Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–20

DOI: 10.1177/0962280220986193
journals.sagepub.com/home/smm
SAGE

Clustered longitudinal data subject to irregular observation

Eleanor M Pullenayegum[1,2], **Catherine Birken**[1,3,4], **Jonathon Maguire**[3,4,5,6] **and The TARGet Kids! Collaboration**[7]

Abstract

Data collected longitudinally as part of usual health care is becoming increasingly available for research, and is often available across several centres. Because the frequency of follow-up is typically determined by the patient's health, the timing of measurements may be related to the outcome of interest. Failure to account for the informative nature of the observation process can result in biased inferences. While methods for accounting for the association between observation frequency and outcome are available, they do not currently account for clustering within centres. We formulate a semi-parametric joint model to include random effects for centres as well as subjects. We also show how inverse-intensity weighted GEEs can be adapted to account for clustering, comparing stratification, frailty models, and covariate adjustment to account for clustering in the observation process. The finite-sample performance of the proposed methods is evaluated through simulation and the methods illustrated using a study of the relationship between outdoor play and air quality in children aged 2–9 living in the Greater Toronto Area.

Keywords

Longitudinal data, inverse-weighting, semi-parametric models, informative observation, clustering

1 Introduction

Longitudinal data frequently feature irregular observation times, where the timing of measurements differs among patients to the extent that no two patients share an observation time. Moreover, the timing and frequency of visits may be correlated with the outcome of interest. For example, in a study of growth in the newborn period using data collected as part of usual care, infants who are slow to regain their birthweight will likely be followed up more frequently. Consequently, if the frequency of visits is ignored, one could expect to underestimate the mean rate of weight gain in the newborn period. Statistical methods for handling informative visit processes are well established but currently do not extend to clustered longitudinal data.

Clustered longitudinal data occur when patients are clustered within families, physicians, hospitals, geographic regions, family health teams, or any other variable for which outcomes are likely to be more similar within clusters than between clusters. As administrative health data become increasingly available, so too does the occurrence of clustered longitudinal data subject to irregular follow-up. For example, in a cluster-randomised registry trial,

[1]Child Health Evaluative Sciences, Hospital for Sick Children, Toronto, ON, Canada
[2]Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada
[3]Department of Paediatrics, University of Toronto, Toronto, ON, Canada
[4]Department of Paediatrics, St Michael's Hospital, Toronto, ON, Canada
[5]Department of Nutritional Sciences, University of Toronto, Toronto, ON, Canada
[6]Li Ka Shing Knowledge Institute, Unity Health Toronto, Toronto, ON, Canada
[7]Research Consortium

Corresponding author:
Eleanor M Pullenayegum, Child Health Evaluative Sciences, Hospital for Sick Children, 555 University Ave, Toronto M5G 1X8, ON, Canada.
Email: eleanor.pullenayegum@sickkids.ca

hospitals are randomised to one of two or more groups, and data are collected as part of existing registries.[1] This can be a cost-effective way of running a trial; however, if the follow-up schedule is according to usual care, there is a risk that the observation times are informative. Similarly, observational studies in rare diseases may use electronic health records as an efficient way of constructing a multi-centre longitudinal cohort; however, visits are likely to be more often when patients are sick, and moreover, the follow-up schedule may vary across centres. For example, in family practice within Canada there is variation across provinces in whether annual physical examinations can be billed.[2]

Several methods to handle non-clustered longitudinal data subject to irregular observation have been proposed. These methods fall into two broad classes: methods based on inverse-intensity weighting assume that the visit and outcome processes are independent given the observed data,[3,4] while semi-parametric joint models assume independence given the covariates in the outcome model and a set of random effects.[5–10] The appropriate analytic approach to this type of data depends on what assumptions are appropriate for describing the dependence between the visit and outcome processes.[11] In the case where visit and outcome processes are independent given observed covariates, inverse-intensity weighted GEEs are appropriate. In the case where visit and outcome processes are independent given baseline covariates and time-invariant random effects, semi-parametric joint models can be used. Recent work also includes methods using a composite likelihood.[12] However, none of these methods has been explored in the context of clustered longitudinal data.

The purpose of this paper is to extend inverse-intensity weighted GEEs to handle clustering, and to develop a semi-parametric joint model that accounts for clustering. The finite-sample performance of the proposed methods is explored using simulation. The methods are illustrated through a study of the relationship between air quality and outdoor play among children aged 2–9 using data from the TARGet Kids! Study.[13]

2 Inverse-intensity weighting

2.1 Notation

Let $Y_{ij}(t)$ be the outcome for subject i in cluster j $(i = 1, \ldots, n_j; j = 1, \ldots, J)$ at time t, and let $N_{ij}(t)$ be a counting process for subject i in cluster j's visit process, with intensity $\lambda_{ij}(t)$. We assume we are interested in the outcome model

$$E(Y_{ij}(t)|X_{ij}(t)) = \mu(X_{ij}(t)\beta) = g^{-1}(X_{ij}(t)\beta) \tag{1}$$

for an observed p-dimensional row vector of covariates $X_{ij}(t)$, a p-dimensional vector β of regression coefficients, and some link function g.

We assume that the covariates X are chosen so as to address the clinical question of interest, and note that there may also be auxiliary covariates that are observed but we choose not to include in X. Suppose also that the intensity model satisfies

$$E(\Delta N_{ij}(t)|Y_{ij}(t), X_{ij}(t), Z_{ij}(t)) = \lambda_{ij}(t; Z_{ij}(t))$$

where $Z_{ij}(t)$ are auxiliary covariates which for notational convenience may include elements of X as well as past observed outcomes. Let $\mathcal{H}_{ij}(t)$ be the observed history of Y_{ij}, X_{ij}, Z_{ij}, N_{ij} up to but not including time t.

2.2 Review of inverse-intensity weighting

Consider for now a single cluster, without loss of generality $j = 1$, for which the visit and outcome processes are conditionally independent given the observed history, i.e. $N_{i1}(t) \perp\!\!\!\perp Y_{i1}(t)|\mathcal{H}_{i1}(t)$. We can then solve the inverse-intensity weighted GEE equations

$$\sum_i \int_0^\tau \frac{\partial\mu(X_{i1}(t)\beta)}{\partial\beta} w_{i1}(t)(Y_{i1}(t) - \mu(X_{i1}(t)\beta))dN_{i1}(t) = 0 \tag{2}$$

with $w_{i1}(t) = s(t)/\lambda_{i1}(t)$ for some stabilising function s.[3] If $\lambda_{ij}(t)$ follows a proportional hazard model and $s(t)$ is taken to be the baseline hazard $\lambda_0(t)$, then we need to estimate only the log hazard ratios to obtain w.[4]

