. Application of WIEE and MWCR for the population of typical members. The cluster size is informative and the working correlation is exchangeable.

No Interaction: $\gamma_3 = 0$

		X non-size balanced TRUE $(\beta_0, \beta_1) = (0.88, 1.23)$			X size balanced TRUE (β_0, β_1)=(1.00,1.00)			
ρ	Method	$\hat{\beta}_0$(ese($\hat{\beta}_0$))	$\hat{\beta}_1$(ese($\hat{\beta}_1$))	CVR[a]($\hat{\beta}_0$,$\hat{\beta}_1$)	$\hat{\beta}_0$(ese($\hat{\beta}_0$))	$\hat{\beta}_1$(ese($\hat{\beta}_1$))	CVR[a]($\hat{\beta}_0$,$\hat{\beta}_1$)	RE[b]($\hat{\beta}_0$,$\hat{\beta}_1$)
	WIEE	0.88(0.099)	1.23(0.115)	(0.95,0.95)	1.00(0.093)	1.00(0.112)	(0.95,0.94)	-
0.2	MWCR	0.92(0.096)	1.15(0.104)	-	1.00(0.090)	1.00(0.101)	(0.95,0.95)	(1.05,1.22)
	WIEE	0.88(0.111)	1.23(0.118)	(0.95,0.95)	1.00(0.105)	1.00(0.113)	(0.96,0.95)	-
0.5	MWCR	0.95(0.103)	1.08(0.087)	-	1.00(0.100)	1.00(0.084)	(0.95,0.95)	(1.10,1.80)
	WIEE	0.88(0.122)	1.23(0.122)	(0.95,0.94)	1.00(0.115)	1.00(0.113)	(0.95,0.95)	-
0.8	MWCR	0.97(0.110)	1.04(0.058)	-	1.00(0.107)	1.00(0.054)	(0.95,0.95)	(1.15,4.30)

With Interaction: $\gamma_3 = 1$

		X non-size balanced TRUE (β_0, β_1)=(0.88,1.35)			X size-balanced TRUE (β_0, β_1)=(1.00,1.00)			
ρ	Method	$\hat{\beta}_0$(ese($\hat{\beta}_0$))	$\hat{\beta}_1$(ese($\hat{\beta}_1$))	CVR[a]($\hat{\beta}_0$,$\hat{\beta}_1$)	$\hat{\beta}_0$(ese($\hat{\beta}_0$))	$\hat{\beta}_1$(ese($\hat{\beta}_1$))	CVR[a]($\hat{\beta}_0$,$\hat{\beta}_1$)	RE[b]($\hat{\beta}_0$,$\hat{\beta}_1$)
	WIEE	0.88(0.099)	1.35(0.139)	(0.95,0.96)	1.00(0.093)	1.01(0.138)	(0.95,0.94)	-
0.2	MWCR	1.95(0.096)	1.21(0.121)	-	1.00(0.089)	1.00(0.120)	(0.95,0.95)	(1.06,1.32)
	WIEE	0.88(0.111)	1.35(0.142)	(0.95,0.95)	1.00(0.105)	1.01(0.138)	(0.94,0.95)	-
0.5	MWCR	0.98(0.104)	1.14(0.105)	-	1.00(0.099)	1.00(0.103)	(0.95,0.95)	(1.10,1.77)
	WIEE	0.88(0.122)	1.35(0.144)	(0.95,0.95)	1.00(0.113)	1.01(0.140)	(0.95,0.94)	-
0.8	MWCR	1.01(0.110)	1.08(0.083)	-	1.00(0.108)	1.00(0.080)	(0.95,0.94)	(1.14,2.95)

a. Empirical Coverage Probability.

b. Relative efficiency compared to WIEE.

auto-regressive working correlation. For Condition 2, the selected scenario is one in which the size of the cluster is associated with the times of observation of the measurements. For Condition 3, in order to isolate the problem, we selected a scenario with constant cluster size but where the times of observation of the measurements differ between clusters. The condition is violated if, after adjusting for covariates, the outcome tends to be higher when members are more widely spaced in time.

