Biometrics 70, 449–456
June 2014

DOI: 10.1111/biom.12151

Methods for Observed-Cluster Inference When Cluster Size Is Informative: A Review and Clarifications

Shaun R. Seaman,[1,*] Menelaos Pavlou,[2] and Andrew J. Copas[3]

[1]MRC Biostatistics Unit, Cambridge CB2 0SR, U.K.
[2]Department of Statistical Science, University College London, London WC1E 6BT, U.K.
[3]MRC Clinical Trials Unit at University College London, London WC2B 6NH, U.K.
email: shaun.seaman@mrc-bsu.cam.ac.uk

Summary. Clustered data commonly arise in epidemiology. We assume each cluster member has an outcome Y and covariates $\boldsymbol{X}$. When there are missing data in Y, the distribution of Y given $\boldsymbol{X}$ in all cluster members ("complete clusters") may be different from the distribution just in members with observed Y ("observed clusters"). Often the former is of interest, but when data are missing because in a fundamental sense Y does not exist (e.g., quality of life for a person who has died), the latter may be more meaningful (quality of life conditional on being alive). Weighted and doubly weighted generalized estimating equations and shared random-effects models have been proposed for observed-cluster inference when cluster size is informative, that is, the distribution of Y given $\boldsymbol{X}$ in observed clusters depends on observed cluster size. We show these methods can be seen as actually giving inference for complete clusters and may not also give observed-cluster inference. This is true even if observed clusters are complete in themselves rather than being the observed part of larger complete clusters: here methods may describe imaginary complete clusters rather than the observed clusters. We show under which conditions shared random-effects models proposed for observed-cluster inference do actually describe members with observed Y. A psoriatic arthritis dataset is used to illustrate the danger of misinterpreting estimates from shared random-effects models.

Key words: Bridge distribution; Immortal cohort inference; Informative missingness; Missing not at random; Mortal cohort inference; Semi-continuous data.

1. Introduction

Clustered data are common in epidemiology. Repeated measures are clustered in individuals; teeth in patients; pups in litters. Suppose interest is in the association between outcome Y and covariates $\boldsymbol{X}$ measured on members of the clusters. Often Y and $\boldsymbol{X}$ are missing for some members of sampled clusters. For simplicity, we assume that a member's $\boldsymbol{X}$ is observed whenever Y is observed. We call members with observed Y "observed members," those with missing Y "missing members," the original clusters "complete clusters," and the subclusters that remain after discarding missing members "observed clusters."

Missing data may arise because although a variable could, in principle, be measured, circumstances meant it was not, for example, because an individual missed a visit. We call such missing data "potentially observable." When missing data are potentially observable, a model can be proposed for the distribution of Y given $\boldsymbol{X}$ in all cluster members, and methods used that, under specified assumptions about the missingness (e.g., missing at random, MAR), give consistent estimates for this model. We call this "complete-cluster inference."

Alternatively, missing data may arise because in a fundamental sense a variable does not exist. We call such missing data "unobservable." Three examples of unobservable Y are measures of: (1) cognitive function of an individual after death; (2) degree of disablement of an individual who is not disabled; (3) health of a tooth that has been lost. Although missing Y could be set to zero when a patient is dead/not disabled/tooth is lost, in practice often a model is instead proposed for Y given $\boldsymbol{X}$ in observed members only (so conditional on alive/disabled/tooth not lost). We call this "observed-cluster inference." Sometimes observed-cluster inference may be of interest even when missing data are potentially observable. When missing data are unobservable "complete-cluster" inference is philosophically problematic: what does it mean to model cognitive function in dead people?

When the size M of complete clusters varies, it is usually assumed that Y is independent of M given $\boldsymbol{X}$. In observed clusters, however, Y and N may be conditionally dependent given $\boldsymbol{X}$, where N is size of observed cluster. For example, in a dental study, the fewer teeth a patient has, the worst their condition tends to be. This is called "informative cluster size" (ICS).

So far we have assumed observed clusters are generated from complete clusters by excluding missing members, but ICS can also arise where observed clusters are complete in themselves. For example, in toxicology, exposed dams who are more sensitive to a toxin may tend to have smaller litters and offspring with greater probability of deformation than less sensitive dams, so that Y (pup being deformed) and N (litter size) are dependent given X (exposure of dam).

We shall show that three of the methods proposed for observed-cluster inference under ICS, viz. weighted and doubly weighted generalized estimating equations (GEE) and shared random effects models, can be seen as actually giving inference for complete clusters. When the Y-$\boldsymbol{X}$ associations in

complete and observed clusters are the same, the distinction is unimportant. However, ICS causes them to differ in general. So, it is important to understand when methods proposed for observed-cluster inference really do describe observed clusters. In the literature on modeling repeated measures in cohorts with high death rates (Dufouil et al., 2004; Kurland et al., 2009) a distinction has been made between complete-cluster (termed "immortal-cohort") inference and observed-cluster ("mortal-cohort") inference. However, conditions under which the two inferences are equivalent have not been set out, and in the wider literature the distinction seems to be less well recognized.

In Section 2 we define notation and discuss methods for complete-cluster inference from observed data. Section 3 defines ICS and discusses how ICS relates to missing-data mechanisms. Section 4 relates two weighted GEE methods, one proposed for complete-cluster inference in the missing-data literature, and one for observed-cluster inference in the ICS literature. We also show that doubly weighted GEE, proposed for observed-cluster inference, actually give complete- rather than observed-cluster inference, and that, moreover, there is no single complete-cluster inference. Shared random-effects models give complete-cluster inference, but have also been used for observed-cluster inference. In Section 5 we discuss when this is valid, and in Section 6 we use a psoriatic arthritis dataset to illustrate that some parameters of such a model may be relevant to observed clusters but others not. In brief, we replicate an analysis of association between disability and covariates, with measurements clustered by patient. Our interest is in how sex affects degree of disability in the "observed clusters" of measurements where degree is greater than zero, that is, given disability. The analysis uses models for probability of disability and for degree of disability given disability which share a random intercept. Because probability of disability is higher in women than in men with the same intercept and other covariates, intercept and sex are not independent given disability and other covariates. Consequently, the effect of sex on degree of disability given disability is less than is suggested by the estimated parameter.

2. Notation and Complete-Cluster Inference

Let K be the number of complete clusters in the sample. When needed we use subscript i to index cluster, but usually omit this. Let M (known) be size of complete cluster. Let Y_j and $\boldsymbol{X}_j$ $(j = 1, \ldots, M)$ be outcome and covariate vector, respectively, for member j of the complete cluster, and $\tilde{\boldsymbol{Y}} = (Y_1, \ldots, Y_M)^T$ and $\tilde{\boldsymbol{X}} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_M)$. Let $R_j = 1$ if Y_j is observed, $R_j = 0$ if Y_j is missing, and $\boldsymbol{R} = (R_1, \ldots, R_M)^T$. $\boldsymbol{R}$ is always observed. Members with $R_j = 1$ are "observed members"; those with $R_j = 0$ are "missing members." Let $N = \sum_{j=1}^{M} R_j$ be size of observed cluster. Assume $(\boldsymbol{R}_i, \tilde{\boldsymbol{X}}_i, \tilde{\boldsymbol{Y}}_i)$ $(i = 1, \ldots, K)$ are i.i.d. For any value $\boldsymbol{r}$ of $\boldsymbol{R}$, partition $\tilde{\boldsymbol{Y}} = (\tilde{\boldsymbol{Y}}_{(r)}, \tilde{\boldsymbol{Y}}_{(\bar{r})})$, where Y_j belongs to $\tilde{\boldsymbol{Y}}_{(r)}$ if $r_j = 1$ and to $\tilde{\boldsymbol{Y}}_{(\bar{r})}$ if $r_j = 0$. For example, $\tilde{\boldsymbol{Y}}_{((1,0,1))} = (Y_1, Y_3)^T$ and $\tilde{\boldsymbol{Y}}_{(\overline{(1,0,1)})} = Y_2$. Partition $\tilde{\boldsymbol{X}}$ likewise, except that if some elements of $\boldsymbol{X}$ are observed even on missing members, these elements belong to $\tilde{\boldsymbol{X}}_{(r)}$.