2.3 Inverse-intensity weighting with clustered longitudinal data

We now extend the inverse-intensity weighting approach to consider inference on all J clusters simultaneously. The IIW-GEE equations can be extended by summing over clusters so as to solve

$$\sum_j \sum_i \int_0^\tau \frac{\partial \mu(X_{i1}(t)\beta)}{\partial \beta} w_{ij}(t)(Y_{ij}(t) - \mu(X_{ij}(t)\beta))dN_{ij}(t) = 0$$

with $w_{ij}(t) = s(t)/\lambda_{ij}(t)$. The left-hand side will continue to have mean zero provided that, $\forall$ i and j, the visit and outcome processes are conditionally independent given the observed history, i.e. $N_{ij}(t) \perp\!\!\!\perp Y_{ij}(t) | \mathcal{H}_{ij}(t)$. The model for the visit intensity can account for cluster effects, for example

$$\lambda_{ij}(t) = \lambda_0(t)\exp(Z_{ij}(t)\gamma_j) \quad \text{covariate stratification} \tag{3}$$

$$\lambda_{ij}(t) = \lambda_0(t)\exp(Z_{ij}(t)\gamma + \alpha_j) \quad \text{cluster adjustment} \tag{4}$$

$$\lambda_{ij}(t) = \lambda_{0j}(t)\exp(Z_{ij}(t)\gamma_j) \quad \text{complete stratification} \tag{5}$$

$$\lambda_{ij}(t) = \lambda_{0j}(t)\exp(Z_{ij}(t)\gamma) \quad \text{baseline hazard stratification} \tag{6}$$

$$\lambda_{ij}(t) = \eta_j \lambda_0(t)\exp(Z_{ij}(t)\gamma) \quad \text{frailty} \tag{7}$$

where η_j are unobserved independent frailty random variables and for notational convenience we allow Z to contain elements of X. In cases (3) and (4), the stabilising function $s(t)$ can be taken to be $\lambda_0(t)$ thus avoiding the need to estimate the baseline hazard. This is no longer possible in (5) and (6), since the baseline hazard is stratified by cluster. In equation (7) using IIW-GEEs requires an estimate of the frailty variable η_j, which can be obtained by assuming a distribution for the frailty variables (e.g. Gamma or log-Normal) and using an empirical Bayes estimate.[14]

To use inverse weighting within standard statistical software, the working variance in the GEE must be set to independent. This is because software generally interprets the weights as indicating heteroscedasticity and therefore as inverse variances. This is implemented in the context of a GEE as follows: let $\mathbf{Y}_j$ be the vector of all observed outcomes for cluster j, $\mathbf{X}_j$ be the matrix whose rows correspond to the covariate values of $X_{ij}(t)$ for each element of $\mathbf{Y}$, $\mathbf{m_j} = \mu(\mathbf{X_j}\beta)$, $\mathbf{\Delta}_j$ be the diagonal matrix of weights corresponding to the elements of $\mathbf{Y_j}$, and $\mathbf{R}_j$ be the working variance matrix. When weighting is introduced in standard statistical software, the working variance $\mathbf{R}_j$ is replaced with $\mathbf{\Delta}_j^{-1/2}\mathbf{R}_j\mathbf{\Delta}_j^{-1/2}$ so the GEE equations become

$$\sum_j \frac{\partial \mathbf{m_j}}{\partial \beta} \mathbf{\Delta}_j^{1/2}\mathbf{R}_j^{-1}\mathbf{\Delta}_j^{1/2}(\mathbf{Y}_j - \mathbf{m_j}) = 0$$

whereas we wish to solve

$$\sum_j \frac{\partial \mathbf{m_j}}{\partial \beta} \mathbf{R}_j^{-1}\mathbf{\Delta}_j(\mathbf{Y}_j - \mathbf{m_j}) = 0$$

The two are, however, the same when the working correlation is set to the identity matrix.

2.4 Sandwich variance error corrections with IIW-GEEs

Standard errors for GEEs with two nested levels of clustering can be estimated by aggregating up to the highest level of clustering. Let $U_j(\beta) = \frac{\partial \mathbf{m_j}}{\partial \beta}\mathbf{R_j}^{-1}(\mathbf{Y_j} - \mathbf{m_j})$ be cluster j's contribution to the pseudo-score function. The GEE equations are

$$U(\beta) = \sum_j \mathbf{U_j}(\beta) = \sum_j \frac{\partial \mathbf{m_j}}{\partial \beta} \mathbf{R}_j^{-1}\mathbf{\Delta}_j(\mathbf{Y}_j - \mathbf{m_j}) = 0$$

By a Taylor expansion of $\sum_j \mathbf{U_j}$ about β, we have

$$\hat{\beta} - \beta \approx -\left(\sum_j \frac{\partial \mathbf{U_j}}{\partial \beta}\right)^{-1} \sum_j \mathbf{U_j}(\beta)$$

and so

$$var(\hat{\beta}) \approx \left(\sum_j \frac{\partial \mathbf{U_j}}{\partial \beta}\right)^{-1} var\left(\sum_j \mathbf{U_j}(\beta)\right)\left(\sum_j \frac{\partial \mathbf{U_j}}{\partial \beta}\right)^{-1}$$

The usual sandwich variance estimator approximates $var(\sum_j \mathbf{U_j}(\beta)) = \sum_j var(\mathbf{U_j}(\beta))$ by $\sum_j \mathbf{U_j}(\hat{\beta})\mathbf{U_j}(\hat{\beta})' = \sum_j \frac{\partial \mathbf{m_j}}{\partial \beta}\mathbf{R_j}^{-1}\mathbf{\Delta}_j(\mathbf{Y_j} - \hat{\mathbf{m}}_\mathbf{j})(\mathbf{Y_j} - \hat{\mathbf{m}}_\mathbf{j})'\mathbf{\Delta_j}\mathbf{R_j}^{-1}\frac{\partial \mathbf{m_j}}{\partial \beta'}$, where $\hat{\mathbf{m}}_\mathbf{j} = \mu(\mathbf{X_j}\hat{\beta})$. However, when the number of clusters is small, replacing β with $\hat{\beta}$ underestimates the true variance.[15,16]

A number of corrections have been proposed to address this problem.[16–23] For the purposes of this paper, we focus on those that do not require a common covarince structure across clusters (i.e. we do not consider the Pan,[20] Gosho,[21] Wang and Long[22] or Westgate and Burchett[23] corrections), as this is not plausible with irregular longitudinal data, and on those that are available in standard statistical software while also allowing for weighting. The Kauermann and Carroll (KC),[16] Mancl and deRouen (MdR),[19] Fay and Graubard (FG)[17] and Morel-Bokossa-Neerchal (MBN)[18] are all available in SAS PROC GLIMMIX. Simulation studies have noted that the KC correction performs well, achieving near nominal coverage of 95% confidence intervals when paired with a correction to the degrees of freedom of $\hat{\beta}$, with as few as eight clusters.[24–27] It can be shown mathematically that the MdR correction overestimates variances but has been noted to achieve near nominal coverage in some settings.[25] Since the KC correction often underestimates variances[26] while the MdR correction overestimates, Ford & Westgate suggest using the average of the two, which they found outperformed both the KC and MdR corrections.[27]

While these corrections have been extensively studied, it is not clear how they behave in the presence of weights. The KC correction is motivated by the scenario where observations are independent and homoscedastic, in which case the correction is exact; however, it is difficult to assess theoretically how it will behave when independence does not hold. The FG correction is motivated by assuming that the working variance matrix is within a scale factor of the truth. In the presence of weighting, this is not a reasonable assumption since the working variance matrix must be set to be independent. The MBN correction adds a model-based variance term to the sandwich variance, with the model-based variance calculated under the assumption that the working variance is correct. Previous work has used simulation to assess how these estimators perform when their motivating assumptions are violated[24–27]; however, this work has not considered the impact of weighting.