4. Illustration

To illustrate the methodology we analyse data from the Delta trial (Delta Coordinating Committee (1996)), a three arm international randomised controlled trial designed to test whether combinations of zidovudine (AZT) with zalcitabine (ddC) or with didanosine (ddI) are more effective than AZT alone in extending survival and delaying disease progression for HIV infected patients.

AIDS-Related Condition (ARC) events experienced during follow-up were recorded. We investigate how the immune status of a patient (of which CD4 count is an indicator) at the times of ARC events is related to the randomisation arm and time since entry to the trial. Since CD4 count had a skewed distribution we model its square root, Y. Events are clustered by patient and we let N denote the number of events experienced by a patient. Let subscript i denote patient (cluster), and let j index the N_i events experienced by patient i. Let X_1 and X_2 be indicator variables of randomisation to the drug combinations AZT+ddC and AZT+ddI, respectively (cluster-constant), and let T be the time to the event from entry to the study in units of 60 days. As interactions between T and X_1 and X_2 were found non-significant, the model we considered was

$$E^C(Y_{ij}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 T_{ij} + \beta_4 T_{ij}^2. \tag{4.1}$$

MWCR can only be applied to datasets where the minimum cluster size is 2 or more. While a thorough examination of immune status at the time of events would clearly involve all events, for the purposes of comparison between methods in our illustration we excluded all patients with one episode. After excluding clusters of size one, 657 clusters remained; the maximum cluster size was 15, the median 3. Among the 657 clusters 32% were of size 2 (cluster size group 1), 39% of size 3 or 4 (group 2) and 29% of size 5 to 15 (group 3). The mean square root CD4 count was 9.25, 8.01, and 6.2, for groups 1, 2, and 3, respectively. This is an initial indication that the cluster size might be informative; patients with more episodes tend to have lower CD4 count than patients with fewer episodes.

We fit a regression model analogous to (4.1), but for population M and including cluster size alongside the covariates of main interest: $E(Y_{ij}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_4 T_{ij}^2 + \beta_N N_i$. We used independence estimating equations to fit this model. The effect of cluster size was found significant ($\hat{\beta}_N = -0.47$, $\mathrm{se}(\hat{\beta}_N) = 0.09$, $p < 0.001$), supporting the initial indication for informative cluster size. We tested for interactions between the cluster size and covariates and these were not found statistically significant.

Model (4.1) was fitted using WIEE and MWCR. WIEE is known to provide consistent inference. So, although the true value of parameters is unknown, parameter estimates and standard errors from MWCR are compared to the corresponding ones from WIEE to assess the evidence of bias in MWCR and possible efficiency gains. MWCR was applied using either (a) an exchangeable working correlation (MWCR(EX)) or (b) an auto-regressive type working correlation with lag 1. The choice (b) is the AR(1) correlation corresponding to treating consecutive events as occurring one time unit apart, e.g. at times 1, 2, and 3 if cluster size is 3. We considered the application of MWCR with the more conventional auto-regressive correlation structure based on the actual times of episodes, but this was not possible because of computational problems. Specifically, because of the highly irregular times of the episodes, for many clusters the working correlation matrix was non-invertible.

Table 2. Application of WIEE and MWCR for population C using data from the Delta trial.

Model: $E^C(Y_{ij}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 T_{ij} + \beta_4 T_{ij}^2$					
Method	$\hat{\boldsymbol{\beta}}_0(\text{se}(\hat{\boldsymbol{\beta}}_0))$	$\hat{\boldsymbol{\beta}}_1(\text{se}(\hat{\boldsymbol{\beta}}_1))$	$\hat{\boldsymbol{\beta}}_2(\text{se}(\hat{\boldsymbol{\beta}}_2))$	$\hat{\boldsymbol{\beta}}_3(\text{se}(\hat{\boldsymbol{\beta}}_3))$	$\hat{\boldsymbol{\beta}}_4(\text{se}(\hat{\boldsymbol{\beta}}_4))$
WIEE	10.96(0.472)	0.71(0.412)	0.97(0.421)	-0.73(0.095)	0.027(0.0053)
MWCR(EX)	11.16(0.353)	0.73(0.414)	0.98(0.426)	-0.63(0.065)	0.016(0.0039)
MWCR(AR-1)	10.95(0.369)	0.70(0.410)	0.96(0.423)	-0.62(0.071)	0.017(0.0042)