Data are missing at random (MAR) if $P(\boldsymbol{R} = \boldsymbol{r} \mid \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}) = \pi(\boldsymbol{r}, \tilde{\boldsymbol{X}}_{(r)}, \tilde{\boldsymbol{Y}}_{(r)})$ $\forall \boldsymbol{r}$ for some function $\pi(.)$ (informally, $P(\boldsymbol{R} \mid \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}) = P(\boldsymbol{R} \mid \tilde{\boldsymbol{X}}_{(R)}, \tilde{\boldsymbol{Y}}_{(R)}, M)$) and missing completely at random (MCAR) if $P(\boldsymbol{R} \mid \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}) = P(\boldsymbol{R} \mid M)$ (Seaman et al., 2013) (note M is a function of $\tilde{\boldsymbol{X}}$, as $\tilde{\boldsymbol{X}}$ has M columns). Otherwise they are missing not at random (MNAR). We say data are missing with equal probability (MWEP) if $P(R_j = 1 \mid \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}, N) = N/M$ $\forall j$. MCAR means that which members are observed does not depend on $\boldsymbol{X}$ or Y values in the cluster. This would be so if, for example, missing data had been lost by the researchers. MAR allows missingness to depend on data on observed members plus any observed data on missing members. For example in a longitudinal study individuals' probability of dropout may depend on past health measurements but not on current health. If it also depends on current health, the data are MNAR. MWEP means the number N of observed members may depend on $\boldsymbol{X}$ and Y but given this number all sets of N observed members are equally likely. This could be so if missingness depends only on cluster-level summaries of $\boldsymbol{X}$ and Y.

The missingness process is monotone if $P(R_{j+1} = 0 \mid R_j = 0, M) = 1$ $\forall j$. (N, M) then defines $\boldsymbol{R}$ and *vice versa*. If $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_M, Y_M)$ are exchangeable given M, we say "members of complete clusters are exchangeable." Indices $\{1, \ldots, N\}$ can then be assigned to observed members and $\{N+1, \ldots, M\}$ to missing members. Missingness is then monotone.

To make "complete-cluster" inference, a model is specified for $\tilde{\boldsymbol{Y}}$ given $\tilde{\boldsymbol{X}}$. To fit this using observed data $(\tilde{\boldsymbol{Y}}_{(R)}, \tilde{\boldsymbol{X}}_{(R)})$, an assumption (e.g., MAR) is made about the missingness process and a method used that is valid under this assumption, for example, inverse probability weighting (IPW) or random-effect models (Albert and Follmann, 2009). We consider two approaches to complete-cluster inference that relate to methods proposed for observed-cluster inference. The first specifies a (marginal) model for $E(Y_j \mid \boldsymbol{X}_j = \boldsymbol{x}, M = m)$ and assumes

$$E(Y_j \mid \boldsymbol{X}_j = \boldsymbol{x}, M = m_1) = E(Y_k \mid \boldsymbol{X}_k = \boldsymbol{x}, M = m_2) \quad \forall 1 \le j \le m_1;\ 1 \le k \le m_2, \tag{1}$$

so that we can define $e_C(\boldsymbol{x}) = E(Y_j \mid \boldsymbol{X}_j = \boldsymbol{x}, M = m)$. This model is fitted to observed clusters using GEE with IPW. The second approach uses a shared random-effects model. This gives cluster-specific inference, but random effects can be integrated out to get $e_C(\boldsymbol{x})$.

3. Informative Cluster Size

3.1. *Semi-Parametric Marginal Models*

For each cluster with $N \ge 1$, let H be the index of a randomly selected member of the observed cluster. So, $P(H = j \mid \boldsymbol{R}, \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}) = R_j/N$. Marginal inference for the *population of typical observed members* and marginal inference for the *population of all observed members* mean estimating the parameters of a model for $e_T(\boldsymbol{x}) = E(Y_H \mid \boldsymbol{X}_H = \boldsymbol{x}, N \ge 1)$ and for $e_A(\boldsymbol{x}) = E(NY_H \mid \boldsymbol{X}_H = \boldsymbol{x}, N \ge 1)/E(N \mid \boldsymbol{X}_H = \boldsymbol{x}, N \ge 1)$, respectively. Whereas $e_T(\boldsymbol{x})$ is the expectation of Y given $\boldsymbol{X} = \boldsymbol{x}$ giving

equal weight to each observed cluster, $e_A(\boldsymbol{x})$ gives equal weight to each observed member. Clusters with $N = 0$ play no role in $e_T(\boldsymbol{x})$ or $e_A(\boldsymbol{x})$.

Hoffman et al. (2001), Williamson et al. (2003) and Benhin et al. (2005) define non-informative cluster size (NICS) as $E(Y_H \mid \boldsymbol{X}_H = \boldsymbol{x}, N = n) = E(Y_H \mid \boldsymbol{X}_H = \boldsymbol{x}, N \geq 1) \ \ \forall \boldsymbol{x},\ n \geq 1$. Otherwise cluster size is informative (ICS). Under NICS, $e_T(\boldsymbol{x}) = e_A(\boldsymbol{x})\ \forall \boldsymbol{x}$. Under ICS, $e_T(\boldsymbol{x}) \neq e_A(\boldsymbol{x})$ in general. They advocate using $e_T(\boldsymbol{x})$. Use of $e_A(\boldsymbol{x})$ has been proposed for mortal cohorts when missing data are due to death, and for modeling degree of disability or health of teeth when missing data are due to non-disabled patients or absent teeth (Dufouil et al., 2004; Kurland et al., 2009; Su et al., 2011; Li et al., 2011). Hoffman et al. (2001) gave an estimator for $e_T(\boldsymbol{x})$. Williamson et al. (2003) and Benhin et al. (2005) gave an asymptotically equivalent and computationally less intensive method: weighted independence estimating equations (WIEE) (see also Wang et al. (2011) for three-level data). The same equations without weighting (IEE) estimate $e_A(\boldsymbol{x})$. We describe WIEE and IEE in Section 4.1.

3.2. *Random-Effects Models*

Dunson et al. (2003), Gueorguieva (2005), Chen et al. (2011), and Neuhaus and McCulloch (2011) consider cluster-specific inference using a linear or generalized linear mixed model (LMM/GLMM). They interpret NICS to mean the random effects $\boldsymbol{u}$ in the mixed model are independent of N, and ICS to mean they are not. NICS in this sense implies NICS in the sense of Hoffman et al., but the converse is not true. To deal with ICS when fitting the LMM/GLMM, several authors have combined it with a model for N or $\boldsymbol{R}$, with the same or correlated random effect (Dunson et al., 2003; Gueorguieva, 2005; Chen et al., 2011; Su et al., 2009, 2011; Li et al., 2011). We discuss this model in Section 5.