It is possible to explore the behaviour of the Mancl & deRouen correction theoretically. Let $\mathbf{\Omega} = \left(\sum_j \frac{\partial \mathbf{U_j}}{\partial \beta}\right)^{-1} = (\sum_j \frac{\partial \mathbf{m_j}}{\partial \beta}\mathbf{R_j}^{-1}\mathbf{\Delta_j}\frac{\partial \mathbf{m_j}'}{\partial \beta})^{-1}$ and $\mathbf{H}_j = \frac{\partial \mathbf{m_j}}{\partial \beta}\mathbf{\Omega}\frac{\partial \mathbf{m_j}'}{\partial \beta}\mathbf{R_j}^{-1}\mathbf{\Delta_j}$. Mancl & deRouen note that the variance of $\mathbf{Y}_j - \mu(\mathbf{X_j}\hat{\beta})$ can be written as

$$(I - \mathbf{H_j})var(\mathbf{Y_j})(I - \mathbf{H_j}) + \sum_{k \neq j} \mathbf{G_{jk}} var(\mathbf{Y_k})\mathbf{G_{jk}}' \tag{8}$$

with $\mathbf{G_{jk}} = \frac{\partial \mathbf{m_j}}{\partial \beta}\mathbf{\Omega}\mathbf{X_k}'\mathbf{R_k}^{-1}\mathbf{\Delta_k}$. Under the assumption that the second term is small, $var(\mathbf{Y}_j)$ can be approximated by

$$(I - \mathbf{H_j})^{-1}(\mathbf{Y_j} - \mu(\mathbf{X_j}\hat{\beta}))(\mathbf{Y_j} - \mu(\mathbf{X_j}\hat{\beta}))'(I - \mathbf{H_j})^{-1}$$

The estimator for $var(\hat{\beta})$ is thus

$$\mathbf{\Omega}(\sum_j \frac{\partial \mathbf{m_j}}{\partial \beta}\mathbf{R_j}^{-1}\mathbf{\Delta_j}(I - \mathbf{H}_j)^{-1}(\mathbf{Y_j} - \mu(\mathbf{X_j}\hat{\beta}))(\mathbf{Y_j} - \mu(\mathbf{X_j}\hat{\beta}))'(I - \mathbf{H}_j)^{-1}\mathbf{\Delta_j}\mathbf{R_j}^{-1}\frac{\partial \mathbf{m_j}}{\partial \beta'})\mathbf{\Omega}$$

Moreover, this is an over-estimate since the second term in equation (8) is a sum of positive definite matrices. The elements of $\mathbf{G_{jk}}$ are typically close to zero unless the weights are extreme, and hence the second term will remain small in the presence of weighting by non-extreme weights.

In addition to adjusting the variance estimator itself, authors have suggested using Wald t rather than Wald z tests when computing p-values and confidence intervals. When coupled with a KC or the average of the KC and MdR corrections, the Wald t-test has been shown to outperform the Wald z-test in maintaining nominal type I error rates.[24,27] Since in the case of clustered longitudinal data there are two levels of clustering, we use the "Between-Within" approach of Schluchter and Elashoff[28] to the degrees of freedom. This divides the residual degrees of freedom into within-cluster and between-cluster portions. For covariates at the cluster level, the degrees of freedom are taken to be the number of clusters minus the number of cluster-specific covariates (i.e. covariates that vary at the cluster-level rather than at the individual or time-within-individual level). In the context of GLMMs for cluster-randomised trials, Li et al.[29] showed that this worked well with as few as 10 clusters and with varying cluster sizes.

3 A semi-parametric joint model

Inverse intensity weighting requires conditional independence within clusters, i.e. $N_{ij}(t)\perp\!\!\!\perp Y_{ij}(t)|\mathcal{H}_{ij}(t)$. We now consider a model with shared random effects, where conditional independence within clusters does not hold. In this section, we generalise the Liang semi-parametric joint model[5] to handle cluster-level random effects in addition to subject-level random effects. Moreover, we show how multiple outputation can be used to accommodate time-dependent covariates in the visit intensity model. The usual formulation of the Liang semi-parametric joint model is

$$\begin{aligned} E(Y_i(t)|X_i(t),\eta_{Ysij}) &= \beta_0(t) + X_i(t)\beta + W_i(t)\eta_{Ysi} \\ \lambda_i(t) &= \eta_{Vsi}\lambda_0(t)\exp(Z_i\gamma) \end{aligned}$$

where η_{Ysi} and η_{Vsi} are potentially correlated random effects with mean 0 and 1, respectively, $E(\eta_{Ysi}|\eta_{Vsi}) = \theta(\eta_{Vsi}-1)$ for some θ, $W_i(t)$ is a vector of covariates, and Z_i is a vector of time-invariant auxiliary covariates. In the context of clustered longitudinal data, we will generalise the Liang model to

$$\begin{aligned} E(Y_{ij}(t)|X_{ij}(t),\eta_{Ysij},\eta_{Ycj}) &= \beta_0(t) + X_{ij}(t)\beta + W_{ij}(t)\eta_{Ysij} + V_{ij}(t)\eta_{Ycj} \\ \lambda_{ij}(t) &= \eta_{Vsij}\eta_{Vcj}\lambda_0(t)\exp(Z_{ij}\gamma) \end{aligned} \tag{9}$$

where V_{ij} is a vector of covariates, $\eta_{Ysij},\eta_{Ycj},\eta_{Vsij},\eta_{Vcj}$ are latent variables (random effects), with $E(\eta_{Ysij}) = E(\eta_{Ycj}) = 0$, $E(\eta_{Vsij}) = E(\eta_{Vcj}) = 1$, $E(\eta_{Ysij}|\eta_{Ycj}) = 0$, $E(\eta_{Vsij}|\eta_{Vcj}) = 1$ for identifiability, with $\eta_{Ycj}\perp\!\!\!\perp\eta_{Vsij}$, $\eta_{Vcj}\perp\!\!\!\perp\eta_{Ysij}$, so that dependence in the visit processes and outcome processes within clusters is due to observed auxiliary covariates and subject-level latent variables only, and with $E(\eta_{Ysij}|\eta_{Vsij}) = \theta(\eta_{Vsij}-1)$, $E(\eta_{Ycj}|\eta_{Vcj}) = \phi(\eta_{Vcj}-1)$ for some θ,ϕ.

3.1 Estimating equation

We now develop estimators for β in equation (9) by setting up a zero-mean estimating function based on the residuals of the outcome process. Extending the approach of Liang et al.,[5] define

$$\begin{aligned} M_{ij}(t;\beta,\theta,\phi,\mathcal{A},\Lambda_0,B^s_{ij},B^c_j) = &\int_0^t (Y_{ij}(s) - X_{ij}(s)\beta - B^s_{ij}(s)\theta + B^c_j(s)\phi))dN_{ij}(s) \\ &- \int_0^t I(C_{ij}\geq s)m_{ij}\frac{d\mathcal{A}(s)}{\Lambda_0(C_{ij})} \end{aligned} \tag{10}$$

where $\mathcal{A}(t) = \int_0^t \beta_0(s)d\Lambda_0(s)$, $B^s_{ij}(t) = W_{ij}(t)E(\eta_{Vsij}-1|m_{ij},C_{ij},i=1,\ldots,N_j)$ and $B^c_j(t) = V_j(t)E(\eta_{Vcj}-1|m_{ij},C_{ij},$ $i=1,\ldots,N_j)$, and note that $E(dM_{ij}(t)) = 0$. Solving $\sum_{ij}dM_{ij}(t) = 0$ yields an estimate for $d\mathcal{A}(t)$, which can be plugged into equation (10) in place of $d\mathcal{A}(s)$ to yield the following estimating equation for β,θ and ϕ