The results from the application of the methods are presented in Table 2. Interestingly, CD4 count at ARC events is on average higher for a typical patient who receives the combination treatment AZT+ddI, compared to a typical patient receiving AZT alone. Also, as would be expected, CD4 count at ARC events for a typical patient decreases over time.

In terms of the performance of MWCR(EX) and MWCR(AR-1), there is some evidence of bias in the estimation of the effects of T and T^2. In particular, for the effect of T^2, the difference between the estimates from MWCR and WIEE is approximately three times the standard error of the estimates when using MWCR(EX) or MWCR(AR-1). The differences between the estimates from WIEE and MWCR of the effects of X_1 and X_2 are negligible. For the intercept term the difference is small when using MWCR(EX) and negligible for MWCR(AR-1). For the intercept term and the effect of T and T^2, the standard errors of the estimates are considerably smaller for MWCR(EX) and MWCR(AR-1) compared to WIEE.

In our illustration, Condition 2 and 3 of our Theorem are satisfied when using MWCR(EX). Condition 2 is not satisfied for MWCR(AR-1) because the correlations specified between members of subclusters are typically smaller for subclusters of larger clusters than for subclusters of smaller clusters. Condition 3 may not be satisfied for MWCR(AR-1) if the gaps between ARC events are associated with the CD4 at events. Condition 1 is not met for either MWCR(EX) or MWCR(AR-1) because the covariates T and T^2 are not size balanced. The mean time in days from entry to the trial for events in cluster size groups 1, 2, and 3 (see earlier) was 495, 465, and 502 respectively, indicating some deviation from size balance. This we view as the main reason for the probable bias in the application of MWCR seen in our results, as Conditions 2 and 3 were satisfied for MWCR(EX) but the bias seems as large for MWCR(EX) as MWCR(AR-1).

5. Discussion

In this paper we have drawn attention to bias in the MWCR method in scenarios where the covariates are non-size balanced, a bias that was not mentioned by Chiang and Lee (2008). Importantly, even for size-balanced covariates we clarify that an exchangeable working correlation is the only safe choice when using MWCR, and we present a variance estimator.

Three conditions have been identified for the consistency of MWCR. Condition 1a relates to the structure of the covariates. If the cluster size is informative and the cluster varying-covariate $\boldsymbol{X}$ is non-size-balanced, Condition 1a is violated. For this reason it is important to consider in advance the likelihood of non-size balanced covariates given the study design and scientific setting. Where size-balance of covariates is assured, MWCR may be a good choice of analysis method, and where non-size-balanced covariates are likely, WIEE will be a natural choice. In scenarios where non-size-balanced covariates are unlikely, but possible, it may be appropriate to explore whether deviations from size-balance have occurred in the data and then select MWCR or WIEE accordingly. One way to empirically check the size-balance assumption is to plot the cluster size against the cluster mean of each component of the vector $\boldsymbol{X}$. Although the definition of size-balance refers to distributions of $\boldsymbol{X}$ rather than expectations, any clear relationship in the plot would be an indication of departure from size-balance. Alternatively, the cluster size could be regressed on the cluster mean of $\boldsymbol{X}$. A significant effect of the cluster mean of $\boldsymbol{X}$ is evidence against size-balance.