3.3. *Relating ICS to Missingness Mechanisms*

Hoffman et al. (2001) wrote that ICS is "closely related" to violation of the MCAR condition. In fact, MCAR is not a sufficient condition for NICS. For example, suppose all complete clusters have size $M = 2$ and have $\tilde{\boldsymbol{Y}} = (0, 1)^T$, there are no covariates, and $P\{\boldsymbol{R} = (1, 1) \mid \tilde{\boldsymbol{Y}}\} = P\{\boldsymbol{R} = (1, 0) \mid \tilde{\boldsymbol{Y}}\} = 1/2$. It is easy to show that $e_T = 1/4$ but $e_A = 1/3$.

Proposition 1
Cluster size will be non-informative if data are MCAR and, moreover, either i) equation (1) holds, or ii) $N \perp\!\!\!\perp M$ and the data are MWEP.

Note (1) is often assumed with GEEs, but $N \perp\!\!\!\perp M$ is unlikely, as $N \leq M$. Proofs of Propositions are in Web Appendices A and E. Just as both ICS and NICS can arise from MCAR mechanisms, so they can from MAR and MNAR (examples in Web Appendix B).

When (1) holds, so $e_C(\boldsymbol{x})$ is defined, a sufficient condition for $e_C(\boldsymbol{x}) = e_T(\boldsymbol{x})$ is MWEP and $P(N \geq 1) = 1$, because the Y-X relation in a randomly chosen member of an observed cluster is then the same as in a random member of the corresponding complete cluster.

4. Weighted and Doubly Weighted GEE

4.1. *Weighted GEE (WGEE)*

Assume (1) holds and $e_C(\boldsymbol{x}) = g^{-1}(\boldsymbol{\beta}^T\boldsymbol{x})$, where g is a link function. If $\tilde{\boldsymbol{Y}}$ and $\tilde{\boldsymbol{X}}$ were observed, $\boldsymbol{\beta}$ could be estimated with GEE. With missing data, WGEE can be used. These weight member j by $R_j/P(R_j = 1 \mid \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}})$. Robins et al. (1995) proposed use of WGEE when M does not vary, missingness is monotone and MAR, and $P(N \geq 1) = 1$.

When data are MWEP and $P(N \geq 1) = 1$, weights $R_j/P(R_j = 1 \mid \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}, N) = R_jM/N$ can be used instead (proof in Web Appendix C). In this case, $e_C(\boldsymbol{x}) = e_T(\boldsymbol{x})$ (Section 3.3), so WGEE with weights R_jM/N also give observed-cluster inference. In fact, with independence working correlation they are the WIEE proposed by Williamson et al. (2003) for estimating $\boldsymbol{\beta}$ in $e_T(\boldsymbol{x}) = g^{-1}(\boldsymbol{\beta}^T\boldsymbol{x})$. So, WIEE have a dual interpretation: they estimate $e_T(\boldsymbol{x})$ under any missingness mechanism; and $e_C(\boldsymbol{x})$ when data are MWEP and $P(N \geq 1) = 1$.

WIEE without weights (IEE) estimate $\boldsymbol{\beta}$ in a model $e_A(\boldsymbol{x}) = g^{-1}(\boldsymbol{\beta}^T\boldsymbol{x})$ (Dufouil et al., 2004).

4.2. *Doubly Weighted GEE (DWGEE)*

If there is ICS and the distribution of $\boldsymbol{X}$ depends on N, interpretation of $e_T(\boldsymbol{x})$ may be awkward, because the Y-$\boldsymbol{X}$ association is confounded by N (Williamson et al., 2003). For example, let X be binary and $E(Y_j \mid X_j, N) = E(Y_j \mid N)$ and $P(X_j = 1 \mid N)$ be increasing functions of N. Then typical members with $X = 1$ tend to come from larger clusters than typical members with $X = 0$, so $e_T(1) > e_T(0)$ even though X has no effect on Y within clusters.

Huang and Leroux (2011) proposed DWGEE1 and DWGEE2. DWGEE1 can be used when $\boldsymbol{X}$ is categorical and every observed cluster contains at least one member with each of the possible values of $\boldsymbol{X}$. DWGEE1 are the same as WIEE except that member j is inversely weighted not by M/N but by the total number of observed members in the same cluster who have $\boldsymbol{X} = \boldsymbol{X}_j$. Thus the total weight of members with $\boldsymbol{X} = \boldsymbol{x}$ is the same for all possible $\boldsymbol{x}$. Rather than estimating $e_T(\boldsymbol{x})$, DWGEE1 estimate $E(Y \mid \boldsymbol{X})$ in the population formed by each cluster in the population contributing one member with each possible value of $\boldsymbol{X}$.

DWGEE2 was proposed for when not all observed clusters contain a member with each possible value of $\boldsymbol{X}$. In DWGEE2 observed member j is inversely weighted by the *expected* (rather than *actual*, as in DWGEE1) number of observed members with $\boldsymbol{X} = \boldsymbol{X}_j$. In Web Appendix D we show that DWGEE2 estimates $E(Y \mid \boldsymbol{X})$ in a population of larger "complete" clusters in which each cluster contains at least one member with each possible value of $\boldsymbol{X}$. Each cluster in the dataset is considered to be the observed component of one of these larger clusters, with the rest being missing. The problem with this is that, unless observed clusters really do arise from larger clusters in which all values of $\boldsymbol{X}$ are represented (which is not so in Huang and Leroux's example), the larger clusters are purely hypothetical and it is unclear why they should be of scientific interest. Further, as shown in Web Appendix D, the distribution of Y given $\boldsymbol{X}$ in the hypothetical population of complete clusters depends on which predictors are included in the model for the expected number

with $\boldsymbol{X} = \boldsymbol{x}$, and there is no obvious reason to prefer one set of predictors to any other.

5. Random-Effect Models

5.1. *LMM, GLMM, and Shared Random Effect Model*

The general form of the LMM is (continuing to omit the subscript i for cluster)

$$Y_j \mid \tilde{\boldsymbol{X}}, \boldsymbol{u} \sim N(\boldsymbol{\beta}^T \boldsymbol{X}_j + \boldsymbol{u}^T \boldsymbol{Z}_j,\ \sigma^2) \quad (j = 1, \ldots, M) \tag{2}$$

$$\boldsymbol{u} \sim f_u(\boldsymbol{u}; \boldsymbol{\alpha}), \qquad E(\boldsymbol{u}) = 0 \qquad \text{and } \boldsymbol{u} \perp\!\!\!\perp \tilde{\boldsymbol{X}} \tag{3}$$

$$Y_1 \perp\!\!\!\perp Y_2 \perp\!\!\!\perp \ldots \perp\!\!\!\perp Y_M \mid \tilde{\boldsymbol{X}}, \boldsymbol{u} \tag{4}$$

where $\boldsymbol{Z}$ is a subvector of $\boldsymbol{X}$, and $\boldsymbol{u}$ a cluster-specific latent variable. This is a model for Y-$\boldsymbol{X}$ association in complete clusters. Assumption $\boldsymbol{u} \perp\!\!\!\perp \tilde{\boldsymbol{X}}$ means that $\boldsymbol{u} \perp\!\!\!\perp M$ and hence that size of *complete* clusters is non-informative. Elements of $\boldsymbol{X}$ not in $\boldsymbol{Z}$ are said to have fixed effects; those in $\boldsymbol{Z}$ have random effects. It follows from (2) and (3) that $e_C(\boldsymbol{x}) = \boldsymbol{\beta}^T \boldsymbol{x}$. So, $\boldsymbol{\beta}$ also has a marginal interpretation in complete clusters. LMMs are a special case of GLMMs. In GLMMs, $Y_j \mid \tilde{\boldsymbol{X}}, \boldsymbol{u}$ is assumed to belong to the exponential family, (2) is replaced by

$$E(Y_j \mid \tilde{\boldsymbol{X}}, \boldsymbol{u}) = g^{-1}(\boldsymbol{\beta}^T \boldsymbol{X}_j + \boldsymbol{u}^T \boldsymbol{Z}_j) \quad (j = 1, \ldots, M) \tag{5}$$

where $g(.)$ is the link function, and (3) and (4) are assumed to hold.