$$\sum_{i,j}\int_0^\tau \begin{pmatrix} X_{ij}(t)-\bar{X}(t) \\ \hat{B}^s_{ij}(t)-\bar{B}^s(t) \\ \hat{B}^c_{ij}(t)-\bar{B}^c(t) \end{pmatrix} (Y_{ij}(t)-X_{ij}\beta-\hat{B}^s_{ij}(t)\theta-\hat{B}^c_j(t)\phi)dN_{ij}(t)=0$$

where $\hat{B}^s_{ij}$, $\hat{B}^c_j$ are estimates of B^s_{ij}, B^c_j, respectively

$$\bar{X}(t)=\frac{\sum_{ij}I(C_{ij}\geq t)X_{ij}(t)m_{ij}/\hat{\Lambda}_0(C_{ij})}{\sum_{ij}I(C_{ij}\geq t)m_{ij}/\hat{\Lambda}_0(C_{ij})}$$
$$\bar{B}^s(t)=\frac{\sum_{ij}I(C_{ij}\geq t)\hat{B}^s_{ij}(t)m_{ij}/\hat{\Lambda}_0(C_{ij})}{\sum_{ij}I(C_{ij}\geq t)m_{ij}/\hat{\Lambda}_0(C_{ij})}$$
$$\bar{B}^c(t)=\frac{\sum_{ij}I(C_{ij}\geq t)\hat{B}^c_j(t)m_{ij}/\hat{\Lambda}_0(C_{ij})}{\sum_{ij}I(C_{ij}\geq t)m_{ij}/\hat{\Lambda}_0(C_{ij})}$$

and $\hat{\gamma}$, $\hat{\Lambda}_0(t)$ are given by solving

$$\sum_{ij}\int_0^\tau (Z_{ij}-\bar{Z}(t,\gamma))dN_{ij}(t)=0$$
$$\hat{\Lambda}_0(t;\hat{\gamma})=\sum_{ij}\int_0^\tau \frac{dN_{ij}(s)}{\sum_{ij}I((C_{ij}>s)\exp(Z_{ij}\gamma)} \text{ where } \bar{Z}(t;\bar{\gamma})=\frac{\sum_{ij}I(C_{ij}\geq t)Z_{ij}(t)\exp(Z_{ij}\gamma)}{\sum_{ij}I(C_{ij}\geq t)\exp(Z_{ij}\gamma)}$$

The conditional expectations $E(\eta_{Vsij}|m_{ij},C_{ij},i=1,\ldots,n_j)$ and $E(\eta_{Vcj}|m_{ij},C_{ij},i=1,\ldots,n_j)$ needed to calculate $\hat{B}^s_{ij}$ and $\hat{B}^c_j$ are given in Appendix 1.

3.2 Multiple outputation

The Liang model requires the covariates in the visit process model to be time-invariant. In many cases, a poor outcome or test results at a visit will precipitate the next visit occurring sooner, making the assumption of time-invariant auxiliary covariates untenable. Because the estimation approach relies on a constant visit rate within subjects, it is not straightforward to generalise to a time-dependent auxiliary covariate.[30] However, multiple outputation can be used to fit the more general model.

Multiple outputation was originally proposed by Hoffman et al.[31] and Follmann et al.[32] to handle informative cluster sizes, and recently adapted for use with irregular longitudinal data.[30] It relies on randomly discarding observations in such a way that the visit process in the resulting thinned dataset does not depend on the auxiliary covariates. This procedure is then repeated multiple times and the results from analysing each thinned dataset are combined so as to use all the data. When combined with a GEE analysis, it has been shown that multiple outputation is asymptotically identical to weighting. As it is computationally more demanding than weighting, multiple outputation is thus primarily helpful for analyses for which weighting is not possible.

The clustered Liang model is one such model in which weighting when the intensity contains time-dependent covariates is not possible. Specifically, $E(dM_{ij}(t))\neq 0$ when the covariates in the intensity model are time-dependent. Just as multiple outputation can be used with the Liang model to allow for time-dependent covariates in the intensity model, so too can multiple outputation be used with the clustered Liang model. Specifically, suppose we wish to fit the model

$$E(Y_{ij}(t)|X_{ij}(t),\eta_{Ysij},\eta_{Ycj})=\beta_0(t)+X_{ij}(t)\beta+W_{ij}(t)\eta_{Ysij}+V_{ij}(t)\eta_{Ycj}$$
$$\lambda_{ij}(t)=\eta_{Vsij}\eta_{Vcj}\lambda_0(t)\exp(Z_{ij}(t)\gamma)$$

but cannot use the clustered Liang model on the full dataset due to the time-dependent auxiliary covariate $Z_{ij}(t)$. Instead, a nested frailty model can be used to estimate γ, then multiple outputation, retaining observations with probability proportional to $\exp(-Z_{ij}(t)\gamma)$, used so as to produce M thinned datasets with

$$\begin{aligned}&E(Y_{ij}(t)|X_{ij}(t), \eta_{Ysij}, \eta_{Ycj}) = \beta_0(t) + X_{ij}(t)\beta + W_{ij}(t)\eta_{Ysij} + V_{ij}(t)\eta_{Ycj}\\ &\lambda_{ij}(t) = \eta_{Vsij}\eta_{Vcj}\lambda_0(t)\end{aligned}$$

The clustered Liang model can then be used to analyse each thinned dataset so as to obtain estimates $\hat{\beta}^{(1)}, \ldots, \hat{\beta}^{(M)}$ and the mean of the results used to estimate β. Note that while the multiple outputation procedure includes an estimate of the variance, this requires an estimate of the variance of each $\hat{\beta}^{(m)}, m = 1, \ldots, M$, which the Liang procedure does not provide. Standard errors must thus be estimated by bootstrapping.

4 Simulation

4.1 Inverse-intensity weighting

We adapt the simulation described by Lin et al.[3] to investigate the finite-sample performance of these methods. The Lin simulation model was designed to mimic data from the Housing and Urban Development – Veterans Affairs Supported Housing (HUD-VASH) randomised trial.[33] We add cluster-level random effects that are correlated between the visit and outcome processes. The simulation set-up is as follows:

Data were generated for 1432 days. The multivariate binary outcome variable $Y_{ij}(t)$ for subject i ($i = 1, \ldots, n_j$) in cluster j ($j = 1, \ldots, J$) indicating whether the subject was homeless at time t, $t = 0, \ldots, 1431$ was given by

$$\begin{aligned}&\text{logit}(E(Y_{ij}(t)|X_j, \eta_{Ysij}, \eta_{Ycj}, \eta_{YXj}))\\ &= -0.5 - 0.075t/30 + (0.001/900)t^2 - 2.27X_j + \eta_{Ysij} + \eta_{Ycj} + \eta_{YXj}X_j\end{aligned}$$

The coefficient of the randomised treatment group X_j was chosen so that the AUC for the treatment group $X = 1$ was half that of the control group $X = 0$. Half the clusters were randomly assigned to be in the intervention group and the remainder were assigned to the control group.

The visit process was simulated using the method of Daley et al.,[34] with intensity function

$$\lambda_{ij}(t) = 0.005\eta_{Vsij}\eta_{VXj}^{X_j}\eta_{Vsij}\exp(-0.0003t + Y_{ij}^{last}(t) - 0.2X_j)$$

for $t = 1, \ldots, 1431$ days, where $Y_{ij}^{last}(t)$ is the last observed value of Y_{ij} prior to t. The coefficient of X_j was chosen so that the difference in visit rate between intervention and control group mimicked that in the HUD-VASH trial, i.e. a difference of -0.2 in log intensity for the intervention vs. the control.