Conditions 2 and 3 relate to the choice of working correlation. Even when Condition 1a is satisfied, Conditions 2 and 3 are necessary for the consistency of MWCR. Exchangeable working correlation satisfies both conditions. Simulation results (see Appendix B) based on a longitudinal data scenario suggest that MWCR(EX) can provide consistent and more efficient estimation than WIEE, even when the true correlation structure is not exchangeable but rather depends on the times of measurement for the members. MWCR(AR) provided biased estimation in both scenarios, due to violation of Conditions 2 and 3. The efficiency gain shown for MWCR(EX) is, however, not guaranteed when the true correlation differs greatly from exchangeability, although other authors (e.g., Park and Sin (1999)) have found that when cluster size is not informative, the use of standard GEE with exchangeable correlation can provide more efficient estimates that IEE when the true correlation is auto-regressive.

When there is a mixture of cluster-constant, cluster-varying size balanced, and cluster-varying non-size balanced covariates, MWCR should not be used; WIEE should be used instead. As was seen in the data example, even where cluster-varying covariates (T and T^2) have means that differ modestly across cluster sizes, the bias from the use of MWCR for the effects of these covariates appeared to be appreciable.

Huang and Leroux (2011) consider further populations for inference when cluster size is informative. MWCR could be extended to these populations.

Though the range of scenarios in which MWCR is unbiased is somewhat limited, and there are restrictions on the choice of working correlation, the method is simple to implement. Due to the possibility of increased efficiency it is worthy of consideration alongside the more generally applicable method (WIEE) based on an independence working correlation. The conditions we have identified for consistent estimation from MWCR can form a useful basis for considering whether the method is appropriate for specific data examples.

Acknowledgements

MP and AJC were supported by grant G0600657, and SRS by grant MC_US_A030_0014, both from the UK Medical Research Council. We thank the Delta Coordinating Committee for the use of the trial data.

Appendix A

In preparation for the proof, note that (2.1) can be rewritten as

$$\sum_{i=1}^{K} \boldsymbol{U}(\boldsymbol{\beta}; \boldsymbol{Y}_i^*, \boldsymbol{X}_i^*) = \sum_{i=1}^{K} \sum_{j=1}^{N_i} \frac{1}{\phi} \boldsymbol{g}_{ij}(Y_{ij} - \mu_{ij}) = \boldsymbol{0}, \tag{A.1}$$

where

$$\boldsymbol{g}_{ij} = \sum_{l=1}^{N_i} \frac{\partial \mu_{il}}{\partial \boldsymbol{\beta}} v(\mu_{il})^{-1/2} \hat{r}_{ilj} v(\mu_{ij})^{-1/2} = \sum_{l=1}^{N_i} \boldsymbol{X}_{il}\ v(\mu_{il})^{1/2} \hat{r}_{ilj} v(\mu_{ij})^{-1/2} \tag{A.2}$$

is an implicit weighting for the j^{th} measurement. As mentioned in Section 2.1, when $h^{-1}(\theta) = \theta$ and there are no covariates, $\boldsymbol{g}_{ij} = \hat{r}_{i+j}$.

Proof of Theorem. We show that the expectation of the contribution from a single cluster to estimating equations (2.3), evaluated at $\boldsymbol{\rho}_0$ and the true value of $\boldsymbol{\beta}$, is zero. Hence, estimating equations (2.3) are consistent.

Analogously to (A.1), (2.3) can be written as

$$\sum_{i=1}^{K} \frac{1}{\Delta_i} \sum_{j=1}^{N_i} \sum_{s \in \Lambda_{ij}} \frac{1}{\phi} \tilde{\boldsymbol{g}}_{ij(s)} \{Y_{ij} - \mu_{ij}^C\}, \tag{A.3}$$

where $\tilde{\boldsymbol{g}}_{ij(s)}$, analogously to $\boldsymbol{g}_{ij}$ in (A.2), is the implicit weighting for the jth member of cluster i when it is in subcluster s, and Λ_{ij} denotes the set of indices of the $\Delta_i m N_i^{-1}$ subclusters containing the j^{th} member of cluster i. Note that the total weight given to cluster i in (A.3) is $\Delta_i^{-1} \sum_{j=1}^{N_i} \sum_{s \in \Lambda_{ij}} \tilde{\boldsymbol{g}}_{ij(s)}$, the average of $\Delta_i m$ values of $m\tilde{\boldsymbol{g}}_{ij(s)}$. In the special case $h^{-1}(\theta) = \theta$ and no covariates, the average value of $m\tilde{\boldsymbol{g}}_{ij(s)}$ is a scalar and is the average, over each of the Δ_i $(m \times m)$ submatrices of $\hat{\boldsymbol{R}}_i$, of the sum of the elements of its inverse matrix.