If Y is binary, $\boldsymbol{Z} = 1$ and $\boldsymbol{u}$ has a bridge distribution with rescaling parameter ϕ ($0 < \phi < 1$), then $e_c(\boldsymbol{x}) = \phi \boldsymbol{\beta}^T \boldsymbol{x}$ and so $\boldsymbol{\beta}$ (in combination with ϕ) has a marginal interpretation in complete clusters (Wang and Louis, 2003). More generally, $\boldsymbol{\beta}$ does not have a marginal interpretation, though $e_C(\boldsymbol{x})$ can be calculated as $e_C(\boldsymbol{x}) = \int g^{-1}(\boldsymbol{\beta}^T \boldsymbol{x} + \boldsymbol{u}^T \boldsymbol{z})\ f_u(\boldsymbol{u}; \boldsymbol{\alpha})\ \mathrm{d}\boldsymbol{u}$.

The MLE of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ from fitting the mixed model to observed clusters is consistent when data are MAR, but not, in general, when MNAR. However, Neuhaus and McCulloch (2011) showed that for LMMs, if (i) $\boldsymbol{X}$ includes an intercept term, (ii) $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_M$ are i.i.d., (iii) $P(\boldsymbol{R} \mid \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}, \boldsymbol{u}) = P(\boldsymbol{R} \mid M, \boldsymbol{u})$, and (iv) the only random effect is an intercept (i.e., $\boldsymbol{Z} = \boldsymbol{1}$), then $\boldsymbol{\beta}$ is consistently estimated except for the intercept. They found the same was approximately true of GLMMs. More generally, they say that if $\boldsymbol{u}_{\text{sub}}$ and $\boldsymbol{X}_{\text{sub}}$ are subvectors of $\boldsymbol{u}$ and $\boldsymbol{X}$ with $\boldsymbol{X}_{\text{sub}} \perp\!\!\!\perp \boldsymbol{u}_{\text{sub}}$ and $P(\boldsymbol{R} \mid \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}, \boldsymbol{u}) = P(\boldsymbol{R} \mid M, \boldsymbol{u}_{\text{sub}})$, then their results suggest that the MLE of elements of $\boldsymbol{\beta}$ corresponding to $\boldsymbol{X}_{\text{sub}}$ will be approximately unbiased.

For MNAR data, a model for $P(\boldsymbol{R} \mid \tilde{\boldsymbol{X}}, \boldsymbol{u})$ can be added to the LMM/GLMM. The result is a shared random-effects model (Albert and Follmann, 2009). When

$$P(\boldsymbol{R} = \boldsymbol{r} \mid \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}, \boldsymbol{u}) = \pi(\boldsymbol{r}, \tilde{\boldsymbol{X}}_{(r)}, \boldsymbol{u}) \qquad \forall \boldsymbol{r} \tag{6}$$

for some function $\pi(.)$, the MLEs of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ from this model are consistent. An indirect way (Su et al., 2009; Li et al., 2011; Su et al., 2011) to model $P(\boldsymbol{R} \mid \tilde{\boldsymbol{X}}, \boldsymbol{u})$ is to introduce another random effect $\boldsymbol{v}$, assume $Y_j \perp\!\!\!\perp \boldsymbol{v} \mid \tilde{\boldsymbol{X}}, \boldsymbol{u}$, and specify models $f_{u,v}(\boldsymbol{u}, \boldsymbol{v}; \boldsymbol{\alpha})$ for the distribution of $(\boldsymbol{u}, \boldsymbol{v})$ and $\pi^*(r, \tilde{\boldsymbol{X}}_{(r)}, \boldsymbol{v})$ for $P(\boldsymbol{R} = \boldsymbol{r} \mid \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}, \boldsymbol{u}, \boldsymbol{v})$. We call the resulting model for $(\tilde{\boldsymbol{Y}}, \boldsymbol{R})$ "a correlated random-effects model." It is a special case of the shared random-effects model, with $\pi(\boldsymbol{r}, \tilde{\boldsymbol{X}}_{(r)}, \boldsymbol{u}) = \int \pi^*(r, \tilde{\boldsymbol{X}}_{(r)}, \boldsymbol{v}) f_v(\boldsymbol{v} \mid \boldsymbol{u}; \boldsymbol{\alpha})\ \mathrm{d}\boldsymbol{v}$ and $f_u(\boldsymbol{u}; \boldsymbol{\alpha}) = \int f_{u,v}(\boldsymbol{u}, \boldsymbol{v}; \boldsymbol{\alpha}) \mathrm{d}\boldsymbol{v}$.

5.2. *Interpretation of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ in Complete Clusters*

Partition $\boldsymbol{X}$ and $\boldsymbol{\beta}$ as $\boldsymbol{X} = (X^{(l)}, \boldsymbol{X}^{(-l)})^T$ and $\boldsymbol{\beta} = (\beta^{(l)}, \boldsymbol{\beta}^{(-l)})$, where $X^{(l)}$ and $\beta^{(l)}$ are the lth elements of $\boldsymbol{X}$ and $\boldsymbol{\beta}$, respectively. If $X^{(l)}$ has a random effect, partition $\boldsymbol{u}$ as $\boldsymbol{u} = (u^{(l)}, \boldsymbol{u}^{(-l)})$, where $u^{(l)}$ corresponds to $X^{(l)}$, and partition $\boldsymbol{Z}$ similarly. If $X^{(l)}$ has a fixed effect, $u^{(l)} = z^{(l)} = 0$, $\boldsymbol{u}^{(-l)} = \boldsymbol{u}$ and $\boldsymbol{z}^{(-l)} = \boldsymbol{z}$. Let $\boldsymbol{I}^{(l)}$ denote a vector of the same length as $\boldsymbol{X}$, with lth element equal to one and all other elements equal to zero.

within-cluster effects

If $X^{(l)}$ is cluster varying with fixed effect, $\beta^{(l)}$ is its within-complete-cluster effect in clusters of size $M \geq 2$. That is, if two members of the same complete cluster have $\boldsymbol{X}$ values that differ only by $\boldsymbol{I}^{(l)}\epsilon$ for some ϵ, then their expected Y values differ by $\beta^{(l)}\epsilon$ for an LMM. In a GLMM, the expected value is transformed by link function g; for example, for logit link, $\beta^{(l)}\epsilon$ is their log odds ratio. If $X^{(l)}$ is cluster varying with random effect, $\beta^{(l)}$ and $\mathrm{Var}(u^{(l)})$ are the mean and variance of the within-cluster effect.