We take $\eta_{Ysij} \sim N(0, 3.6^2)$, $\eta_{Vsij} \sim \Gamma(10, 10)$, $\eta_{Vcj} \sim \Gamma(5, 5)$, $\eta_{VXj} \sim \Gamma(5, 5)$, $\eta_{Ycj}|\eta_{Vcj} \sim N(\eta_{Vcj} - 1, 0.5^2)$, and $\eta_{YXj}|\eta_{VXj} \sim N(\eta_{VXj} - 1, 0.5^2)$. The variance of 3.6^2 on the subject-level random effect in the outcome model creates a within-subject correlation of 0.6 in the outcomes as used in Lin et al.[3] The effect of the subject-level Gamma frailties falls between a 50% decrease in event rate and a 70% increase with 95% probability, while 95% of the effects of the cluster-level Gamma frailties fall between a 60% decrease and a doubling of event rate. The 2.5th percentile of distribution of the cluster-level random effects represents a 70% reduction in odds of the outcome, while the 97.5th percentile represents a four-fold increase in the odds.

The target of inference was the difference between groups in the mean of the total number of days homeless over the 1432 days, i.e. the between-group difference in the area under the curve of $E(Y_{ij}(t))$ vs. t. Since this is a randomised trial setting, the treatment effect has a causal interpretation. By numerical integration, this is -249.343.

The AUC was computed using inverse-intensity weighting both with and without accounting for clustering. The GEEs regressed Y onto a quartic polynomial of time with an identity link function. The GEEs were fit in SAS using PROC GLIMMIX with working independence, using maximum pseudo-likelihood. Since the link function is linear, the pseudo-likelihood equations are identical to the usual GEE equations. The intensity weights were calculated using (i) a Cox model with clusters as fixed effects, (ii) a frailty model with cluster-level frailties, (iii) a stratified Cox model, and, for comparison, (iv) a Cox model ignoring clustering. We considered Wald-z and Wald-t based confidence intervals, and used the KC, MdR, MBN and the average of the KC and MdR corrections. We initially intended to use the FG correction as well; however, this led to extreme estimated variances (over 10-fold overestimation). This degree of over-estimation should not have occurred given that the estimator contains a parameter to control the extent of the correction; this was set to the default in SAS, which should have limited the standard error inflation to a maximum of two-fold. SAS support acknowledged that there is a problem with the FG estimator in PROC GLIMMIX. In evaluating the sandwich variance corrections, we considered a sample size of 100. When evaluating the choices of intensity weights, we considered sample sizes of 100 and 500. For each sample size, we considered 4, 10 and 20 equally sized clusters. One thousand datasets were generated for each scenario.

4.2 Semi-parametric joint modelling

We then considered models with correlated random effects at the subject level, i.e. for which the assumption of conditional independence within clusters given covariates in the visit process model does not hold. To begin, we adopted an intensity model with no time-dependent covariates. The previous model was thus adapted as follows:

The visit process had intensity

$$\lambda_{ij}(t) = 0.005\eta_{Vsij}\eta_{Vcj}\eta_{VXjX}^{X_j}\exp(-0.0003t - 0.2X_j)$$

where $\eta_{Vsij}\sim\Gamma(10,10), \eta_{Vcj}\sim\Gamma(5,5), \eta_{VXj}\sim\Gamma(5,5)$ and X_j is a binary treatment group indicator as before.

For subject i in cluster j at time t, the probability of homelessness conditional on random effects was as before, but with $\eta_{Ysij}|\eta_{Vsij}\sim N(\eta_{Vsij} - 1, 3.6^2)$ instead of $\eta_{Y\ sij}$ & $\eta_{V\ sij}$ being independent.

The target of inference was the difference in AUC for the treatment group compared to the control group. This was computed by using the semi-parametric clustered Liang model to regress Y onto X interacted with a quartic polynomial of time with random effects for time, time^2, time^3, time^4, X, and the interaction terms between X and the powers of time. These covariates were also used for both subject-level and cluster-level random effects. For comparison, we also computed the Liang estimator that ignores clustering.

To explore the potential of multiple outputation to handle time-dependent covariates in the visit process model, we also simulated data with the last observed outcome in the visit intensity model

$$\lambda_{ij}(t) = 0.005\eta_{Vsij}\eta_{Vcj}\eta_{VXj}^{X_j}\exp(-0.0003t + Y_{ij}^{last}(t) - 0.2X_j),$$

while using the same outcome model.

The visit intensity model was estimated using a proportional hazard model with subject-level frailties and fixed effects for both last observed outcome and cluster. Multiple outputation was used with 20 outputations, and with probability of retention for subject i in cluster j at time t proportional to $\exp(-\hat{\gamma}_1 Y_{ij}^{last}(t))$, with $\hat{\gamma}$ the estimated regression coefficient of the last observed outcome. The resulting thinned datasets were analysed using the clustered Liang approach as well as the usual Liang approach.

For each of the two simulation scenarios, we considered total sample sizes of 100 and 500, and for each total sample size we considered 4, 10, and 20 clusters.

5 Results

5.1 Inverse-intensity weighted GEEs

For the inverse-intensity weighted GEEs, we consider first the sandwich variance adjustments, for simplicity focussing on the weights accounting for cluster as a factor (results were similar for other weights). As can be seen from Table 1, the usual sandwich variance underestimates the true variance. The MdR correction overestimates, but is closer to the true variance, while the KC correction underestimates but is also closer. The MBN correction underestimates for 4 and 10 clusters and overestimates for 20 clusters. The average of the KC and MdR corrections overestimates, but to a smaller degree than the MdR correction.

Turning to coverage probabilities, no estimator had good coverage for four clusters, while for 10 clusters the t-based confidence intervals for the MdR, MBN and the average of the KC and MdR corrections had close to

Table 1. Simulation under within-cluster conditional independence of visit and outcome processes given the observed history.

	Four clusters, 25/cluster			Ten clusters, 10/cluster			Twenty clusters, 5/cluster		
ESE	138			122			115		
Correction	ASE	CPZ	CPt	ASE	CPZ	CPt	ASE	CPZ	CPt
Uncorrected	88	70	92	105	88	91	106	91	94
MdR	182	91	98	133	93	96	120	95	96
KC	126	84	98	116	90	93	112	93	95
MBN	118	83	98	120	91	95	119	95	96
avKCMdR	157	88	99	125	91	95	116	94	96

ESE: Empirical standard error, the standard deviation of the 1000 estimated values; ASE: Average standard error, the mean of the 1000 estimates of standard error; KC: mean of the 1000 Kauermann & Carroll estimated standard errors; MdR: mean of the 1000 Mancl & deRouen estimated standard errors; MBN: mean of the 1000 Morel, Bokossa & Neerchal estimated standard errors; avKCMdR: mean of the 1000 square roots of the average of the KC and MdR variances; CPz: coverage of the 95% Wald z confidence intervals; CPt: coverage of the 95% Wald t confidence intervals.

Table 2. Simulation under within-cluster conditional independence of visit and outcome processes given the observed history.