For the sampling mechanism described immediately before the Theorem, let $\tilde{\boldsymbol{g}}$ denote the resulting implicit weighting for the chosen member when it is in the chosen subcluster (see (A.3)). Denote expectations of the distributions of Y, $\boldsymbol{X}$, $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{g}}$ under this sampling mechanism by $E^S(.)$. Note that $E^S(Y|\boldsymbol{X}) = E^C(Y|\boldsymbol{X})$.

It can be seen that the expectation of the contribution of a single cluster to (A.3), at $\boldsymbol{\rho} = \boldsymbol{\rho}_0$ and the true value of $\boldsymbol{\beta}$, is $\phi^{-1} m E^S_{Y,\boldsymbol{X},\tilde{\boldsymbol{X}},\tilde{\boldsymbol{g}},N}[\tilde{\boldsymbol{g}}\{Y - \mu^C(\boldsymbol{X})\}]$.

Now,

$$\begin{aligned}E^S_{Y,\boldsymbol{X},\tilde{\boldsymbol{X}},\tilde{\boldsymbol{g}},N}[\tilde{\boldsymbol{g}}\{Y-\mu^C(\boldsymbol{X})\}] &= E^S_{\boldsymbol{X},\tilde{\boldsymbol{X}},\tilde{\boldsymbol{g}}} E^S_{N|\boldsymbol{X},\tilde{\boldsymbol{X}},\tilde{\boldsymbol{g}}} E^S_{Y|\boldsymbol{X},\tilde{\boldsymbol{X}},\tilde{\boldsymbol{g}},N}[\tilde{\boldsymbol{g}}\{Y - \mu^C(\boldsymbol{X})\}] \\ &= E^S_{\boldsymbol{X},\tilde{\boldsymbol{X}},\tilde{\boldsymbol{g}}}[\tilde{\boldsymbol{g}} E^S_{N|\boldsymbol{X},\tilde{\boldsymbol{X}},\tilde{\boldsymbol{g}}}\{E^S(Y|\boldsymbol{X},\tilde{\boldsymbol{X}},\tilde{\boldsymbol{g}},N) - \mu^C(\boldsymbol{X})\}].\end{aligned}$$

From Conditions 2 and 3, respectively, it follows that $N \perp\!\!\!\perp \tilde{\boldsymbol{g}} \mid \boldsymbol{X}, \tilde{\boldsymbol{X}}$ and $Y \perp\!\!\!\perp \tilde{\boldsymbol{g}} \mid N, \boldsymbol{X}, \tilde{\boldsymbol{X}}$. So,

$$\begin{aligned}E^S_{Y,\boldsymbol{X},\tilde{\boldsymbol{X}},\tilde{\boldsymbol{g}},N}[\tilde{\boldsymbol{g}}\{Y - \mu^C(\boldsymbol{X})\}] &= E^S_{\boldsymbol{X},\tilde{\boldsymbol{X}},\tilde{\boldsymbol{g}}}[\tilde{\boldsymbol{g}} E^S_{N|\boldsymbol{X},\tilde{\boldsymbol{X}}}\{E^S(Y|\boldsymbol{X},\tilde{\boldsymbol{X}},N) - \mu^C(\boldsymbol{X})\}] \\ &= E^S_{\boldsymbol{X},\tilde{\boldsymbol{X}},\tilde{\boldsymbol{g}}}[\tilde{\boldsymbol{g}}\{E^S(Y \mid \boldsymbol{X},\tilde{\boldsymbol{X}}) - \mu^C(\boldsymbol{X})\}].\end{aligned}$$