between-cluster effects

$\beta^{(l)}$ and $\boldsymbol{\alpha}$ can be interpreted in terms of differences between expected Y in members of different complete clusters. That is, if for some ϵ, two complete clusters are randomly sampled conditional on one containing a member with $\boldsymbol{X} = \boldsymbol{x}$ and the other a member with $\boldsymbol{X} = \boldsymbol{x} + \boldsymbol{I}^{(l)}\epsilon$, then the difference between the expected Y values of these two members is

$$\int \{g^{-1}(\boldsymbol{\beta}^T \boldsymbol{x} + \boldsymbol{u}^T \boldsymbol{z} + \beta^{(l)}\epsilon + u^{(l)}\epsilon) - g^{-1}(\boldsymbol{\beta}^T \boldsymbol{x} + \boldsymbol{u}^T \boldsymbol{z})\}\ f_u(\boldsymbol{u}; \boldsymbol{\alpha})\ \mathrm{d}\boldsymbol{u} \tag{7}$$

This reduces to $\beta^{(l)}\epsilon$ for the LMM and to $\phi\beta^{(l)}\epsilon$ for the GLMM with bridge distribution.

causal effects

If $X^{(l)}$ is manipulable, for example, treatment, $\beta^{(l)}$ may be interpretable as a causal effect in complete clusters. Let $Y_j(x, \boldsymbol{X}_j^{(-l)})$ be the potential outcome of member j when $X_j^{(l)}$ is manipulated to equal x. We make the following "causal assumptions" (Vansteelandt, 2007). First, $P\{Y_j = Y_j(X_j^{(l)}, \boldsymbol{X}_j^{(-l)})\} = 1$, that is, observed outcome equals outcome that would be seen if $X^{(l)}$ were set to its observed value. Second, manipulating $X_j^{(l)}$ does not affect $\boldsymbol{X}_j^{(-l)}$ or $\boldsymbol{X}$ or Y values of other members. Third, $\{Y_j(x, \boldsymbol{X}_j^{(-l)}) : x \in \mathcal{X}\} \perp\!\!\!\perp \tilde{\boldsymbol{X}} \mid \boldsymbol{X}_j^{(-l)}$, where $\mathcal{X}$ is set of possible values of $X^{(l)}$. With these assumptions, the conditional expected causal effect $E\{Y_j(x, \boldsymbol{X}_j^{(-l)}) - Y_j(0, \boldsymbol{X}_j^{(-l)}) \mid \boldsymbol{X}_j^{(-l)}, \boldsymbol{u}\}$ of $X_j^{(l)}$ given $\boldsymbol{X}_j^{(-l)}$ and $\boldsymbol{u}$ is $c(x, \boldsymbol{X}_j^{(-l)}, \boldsymbol{u}) = g^{-1}\{(\boldsymbol{\beta}^{(-l)})^T \boldsymbol{X}_j^{(-l)} + (\boldsymbol{u}^{(-l)})^T \boldsymbol{z}^{(-l)} + (\beta^{(l)} + u^{(l)})x\} - g^{-1}\{(\boldsymbol{\beta}^{(-l)})^T \boldsymbol{X}_j^{(-l)} + (\boldsymbol{u}^{(-l)})^T \boldsymbol{z}^{(-l)}\}$. For LMMs, $c(x, \boldsymbol{X}^{(-l)}, \boldsymbol{u})$ reduces to $(\beta^{(l)} + u^{(l)})x$. The conditional expected causal effect $E\{Y_j(x, \boldsymbol{X}_j^{(-l)}) - Y_j(0, \boldsymbol{X}_j^{(-l)}) \mid \boldsymbol{X}_j^{(-l)}\}$ of $X^{(l)}$

given $\boldsymbol{X}_j^{(-l)}$ is $c^*(x, \boldsymbol{X}_j^{(-l)}) = \int c(x, \boldsymbol{X}_j^{(-l)}, \boldsymbol{u}) f_u(\boldsymbol{u}; \boldsymbol{\alpha})\, \mathrm{d}\boldsymbol{u}$, which reduces to $\beta^{(l)}x$ for LMMs and to $\phi\beta^{(l)}x$ for GLMMs with bridge distribution.

5.3. *Interpretation of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ in Observed Clusters*

Section 5.2 discussed how $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ in the model defined by (2)–(4) or (3)–(5) describe the Y-$\boldsymbol{X}$ association in *complete* clusters. Now we discuss how the *same* $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ relate to associations in *observed* clusters.

within-cluster fixed effects

When (6) holds and $X^{(l)}$ is cluster varying with fixed effect, $\beta^{(l)}$ is not only the within-complete-cluster effect of $X^{(l)}$, it is also the within-observed-cluster effect, which is the same in all observed clusters of size $N \geq 2$. That is, if two members of the same *observed* cluster of size $N \geq 2$ have $\boldsymbol{X}$ values that differ only by $\boldsymbol{I}^{(l)}\epsilon$ for some ϵ, then their expected values (transformed by link function g in the case of the GLMM) of Y differ by $\beta^{(l)}\epsilon$.

When considering within-observed-cluster effects of covariates with random effects, between-observed-cluster effects and causal effects, we find it convenient to introduce the concept of the LMM/GLMM given by equations (2)–(4) or (3)–(5) "describing observed random subclusters." For a cluster with $N \geq n$, let H_n denote the set of indices of a simple random sample of size n from the N observed members, and let $\tilde{\boldsymbol{X}}_{(H_n)} = \{\boldsymbol{X}_j : j \in H_n\}$. Note that H_1 is the same as what we denoted in Section 3 by H. We say "the LMM given by (2)–(4) describes observed random subclusters of size n from observed clusters of size $\geq n$" (or, more concisely, "the LMM describes observed random subclusters of size n") if

$$Y_j \mid \tilde{\boldsymbol{X}}_{(H_n)}, \boldsymbol{u}, N \geq n \;\sim\; N(\boldsymbol{\beta}^T\boldsymbol{X}_j + \boldsymbol{u}^T\boldsymbol{Z}_j,\ \sigma^2) \quad \forall j \in H_n \tag{8}$$

$$\boldsymbol{u} \perp\!\!\!\perp \tilde{\boldsymbol{X}}_{(H_n)} \;\mid\; N \geq n \tag{9}$$

$$\boldsymbol{u} \mid N \geq n \;\sim\; f_u(\boldsymbol{u}; \boldsymbol{\alpha}) \tag{10}$$

$$\{Y_j : j \in H_n\} \text{ are independent given } \tilde{\boldsymbol{X}}_{(H_n)}, \boldsymbol{u}, N \geq n \tag{11}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ in (8)–(11) are the *same* parameters (i.e., have the same values) as in equations (2)–(4). Similarly, "the GLMM (given by (3)–(5)) describes observed random subclusters of size n" if

$$E(Y_j \mid \tilde{\boldsymbol{X}}_{(H_n)}, \boldsymbol{u}) = g^{-1}(\boldsymbol{\beta}^T\boldsymbol{X}_j + \boldsymbol{u}^T\boldsymbol{Z}_j) \quad \forall j \in H_n \tag{12}$$

and (9)–(11) hold. If (8)–(11) or (9)–(12) hold for one or more values of n, we have a basis for interpreting the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ obtained by fitting the LMM/GLMM given by (2)–(5) (which describes *complete* clusters) in terms of effects in *observed* clusters. We give these interpretations below. Later (Proposition 2) we give sufficient conditions for the LMM/GLMM to describe observed random subclusters of size n and (Section 5.4) show what can happen when these conditions are not satisfied. Note that the statement that LMM/GLMM describes random subclusters of size n is a statement about the Y-$\boldsymbol{X}$ relation only in observed members of clusters with $N \geq n$; the association in missing members or in clusters with $N < n$ is not relevant. We shall focus on $n = 1$ when discussing between-cluster effects, but for within-cluster effects we need $n \geq 2$, because within-cluster comparisons only make sense in clusters with at least two members. In most realistic settings, if the sufficient conditions (Proposition 2) are satisfied for n, they are also satisfied for $n^* < n$.