	Four clusters, 25/cluster				Ten clusters, 10/cluster				Twenty clusters, 5/cluster			
Weight	Mean	ESE	MBN	MdR	Mean	ESE	MBN	MdeR	Mean	ESE	MBN	MdR
Naive	−250	148	104	190	−235	136	118	143	−238	132	125	134
Frailty	−252	139	114	183	−244	123	120	134	−247	116	119	121
Factor	−252	138	118	182	−245	122	120	133	−249	115	119	120
Strata	−271	214	129	248	−264	215	157	228	−263	218	170	223
	4 clusters, 125/cluster				10 clusters, 50/cluster				20 clusters, 25/cluster			
Weight	Mean	ESE	MBN	MdR	Mean	ESE	MBN	MdR	Mean	ESE	MBN	MdR
Naive	−247	109	66.2	127	−235	89.6	73.2	93.7	−235	72.5	69.9	76.1
Frailty	–253	99.9	75.4	122	−249	73.8	72.1	80.6	−247	60.7	65.0	65.8
Factor	−253	99.6	76.9	121	−249	73.6	72.1	80.3	−248	60.5	64.9	65.6
Strata	−264	160	91.6	170	−253	163	121	163	−253	155	131	160

Note: The true effect is −249. ESE: Empirical standard error, the standard deviation of the 1000 estimated values; MdR: mean of the 1000 Mancl & deRouen standard error estimates; MBN: mean of the 1000 Morel, Bokossa & Neerchal standard error estimates.

nominal coverage. For 20 clusters, all the t-based confidence intervals performed well, as did the z-based confidence intervals for the MdR, MBN, and average of the KC and MdR corrections. The KC correction is significantly slower to compute than the other two, as it involves taking the square root of a high dimensional matrix. Thus, in what follows we consider the MBN and MdR corrections only.

Failing to account for clustering resulted in biased estimates of the treatment effect on the AUC (Table 2), while for 10 and 20 clusters using weights that accounted for clustering through either a factor variable in the proportional hazards regression or a frailty variable led to reduced bias as well as slightly smaller standard errors. The frailty approach generally had larger bias than the weights that captured cluster effects as a factor. Using weights stratified on cluster led to more bias than the weights that ignored clustering, as well as increased standard errors. The bias did, however, reduce with the larger sample size.

5.2 Semi-parametric joint models

With just four clusters, the clustered Liang method led to large bias and standard errors. This was caused by a few very large estimates in the simulation; as can be seen in Table 3 the trimmed mean was close to the true difference in AUC between randomised groups. With a larger number of clusters, the clustered Liang method had smaller bias, while the usual Liang method that ignores clustering underestimated the decrease in AUC due to the intervention. Furthermore, the bias in the usual Liang method did not improve with increasing sample size.

Table 3. Simulation under correlated random effects at both the subject and cluster level, with no time-dependent covariates in the visit process model.

	Four clusters, 25/cluster			Ten clusters, 10/cluster			Twenty clusters, 5/cluster		
Method	Mean	Trimmed Mean	ESE	Mean	Trimmed Mean	ESE	Mean	Trimmed Mean	ESE
Cluster Liang	6.9×10^{10}	−249	1.8×10^{12}	−243	−248	141	−248	−248	125
Regular Liang	−227	−227	152	−221	−221	135	−220	−220	124
	4 clusters, 125/cluster			10 clusters, 50/cluster			20 clusters, 25/cluster		
Method	Mean	Trimmed Mean	ESE	Mean	Trimmed Mean	ESE	Mean	Trimmed Mean	ESE
Cluster Liang	2.0×10^{7}	−248	7.8×10^{8}	−248	−248	81	−244	−244	68
Regular Liang	−225	−226	102	−222	−222	76	−217	−217	64

Note: The true effect is −245. ESE: Empirical standard error, the standard deviation of the 1000 estimated AUCs.

Table 4. Simulation under correlated random effects at both the subject and cluster levels, with a time-dependent covariate in the visit process model.

	Four clusters, 25/cluster			Ten clusters, 10/cluster			Twenty clusters, 5/cluster		
Method	Mean	Trimmed Mean	ESE	Mean	Trimmed Mean	ESE	Mean	Trimmed Mean	ESE
Cluster Liang	1.40×10^{7}	−273	3.3×10^{8}	−255	−255	128	−253	−253	119
Regular Liang	−230	−230	152	−238	−238	130	−238	−239	119
	4 clusters, 125/cluster			10 clusters, 50/cluster			20 clusters, 25/cluster		
Method	Mean	Trimmed Mean	ESE	Mean	Trimmed Mean	ESE	Mean	Trimmed Mean	ESE
Cluster Liang	−421	−249	3323	−244	−244	79	−250	−250	68
Regular Liang	−229	−230	111	−229	−230	80	−235	−235	69

Note: The true effect is −245. ESE: Empirical standard error, the standard deviation of the 1000 estimated AUCs.

Similar results held when introducing a time-dependent covariate to the visit intensity process and using multiple outputation before implementing the two Liang procedures (Table 4). For four clusters, the clustered Liang method was badly biased. For larger numbers of clusters, the clustered Liang method had a much smaller bias that decreased on increasing the total sample size to 500. By comparison, using the Liang method without accounting for clustering resulted in larger bias as the sample size increased.

6 Example: Is air quality associated with outdoor play in children aged 2–9 years?

To illustrate these methods, we examine the relationship between air quality and outdoor play in a longitudinal cohort of children aged 2–9 years. Many children do not meet guidelines on physical activity,[35] and there is evidence of both a positive association between outdoor play and physical activity,[36,37] and that increasing opportunity for outdoor play increases moderate to vigorous physical activity.[38] Opportunities for outdoor play may be limited by poor air quality. During periods of poor air quality, advisories urge particular caution around vigorous physical activity, with children and those with respiratory or cardiovascular disease identified as particularly at risk.[39]

Children were participants in the TARGet Kids! study, which recruits healthy children aged 0–5 through their family practice or pediatrician's office.[13] As part of usual care, children's parents are advised to return for annual well-child visits, in addition to visits as needed. The aim was to administer TARGet Kids! questionnaires at the child's well-child visit. There is variability in when parents schedule these visits, and some years parents may not schedule a well-child visit. In the absence of a well-child visit, questionnaires can be administered at an as-needed visit, if this occurs. The questionnaires collect information on diet, physical activity, screen time, behavioural characteristics, as well as demographics. We used data from 1 January 2012 through 9 May 2019. While some

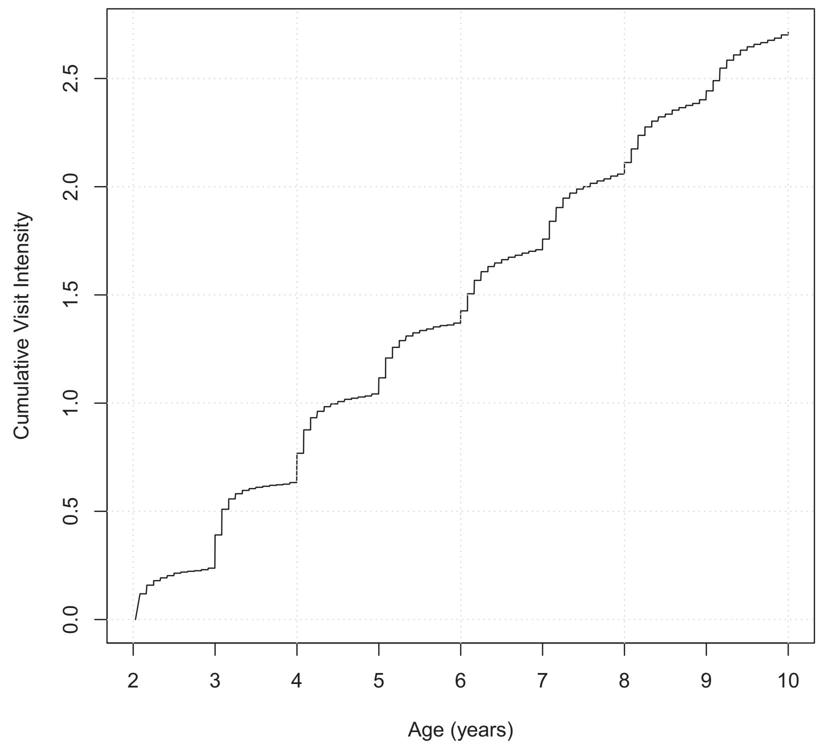

Figure 1. Cumulative visit intensity for the TARGet Kids! cohort. The cumulative visit intensity is the integrated hazard and gives the cumulative mean number of visits per patient on attaining each age.

information is available from all visits (e.g. height, weight), this analysis includes only "research visits", i.e. visits at which the TARGet Kids! questionnaires were administered.