Then, using Condition 1a,

$$\begin{aligned}E^S_{Y,\boldsymbol{X},\tilde{\boldsymbol{X}},\tilde{\boldsymbol{g}},N}[\tilde{\boldsymbol{g}}\{Y - \mu^C(\boldsymbol{X})\}] &= E^S_{\boldsymbol{X},\tilde{\boldsymbol{X}},\tilde{\boldsymbol{g}}}[\tilde{\boldsymbol{g}}\{E^S(Y \mid \boldsymbol{X}) - \mu^C(\boldsymbol{X})\}] \\ &= E^S_{\boldsymbol{X},\widetilde{\boldsymbol{X}},\tilde{\boldsymbol{g}}}[\tilde{\boldsymbol{g}}\{E^C(Y \mid \boldsymbol{X}) - \mu^C(\boldsymbol{X})\}] \\ &= E^S_{\boldsymbol{X},\tilde{\boldsymbol{X}},\tilde{\boldsymbol{g}}}[\tilde{\boldsymbol{g}} \times 0] = \mathbf{0}.\end{aligned}$$

Appendix B

We used two simulation scenarios to demonstrate bias from the application of MWCR with an auto-regressive working correlation (MWCR(AR)) when Conditions 2 and 3 are violated.

Bias from violation of Condition 2.

We considered a simple longitudinal scenario where the timing of the measurements is associated with the size of the clusters, so Condition 2 is violated if auto-regressive is selected as the working correlation.

We conducted 1,000 simulations and, for each simulation, generated 100 clusters. Date were generated for cluster i as follows.

(1) Generate $B_i \sim N(0,1)$ and $V_i = \text{logit}^{-1}(B_i)$.

(2) Generate $N_i \sim Binomial(2, V_i) + 2$. So, $N_i = 2, 3$ or 4.

(3) The times of observations are $\boldsymbol{T}_i = (1,6)^T$ if $N_i = 2$, $\boldsymbol{T}_i = (1,3,5)^T$ if $N_i = 3$ and $\boldsymbol{T}_i = (1,2,3,4)^T$ if $N_i = 4$.

Table 3. Bias in MWCR(AR) from the violation of Condition 2.

Method	$\hat{\beta}_0(ese(\hat{\beta}_0))$	$\hat{\beta}_1(ese(\hat{\beta}_1))$	RE$(\hat{\beta}_0, \hat{\beta}_1)$	CVR$(\hat{\beta}_0, \hat{\beta}_1)$
WIEE	0.50 (0.146)	1.00 (0.241)	-	(0.95,0.94)
MWCR(EX)	0.50 (0.139)	1.00 (0.175)	(1.11,1.90)	(0.95,0.94)
MWCR(AR)	0.35 (0.133)	1.06 (0.165)	-	(0.83,0.86)

Table 4. Bias in MWCR(AR) from the violation of Condition 3.

Method	$\hat{\beta}_0(ese(\hat{\beta}_0))$	$\hat{\beta}_1(ese(\hat{\beta}_1))$	RE$(\hat{\beta}_0, \hat{\beta}_1)$	CVR$(\hat{\beta}_0, \hat{\beta}_1)$
WIEE	0.50 (0.148)	1.00 (0.234)	-	(0.96,0.94)
MWCR(EX)	0.50 (0.138)	1.00 (0.150)	(1.16,2.41)	(0.96,0.95)
MWCR(AR)	0.76 (0.139)	0.74 (0.144)	-	(0.54,0.56)

(4) Generate $X_{ij} \sim Bernouli(0.5)$.

(5) Calculate $\mu_{ij} = \gamma_0 + \gamma_1 X_{ij} + \gamma_2 B_i + \gamma_3 B_i X_{ij}$.

(6) Generate $\boldsymbol{y}_i \sim MVN(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i(\rho))$, where $\boldsymbol{\Sigma}_i(\rho)$ is the matrix with $(k,l)^{th}$ element equal to $\rho^{|T_{il}-T_{ik}|}$ and T_{ik}, T_{il} are the times of the k^{th} and l^{th} measurements in cluster i.