within-cluster random effects

If the LMM/GLMM describes observed random subclusters of size n (with $n \geq 2$) and $X^{(l)}$ is a cluster-varying covariate with random effect, then $\beta^{(l)}$ and $\mathrm{Var}(u^{(l)})$ are the mean and variance of the within-observed-cluster effect of $X^{(l)}$. That is, if an observed cluster is randomly sampled conditional on $N \geq n$ and on n members randomly chosen from it having $\boldsymbol{X}$ values that differ only in $X^{(l)}$, then the expected values (transformed by link function g) of Y of any pair of these n members differ by $(\beta^{(l)} + u^{(l)})\epsilon$, where ϵ is the difference between their $X^{(l)}$ values, and the distribution of $u^{(l)}$ is given by $\boldsymbol{u} \sim f_u(\boldsymbol{u}; \boldsymbol{\alpha})$.

between-cluster effects

If the LMM/GLMM describes observed random subclusters of size $n = 1$, $\boldsymbol{\beta}$ are the between-observed-cluster effects of $\boldsymbol{X}$. That is, if two clusters each with $N \geq 1$ are randomly sampled conditional on $\boldsymbol{X}_H = \boldsymbol{x}$ in one cluster and $\boldsymbol{X}_H = \boldsymbol{x} + \boldsymbol{I}^{(l)}\epsilon$ in the other, then the difference between the expectations of Y_H in the two clusters is

$$\int \{g^{-1}(\boldsymbol{\beta}^T\boldsymbol{x}_H + \boldsymbol{u}^T\boldsymbol{z}_H + \beta^{(l)}\epsilon + u^{(l)}\epsilon) - g^{-1}(\boldsymbol{\beta}^T\boldsymbol{x}_H + \boldsymbol{u}^T\boldsymbol{z}_H)\}\, f_u(\boldsymbol{u}; \boldsymbol{\alpha})\, \mathrm{d}\boldsymbol{u}. \tag{13}$$

Since (13) has the same form as (7), between-cluster effects in observed and complete clusters are equal and $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ describe them both. As with (7), (13) reduces to $\beta^{(l)}\epsilon$ for the LMM. When $X^{(l)}$ has fixed effect, this is true even if $\boldsymbol{u}$ is not independent of N, so (10) is not necessary for $\beta^{(l)}$ to be interpreted as a between-observed-cluster fixed effect in a LMM.

causal effects

Let $X^{(l)}$ be manipulable and the "causal assumptions" of Section 5.2 hold. Let $\tilde{\boldsymbol{X}}^{(-l)} = (\boldsymbol{X}_1^{(-l)}, \ldots, \boldsymbol{X}_M^{(-l)})$ and $\mathcal{Y} = \{Y_j(x_j, \boldsymbol{X}_j^{(-l)}) : j = 1, \ldots, M; x_j \in \mathcal{X}\}$. If the LMM/GLMM describes observed random subclusters of size n ($n \geq 1$) and $P(\boldsymbol{R} \mid \tilde{\boldsymbol{X}}, \mathcal{Y}, \boldsymbol{u}) = P(\boldsymbol{R} \mid \tilde{\boldsymbol{X}}^{(-l)}, \boldsymbol{u})$, then $\beta^{(l)}$ and $\boldsymbol{\alpha}$ describe a causal effect of $X^{(l)}$ in observed random subclusters of size n. That is, the expected causal effect given $\tilde{\boldsymbol{X}}_{(H_n)}$ and $\boldsymbol{u}$ in the members whose indices belong to H_n is equal to $c(\boldsymbol{x}, \boldsymbol{X}_j^{(-l)}, \boldsymbol{u})$ with $\boldsymbol{u} \sim f_u(\boldsymbol{u}; \boldsymbol{\alpha})$, and the expected causal effect given $\tilde{\boldsymbol{X}}_{(H_n)}$ is equal to $c^*(\boldsymbol{x}, \boldsymbol{X}_j^{(-l)})$. For the LMM when $X^{(l)}$ has fixed effect, $c(\boldsymbol{x}, \boldsymbol{X}_j^{(-l)}, \boldsymbol{u})$ reduces to $\beta^{(l)}$ even if (10) does not hold. Note that if $P(\boldsymbol{R} \mid \tilde{\boldsymbol{X}}, \mathcal{Y}, \boldsymbol{u})$ depends on $X_1^{(l)}, \ldots, X_M^{(l)}$, this causal interpretation is problematic because membership of observed clusters may change as $X^{(l)}$ is manipulated, that is, some observed members would not have been observed if their $X^{(l)}$ values had been otherwise, while some missing members would have been observed.

Proposition 2
The LMM/GLMM describes observed random subclusters of size n if (i) $P(\boldsymbol{R} \mid \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}, \boldsymbol{u}) = P(\boldsymbol{R} \mid \boldsymbol{X}_{\text{con}}, \boldsymbol{u})$, where $\boldsymbol{X}_{\text{con}}$ is a cluster-constant subvector of $\boldsymbol{X}$; either (iia) $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_M)$ are exchangeable given M or (iib) $P(\boldsymbol{R} = \boldsymbol{r} \mid \boldsymbol{X}_{\text{con}}, \boldsymbol{u}) = P(\boldsymbol{R} = \boldsymbol{r}' \mid \boldsymbol{X}_{\text{con}}, \boldsymbol{u})$ whenever $\boldsymbol{r}'$ is a permutation of $\boldsymbol{r}$; and (iii) $P(N \geq n \mid \boldsymbol{X}_{\text{con}}, \boldsymbol{u}) = P(N \geq n)$.

Note that (iii) holds if the minimum possible observed cluster size is $\geq n$, but is unlikely to hold otherwise; and if (iii) is replaced by the weaker condition $P(N \geq n \mid \boldsymbol{X}_{\text{con}}, \boldsymbol{u}) = P(N \geq n \mid \boldsymbol{u})$, then (8), (9) and (11) still hold, but (10) may not.

5.4. *Situations Where Complete- and Observed-Cluster Effects Differ*

With the exceptions mentioned above (i.e., within-cluster fixed effects, and between-cluster and causal fixed effects in LMMs when (9) holds), $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ may not be so interpretable in terms of effects in observed clusters if (9) or (10) do not hold.