Outdoor play was measured by parent report. Specifically, parents were asked, "Aside from time in day care and preschool, on a typical weekday, how much time does your child spend outside in unstructured free play?" Air quality was measured using publicly available records recorded at one of five stations in the Greater Toronto Area. The measure of air quality changed in 2015, and thus from 2015 on, the exposure of interest is the Air Quality Health Index (AQHI)[40] recorded at 4 p.m. at the station closest to the child's home (ascertained using the Forward Sortation Area – the first three digits of the child's postal code) over the two weeks prior to the visit date; for each visit, we determined the number of days in the previous two weeks for which the AQHI was between 4 and 6 (moderate risk), 7 and 10 (high risk), and >10 (very high risk). From January 2012 to 31 December 2014, the exposure of interest is the Air Quality Index (AQI)[41] recorded at 4 p.m. at the station closest to the child's home over the two weeks prior to the visit date. For each visit, we determined the number of days in the previous two weeks for which the AQI was between 32 and 49 (moderate), 50 and 99 (poor), and >99 (very poor). Due to the change in measure of exposure, analyses assessed whether there was an interaction between air quality category (moderate, poor/high risk, very poor/very high risk) and era (pre vs/post 1 January 2015).

Since the exposure of interest, air quality, is an external variable, we sought to adjust for other correlated external variables (temperature, calendar month, precipitation, year), but not subject-level covariates, as these would not be confounders. The exception is the child's age, as this is associated with calendar year. We thus included child age and birth year to account for potential trends in physical activity by age as well as by birth cohort.

There is significant variation within the Greater Toronto Area with respect to traffic density, housing density, green space, and socioeconomic status. This was modelled by clustering by neighbourhood. Since the methods assume that random effects are nested (subject within cluster), children were censored when they moved out of their original neighbourhood. Multiple imputation, implemented using the mice package in R,[42] was used to handle missing data.

Analysis began by modelling the visit frequency. Visit intensity was highest in the months following the child's birthday (Figure 1). Potential predictors of visit intensity were entered into a multivariable model; those that were not significant were then removed. The resulting visit intensity model was used to fit an inverse-intensity weighted GEE. Due to the large number of clusters ($J = 63$), cluster robust standard errors were not needed.

Table 5. Baseline characteristics of the 5372 TARGet Kids!

Variable	*n* (%) or Median (IQR)	Number missing (%)
Child age (months)	46 (25, 64)	0
Male	2779 (52%)	4 (0.1%)
Child BMI z-score	0.20 (−0.46, 0.87)	70 (1%)
Child birthweight (kg)	3.3 (2.9, 3.7)	290 (5%)
Maternal age at birth (years)	34 (31, 36)	350 (7%)
Maternal education		76 (1%)
Apprenticeship/trades cert/diploma	59 (1%)	
College/other non-uni diploma	703 (13%)	
High school certificate or equivalent	359 (7%)	
No certificate, diploma or degree	55 (1%)	
University certificate, diploma or degree	4120 (78%)	
Mother born in Canada	3383 (66%)	206 (4%)
Self-reported income (CAD)		174 (3%)
<10 K	62 (1%)	
10–19 K	105 (2%)	
20–29 K	127 (2%)	
30–39 K	156 (3%)	
40–49 K	178 (3%)	
50–60 K	182 (4%)	
60–80 K	354 (7%)	
80–100 K	466 (9%)	
100–150 K	1103 (21%)	
>150 K	2465 (47%)	
Exposure to second-hand smoke	553 (10%)	80 (1%)
Daycare	1698 (33%)	204 (8%)
School age	2118 (39%)	0
Minutes of outdoor play	35 (23, 60)	335 (6%)
Wheezing in last 12 months	1095 (21%)	47 (1%)
Asthma	313 (7%)	524 (10%)
Number of visits	2 (1, 3)	0
Duration of follow-up	68 (39, 79)	0

Note: Participants aged 2–9 years living in the Greater Toronto Area. IQR: Inter-quartile range; CAD: Canadian dollars; BMI: body mass index.

We used the clustered Liang model to relax the assumption that random effects are independent for the visit and outcome processes. We added neighbourhood as a fixed effect to the visit intensity model to derive an estimate of the within-subject variance, as well as to estimate B_{ij}^{s} (see Appendix 1). B_{j}^{c} was estimated using Monte Carlo simulation with 1000 iterations. A limitation of the clustered Liang method with this dataset is that multiple outputation cannot be used with a lagged covariate in the visit process model when the first observed visit time is stochastic: the lagged covariate for the first visit time is missing. While discarding the first visit would fix this problem (and is the approach used with inverse-intensity weighted GEEs), semi-parametric joint models require the censoring time to be independent of the outcomes and observation times given the covariates. This is plausible when the start time of observation is the larger of 24 months and the date of study enrollment and the stop time of observation is the date the data was cut, but is no longer true when the start time of observation is the second visit after age 24 months. For this reason, when fitting the Liang model, we considered a simpler (albeit incorrect) model for the visit times that used only time-invariant covariates. The clustered Liang method requires an estimate of the covariates not just at the observation times, but also for every age under the time frame of observation (in this case, 24-120 months). The air quality covariate is available on a daily basis while age is to the nearest month. Birth date was estimated as the mean of the date minus the age at each visit and then air quality at each month of age was computed based on the date at which the child reached that age. Standard errors for the Liang model were estimated using a non-parametric bootstrap, with clusters as the sampling units.

There were 5428 children in the analysis, with a total of 11,717 observations (11,307 after censoring due to moving). Their characteristics at baseline are given in Table 5. The model for the visit intensity is given in Table 6. Children whose mothers were born in Canada, and who had eaten vegetables in the past three days were likely to

Table 6. Visit process model.