We chose the simulation parameters as $\gamma_0 = 0.5$, $\gamma_1 = \gamma_2 = \gamma_3 = 1$ and $\rho = 0.7$. We fit the model $E(Y \mid X = x) = \beta_0 + \beta_1 x$ for the population of typical members, using WIEE, MWCR(AR), and MWCR(EX). Table 3 shows the results. It can be seen that MWCR(AR) gave biased estimation. MWCR(EX) provided unbiased estimation and efficiency gains compared to WIEE even though the working correlation assumption was false.

Bias from violation of Condition 3.

In this scenario, to isolate Condition 3 we considered constant cluster sizes. The times of measurements were allowed to be different for different clusters and, after adjusting for X, the outcome was higher, on average, when members were more widely spread in time. So, Condition 3 is violated if auto-regressive is selected as the working correlation. The simulation steps were as follows.

(1) Same as step (1) earlier.

(2) N_i=3.

(3) There are two measurement-time patterns. If $B_i < 0$, $\boldsymbol{T}_i = (1,2,3)^T$, and if $B_i > 0$, $\boldsymbol{T}_i = (1,3,6)^T$.

(4)-(6) Same as steps (4)-(6) earlier.

The values of parameters were the same as in the scenario above. Again, MWCR(AR) was biased, while MWCR(EX) provided unbiased estimates with increased efficiency compared to WIEE (see Table 4).

References

Benhin, E., Rao, J. N. K. and Scott, A. J. (2005). Mean estimating equation approach to analysing cluster-correlated data with non-ignorable cluster sizes. *Biometrika* **92**, 435-450.

French, B., Farjah, F., Flum, D. R. and Heagerty, P. (2012). A general framework for estimating volume-outcome associations from longitudinal data. *Statist. Medicine* **31**, 366-382.

Chiang, T. C. and Lee, K. Y. (2008). Efficient methods for informative cluster size data. *Statist. Sinica* **80**, 121-123.

Delta Coordinating Committee (1996). Delta: A Randomized Double-Blind Controlled Trial Comparing Combinations of Zidovudine Plus Didanosine or Zalcitabine With Zidovudine Alone in HIV-Infected Individuals. *Lancet* **348**, 283-291.

Hoffman, E. B., Sen, P.K. and Weinberg, C. R. (2001). Within-cluster resampling. *Biometrika* **88**, 1121-1134.

Huang, Y. and Leroux, B. (2011). Informative cluster sizes for subcluster-level covariates and weighted generalized estimating equations. *Biometrics* **67**, 843-851.

Liang, K-Y. and Zeger, S. L. (1986). Longitudinal data analysis with generalized linear models. *Biometrika* **73**, 13-22.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, Second Edition. Wiley, New York.

Park, T. and Sin, D-Y. (1999). On the use of working correlation matrices in the GEE approach for longitudinal data. *Comm. Statist. Simulation Comput.* **28**, 1011-1029.

Pepe, M. S. and Anderson, G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Comm. Statist.* **32**, 1939-1951.

Preisser, J. S., Lohman, K. K. and Rathouz, P. J. (2002). Performance of weighted estimating equations for longitudinal binary data with dropouts missing at random. *Statist. Medicine* **21**, 3035-3054.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Amer. Statist. Assoc.* **90**, 106-121.

Williamson, J. M., Datta, S. and Satten, G. A. (2003). Marginal analyses of clustered data when cluster size is informative. *Biometrics* **59**, 36-42.

Research Department of Infection and Population Health, University College London, London WC1E 6JB, UK.

E-mail: m.pavlou@ucl.ac.uk

MRC Biostatistics Unit, Institute of Public Health, University Forvie Site, Robinson Way, Cambridge CB2 0SR, UK.

E-mail: shaun.seaman@mrc-bsu.cam.ac.uk

MRC Clinical Trials Unit, Hub for Trials Methodology Research, Aviation House, 125 Kingsway, London WC2B 6NH, UK.

E-mail: anc@ctu.mrc.ac.uk

(Received October 2011; accepted August 2012)