Suppose that (10) with $n = 1$ does not hold and $X^{(l)}$ has a random effect. The between-observed-cluster effect of $X^{(l)}$ is given by (13) with $f_u(\boldsymbol{u}; \boldsymbol{\alpha})$ replaced by $f_u(\boldsymbol{u} \mid N \geq 1; \boldsymbol{\alpha})$. In particular, it does not reduce to $\beta^{(l)}\epsilon$ for the LMM unless $E(\boldsymbol{u} \mid N \geq 1) = 0$. Similarly, the observed-cluster causal effect $\int c(x, \boldsymbol{X}^{(-l)}, \boldsymbol{u}) f_u(\boldsymbol{u} \mid N \geq 1; \boldsymbol{\alpha})\, \mathrm{d}\boldsymbol{u}$ is, in general, not the same as the complete-cluster causal effect $c^*(\boldsymbol{x}, \boldsymbol{X}^{(-l)})$; and the within-observed-cluster effect will not, in general, have mean $\beta^{(l)}$ and variance implied by $f_u(\boldsymbol{u}; \boldsymbol{\alpha})$.

In the following example, (9) does not hold for $n = 2$. Suppose clusters are old people in a cohort study of cognitive function Y. A LMM is used, with a random effect for time because rate of cognitive decline varies between people. Assume a fixed effect for the intercept. The only missing data are due to death: $R_{ij} = 1$ if person i is alive at time j; $R_{ij} = 0$ if dead. So, $\boldsymbol{X}_j = (1, j)^T$, $\boldsymbol{Z} = X^{(2)}$, $\boldsymbol{u} = u^{(2)}$ and missingness is monotone. Suppose people with more rapid decline (more negative $u^{(2)}$) tend to die earlier. The within-complete-cluster effect of $X^{(2)}$ has mean $\beta^{(2)}$ and variance $\mathrm{Var}(u^{(2)})$. The mean and variance of the within-observed-cluster effect are functions of $X^{(2)}$: they both diminish as $X^{(2)}$ increases. This is because the subsample still alive at later times is enriched for high $u^{(2)}$. In this setting "complete-cluster" inference has been called inference for a hypothetical immortal cohort, and it has been suggested that "observed-cluster" inference (describing the population still alive at each timepoint) is of more interest (Dufouil et al., 2004). See Section 6 and Web Appendix F for examples of between-cluster or causal effects differing in complete and observed clusters.

5.5. *Observed Clusters Without Complete Clusters*

Dunson et al. (2003), Chen et al. (2011) and Gueorguieva (2005) wanted observed-cluster inference when "complete clusters" do not exist, for example, toxicology experiments where clusters are litters. Dunson et al. and Gueorguieva assumed cluster-constant $\boldsymbol{X}$, $P(N \geq 1) = 1$ and $P(N \mid \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}, \boldsymbol{u}) = P(N \mid \boldsymbol{X}, \boldsymbol{u})$. Chen et al. assumed $\boldsymbol{X}$ was cluster constant or a function of j (e.g., $\boldsymbol{X}_j = (1, j)^T$), $P(N \mid \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}, \boldsymbol{u}) = P(N \mid \boldsymbol{u})$ and $\boldsymbol{Z} = 1$. It can be seen that these methods give complete-cluster inference for a hypothetical population of complete clusters in which $M_i = \max(N_1, \ldots, N_K)$ and from which the population of observed clusters would be generated by applying monotone missingness mechanism $P(N \mid \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}, \boldsymbol{u})$. However, they do not *only* provide complete-cluster inference. When, as in Dunson et al. and Gueorguieva, $\boldsymbol{X}$ is cluster constant and $P(N \geq 1) = 1$, conditions (i), (iia) and (iii) of Proposition 2 hold with $n = 1$, so $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are also between-cluster or causal effects in observed clusters. When, as in Chen et al., $\boldsymbol{X}$ is cluster varying, $\boldsymbol{Z} = 1$ and $P(N \mid \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}, \boldsymbol{u}) = P(N \mid \boldsymbol{u})$, non-intercept elements of $\boldsymbol{\beta}$ are within-observed-cluster effects.

6. Example: Psoriatic Arthritis

This example shows a model that ostensibly describes observed clusters but some of whose parameters relate only to a population of complete clusters with no obvious meaning. Husted et al. (2007) analyzed a cohort of 382 psoriatic arthritis (PsA) patients. Physical function was measured by the health assessment questionnaire score (HAQ). HAQ is semi-continuous: it is zero (no disability) with positive probability and otherwise varies continuously up to 3 (severe disability). 31% of the 2107 HAQ scores were zero. They separately modeled $P(\text{HAQ} > 0)$ (the "binary-part") and HAQ given $\text{HAQ} > 0$ (the "continuous-part"), using, respectively, logistic regression with random intercept $v^{(1)}$ and linear regression with random intercept $u^{(1)}$. Both parts had the same covariates (sex, time since onset, etc.), and all covariates had fixed effects. Among the conclusions was that being female predicted higher HAQ when $\text{HAQ} > 0$, adjusting for other covariates.

Here, clusters are patients and "observed cluster" means a patient's set of non-zero scores. Su et al. (2009) noted that estimates for the continuous part might be biased because separate modeling of binary and continuous parts did not account for ICS caused by the model for the binary part determining the observed cluster size in the continuous part. So, they modified Husted et al.'s model by replacing $v^{(1)}$ by $\psi u^{(1)}$, where ψ is unknown. They called this shared random-effect model the "latent-process model" (SAS code provided in Web Appendix G). They also used a correlated random effects model, but results were similar.

In the original (misspecified) model of Husted et al., the estimated sex effect in the continuous part was 0.181 (SE 0.051). In the latent-process model, it was 0.246 (SE 0.052) (Table 1). We focus on the meaning of this latter estimate. We emphasize there is nothing intrinsically wrong with the latent-process model. It can validly be used to predict HAQ. What is important is not to misinterpret the parameters in the continuous part. As this is an LMM and sex is cluster-constant with fixed effect, the estimated sex effect, 0.246, describes the between-cluster effect in "complete clusters," that is, in a hypothetical world in which all scores are somehow non-zero. The meaning and scientific interest of this hypothetical world, analogous to the world of "immortal cohorts," is unclear.

Su et al. (2009) do not comment on the meaning of their estimated sex effect, but suppose one wished to interpret it as an effect in observed clusters, as done in Husted et al. (2007). As all the covariates have fixed effects, estimates for cluster-varying covariates can be interpreted unproblematically as within-cluster effects in complete or observed clusters. However, sex is cluster-constant. To illustrate the problem with

Table 1
Estimates for latent process model and marginal model fitted to psoriatic arthritis data

	latent process model				marginal model	
	binary part		continuous part			
Parameter	estim	SE	estim	SE	estim	SE
Intercept	−0.9909	0.3556	0.1748	0.0555	0.263	0.0669
Age at onset	0.6392	0.1538	0.0984	0.0250	0.115	0.0267
Female	2.0037	0.3149	0.2461	0.0523	0.100	0.0580
PsA disease duration	0.0166	0.0220	0.0044	0.0032	0.004	0.0041
Actively inflamed joints	0.1380	0.0465	0.0243	0.0027	0.023	0.0045
Clinically deformed joints	0.0179	0.0238	0.0051	0.0031	0.007	0.0037
PASI score	0.1543	0.1017	0.0257	0.0134	−0.005	0.0237
Morning stiffness	1.5691	0.2018	0.1620	0.0262	0.273	0.0444
ESR	0.2971	0.1103	0.0374	0.0126	0.065	0.0232
Medication:						
NSAIDs	0.2960	0.2439	−0.0181	0.0280	−0.235	0.0467
DMARDs	0.3138	0.2197	0.0226	0.0272	0.003	0.0442
steroids	0.9927	0.4355	0.0481	0.0441	0.049	0.0553
Actively inflamed joints × disease duration	0.0003	0.0031	−0.0005	0.0002	0.0000	0.0002
Clinically deformed joints × disease duration	0.0018	0.0011	0.0003	0.0001	0.0000	0.0001
Var(u)	4.2641	0.9001				
ψ			0.2074	0.0210		
σ^2			0.0779	0.0039		

interpreting the estimated sex effect, 0.246, as a between-cluster effect in observed clusters, we obtained the empirical Bayes estimate of each patient's random intercept $u^{(1)}$. While the means of $u^{(1)}$ were 0.005 and 0.016 for men and women, respectively, means of $u^{(1)}$ for observations on men and women when HAQ > 0 were 0.165 and 0.043. This difference arises because in the binary part of the model the estimated sex effect is 2.00 (SE 0.31), meaning that a woman was more likely to have HAQ > 0 than a man with the same values of other covariates. So, if we compare a man and woman who both have HAQ > 0 and have the same time since onset and other covariate values, we expect the woman's HAQ to be not 0.246 greater but only $0.246 - (0.165 - 0.043) = 0.124$ greater. Note that in Su et al.'s model, none of the conditions of Proposition 2 hold for any n.

We also used IEE to fit a model for $e_A(\boldsymbol{x})$, the conditional mean of HAQ given sex, time since onset, etc. and HAQ > 0 (Table 1). The estimated sex effect is 0.100 (SE 0.031), which is close to the effect, 0.124, worked out above using empirical Bayes estimates.

In conclusion, the estimated sex effect in the continuous part of the latent-process model (and correlated random-effects model) describes the association between sex and HAQ in a hypothetical population of little scientific interest; for this dataset it overstates the size of the effect in the population of scientific interest. In further work, Su et al. (2011) found an association of genotype HLA-B27 with HAQ when HAQ > 0. The same interpretation problem applies here: this association refers to the hypothetical "complete" clusters.

7. Discussion

We have shown that shared random-effect models do not always describe observed clusters, except for cluster-varying covariates with fixed effects or under the conditions of Proposition 2. The models of Dunson et al. (2003), Gueorguieva (2005) and Chen et al. (2011) are unnecessarily restrictive. They assume either cluster-constant $\boldsymbol{X}$ or that N does not depend on $\boldsymbol{X}$. Proposition 2 shows $\boldsymbol{X}$ can be cluster varying if N depends only on cluster-constant elements. The assumptions required do, however, remain restrictive. WIEE relate to IPW for missing data. DWGEE2 give inference for a hypothetical population of complete clusters that is, in general, neither unique nor of scientific interest.

For binary Y, Li et al. (2011) used a correlated random-intercepts model with bridge distributions, so that $e_C(\boldsymbol{x}) = \phi\boldsymbol{\beta x}$. For a single binary X, they compared the log odds ratios in complete and observed clusters. They found the difference was small when the variance of the random intercepts or the correlation between them was small. However, when random-intercept variances and/or correlation are small, cluster size is only weakly informative; when size is strongly informative, inferences for complete and observed clusters will differ more. We replicated Li et al's study and found the two log odds ratios could differ by as much as 25% when $\phi = 0.6$, and 56% when $\phi = 0.2$ (see Web Appendix H).

We have assumed Y and $\boldsymbol{X}$ are observed in all members for which we wish to make inference. Dufouil et al. (2004) and Shardell and Miller (2008) give methods for when this is not so.

Having illustrated the danger of misinterpreting estimates, we recommend careful thought about which inference is of scientific interest and which analysis method will give it.

8. Supplementary Materials

Web Appendices referenced in Sections 3–7 are available with this paper at the Biometrics website on Wiley Online Library.

Acknowledgements

SRS is funded by MRC grants U1052 60558 and MC_US_A030_0015, AJC and MP by MRC grant G0600657. We thank Brian Tom for helpful comments on a draft of this article, and Li Su for providing the PsA data and advising on the use of SAS.

References

Albert P. and Follmann D. (2009). Shared-parameter models. In *Longitudinal Data Analysis*, G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs (eds). Chapman & Hall/CRC, Boca Raton, Florida, pp. 433–452.

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

Chen, Z., Zhang, B., and Albert, P. (2011). A joint modeling approach to data with informative cluster size: Robustness to the cluster size model. *Statistics in Medicine* **30**, 1825–1836.

Dufouil, C., Brayne, C., and Clayton, D. (2004). Analysis of longitudinal studies with death and drop-out: A case study. *Statistics in Medicine* **23**, 2215–26.

Dunson, D., Chen, Z., and Harry, J. (2003). A Bayesian approach for joint modeling of cluster size and subunit-specific outcomes. *Biometrics* **59**, 521–530.

Gueorguieva, R. (2005). Comments about joint modelling of cluster size and binary and continuous subunit-specific outcomes. *Biometrics* **61**, 862–867.

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

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

Husted, J., Tom, B., Farewell, V., Schentag, C., and Gladman, D. (2007). A longitudinal study of the effect of disease activity and clinical damage on physical function over the course of psoriatic arthritis: Does the effect change over time? *Arthritis and Rheumatism* **56**, 840–849.

Kurland, B., Johnson, L., Egleston, B., and Diehr, P. (2009). Longitudinal data with follow-up truncated by death: Match the analysis method to research aims. *Statistical Science* **24**, 211–222.

Li, X., Bandyopadhyay, D., Lipsitz, S., and Sinha, D. (2011). Likelihood methods for binary responses of present components in a cluster. *Biometrics* **67**, 629–635.

Neuhaus, J. and McCulloch, C. (2011). Estimation of covariate effects in generalized linear mixed models with informative cluster sizes. *Biometrika* **98**, 147–162.

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

Seaman, S., Galati, J., Jackson, D., and Carlin, J. (2013). What is meant by 'missing at random'? *Statistical Science* **28**, 257–268.

Shardell, M. and Miller, R. (2008). Weighted estimated equations for longitudinal studies with death and non-monotone missing time-dependent covariates and outcomes. *Statistics in Medicine* **27**, 1008–25.

Su, L., Tom, B., and Farewell, V. (2009). Bias in 2-part mixed models for longitudinal semicontinuous data. *Biostatistics* **10**, 374–389.

Su, L., Tom, B., and Farewell, V. (2011). A likelihood-based two-part marginal model for longitudinal semi-continuous data. *Statistical Methods in Medical Research*, DOI: 10.1177/0962280211414620. Available at: http://smm.sagepub.com/content/early/2011/08/25/0962280211414620.abstract.

Vansteelandt, S. (2007). On confounding, prediction and efficiency in the analysis of longitudinal and cross-sectional clustered data. *Scandinavian Journal of Statistics* **34**, 478–498.

Wang, M., Kong, M., and Datta, S. (2011). Inference for marginal linear models for clustered longitudinal data with potentially informative cluster sizes. *Statistical Methods in Medical Research* **20**, 347–367.

Wang, Z. and Louis, T. (2003). Matching conditional and marginal shapes in binary random intercept models with a bridge distribution function. *Biometrika* **90**, 765–775.

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

Received April 2013. *Revised November* 2013.
Accepted January 2014.