Covariate	Hazard ratio (95% Confidence Interval)
zBMI at last visit	0.97 (0.94, 1.00)
Family income at last visit (CAD)	
Less than 10,000	0.74 (0.45, 1.20)
10,000 to 19,999	ref
20, 000 to 29,999	1.19 (0.84, 1.68)
30,000 to 39,999	1.17 (0.89, 1.53)
40,000 to 49,999	1.15 (0.86, 1.54)
50,000 to 59,999	1.35 (0.97, 1.88)
60,000 to 79,999	1.50 (1.09, 2.08)
80,000 to 99,999	1.42 (1.07, 1.89)
100,000 to 149,999	1.57 (1.15, 2.13)
150,000 or more	1.53 (1.16, 2.02)
Mother born in Canada	1.07 (1.01, 1.13)
Outdoor play at last visit (h)	0.95 (0.92, 0.98)
Had eaten vegetables in past three days at last visit	1.23 (1.09, 1.40)
Maternal age at birth (years)	
12–19	0.81 (0.51, 1.29)
20–29	0.81 (0.74, 0.89)
30–39	ref
40–49	0.95 (0.85, 1.06)
50–59	1.26 (0.49, 3.21)
Primary Care Practice	
A	0.69 (0.63, 0.75)
B	0.83 (0.77, 0.90)
C	0.95 (0.88, 1.02)
D	0.27 (0.17, 0.46)
E	0.76 (0.66, 0.87)
F	0.34 (0.20, 0.56)
G	0.23 (0.19, 0.28)
H	0.99 (0.90, 1.08)
I	0.65 (0.56, 0.77)
J	ref

return sooner, while children with larger zBMI and children who spent more time playing outdoors were likely to visit less often. Higher income was associated with more frequent visits, while children whose mothers were in their twenties at birth tended to visit less frequently than children whose mothers were in their thirties. There was also variation in visit frequency by primary care practice.

The initial intent was to regress minutes of outdoor play onto age, birth year (to address cohort effects), precipitation, air quality, temperature and calendar month. However, temperature and calendar month were colinear, and so we used calendar month without temperature. In no case was the interaction between year and air quality significant, and so the following results present models with main effects only. There were no days with very high risk air quality.

Results for inverse-intensity weighting showed more outdoor play in the summer, decreasing outdoor play with increasing age, and less outdoor play in periods with high-risk air quality (Table 7). This last association was statistically significant at the 5% level when accounting for clustering, but not when clustering was ignored, nor when using an unweighted GEE that ignores the informative observation process. The clustered Liang method produced results that were in general qualitatively similar to the IIW-GEEs. There continued to be more outdoor play in the summer months, although to a lesser degree. There was much less play in the month of December than other months. Moderate risk air quality was associated with more outdoor play, while high risk air quality continued to be associated with less outdoor play, although not significantly so. There was clear association between the subject-level random effects in the visit and outcome process (captured by θ, see Table 7 and equation (9)); the association between cluster-level random effects (captured by ϕ) was weaker.

Table 7. Regression coefficients and standard errors (SEs) on regressing typical weekday free play (in minutes) onto month, precipitation and air quality using (a) ignoring clustering and other predictors of visit frequency, i.e. with no weights; inverse intensity weighting (b) accounting for clustering using frailty weights, (c) accounting for clustering using factor weights, (d) ignoring clustering, and (e) the clustered Liang method.

	Regression coefficient (SE)				
Covariate	GEE (no weights)	Frailty	Factor	Ignoring clusters	Clustered Liang
Age (months)	−0.19 (0.03)	−0.19 (0.04)	−0.19 (0.04)	−0.19 (0.03)	n/e
Birth year	−0.55 (0.35)	−0.42 (0.43)	−0.46 (0.43)	−0.63 (0.34)	−0.58 (0.41)
Month					
January	Reference	Reference	Reference	Reference	
February	1.73 (2.46)	0.48 (3.05)	0.40 (3.07)	2.55 (2.47)	0.19 (3.67)
March	3.30 (2.61)	2.11 (2.86)	2.14 (2.90)	4.41 (2.57)	−0.17 (3.84)
April	6.13 (2.77)	7.24 (2.94)	7.18 (2.93)	5.25 (2.60)	4.48 (3.81)
May	10.53 (2.79)	11.44 (3.62)	11.37 (3.63)	9.90 (2.59)	8.21 (3.46)
June	21.22 (3.12)	23.48 (3.69)	23.61 (3.72)	20.62 (3.02)	18.0 (3.92)
July	19.06 (3.24)	20.25 (3.78)	20.23 (3.77)	19.09 (3.20)	12.8 (4.68)
August	23.55 (3.40)	26.50 (4.16)	26.76 (4.23)	22.65 (3.25)	12.1 (3.92)
September	12.70 (2.74)	14.19 (2.95)	14.19 (2.93)	11.90 (2.60)	12.8 (3.75)
October	5.96 (2.88)	7.07 (3.62)	7.20 (3.66)	5.82 (2.72)	10.1 (3.32)
November	2.60 (2.46)	2.77 (3.24)	2.70 (3.20)	3.34 (2.41)	8.52 (3.07)
December	−1.45 (2.53)	−1.68 (2.39)	−1.61 (2.38)	−1.03 (2.50)	−27.7 (5.09)
Days in past two weeks with >1 mm precipitation	0.24 (0.30)	0.30 (0.44)	0.31 (0.44)	0.25 (0.29)	1.49 (0.50)
Air quality at					
Low risk	Reference	Reference	Reference	Reference	Reference
Moderate risk	0.39 (0.35)	0.17 (0.39)	0.20 (0.40)	0.53 (0.33)	2.25 (0.66)
High risk	−6.08 (3.88)	−6.87 (3.38)	−6.95 (3.39)	−4.99 (4.14)	−4.59 (4.22)
θ					−21.4 (4.6)
ϕ					11.0 (10.6)

Note: θ captures association between subject-level random effects in the visit and outcome processes, and ϕ captures association between cluster-level random effects in the visit and outcome processes. n/e: Not estimated. (The Liang method treats time as a nuisance parameter; the model allows for a non-parametric association that cancels out in the estimating equations so that the remaining coefficients are adjusted for the effect of time, in this case age, but the effect of time itself is not estimated.)

Table 8. Comparison of methods for handling clustered longitudinal data subject to irregular observation.

Method	Baseline auxiliary covariates	Time-dependent auxiliary covariates	Subject-level random effects	Clusters
IIW-GEE – weights ignoring clusters	✓	✓	✗	✗
IIW-GEE – factor or frailty weights	✓	✓	✗	✓
Liang semi-parametric joint model	✓	✗	✓	✗
Clustered semi-parametric joint model	✓	✗	✓	✓
Liang semi-parametric joint model with multiple outputation	✓	✓	✓	✗
Clustered semi-parametric joint model with multiple outputation	✓	✓	✓	✓

Note: For each method, the visit and outcome processes are required to be independent conditional on a set of variables. The table indicates which variables are included in that conditioning set. The more variables can be included, the less strict the independence assumptions.

7 Discussion

In this paper, we demonstrated that ignoring clustering can lead to biased estimates of regression coefficients. Furthermore, we developed methods for accounting for clustering when analysing longitudinal data subject to irregular observation, using both inverse-intensity weighting and a semi-parametric joint model, and showed that our proposed methods resolve bias due to clustering for 10 or more clusters.

Clustering can induce dependence between the measurement intensity and outcome processes, thus violating the assumptions of traditional inverse-intensity weighting and semi-parametric joint models. By accounting for clustering in calculating the intensity weights, or by adding cluster-level random effects to the Liang model, conditional independence given clusters and/or random effects is restored.

With 10 and 20 clusters, all the variance correction methods achieved near nominal coverage when used in conjunction with a t-distribution. The KC correction tended to under-estimate variances while the MdR correction over-estimated, similar to previous work in the absence of weighting.[25–27] As in other work,[24] the MBN correction sometimes under-estimated and sometimes over-estimated variances. Inverse-intensity weights using clusters as a factor or as a frailty both performed well. In practice, unless there are a large number of clusters, it may be preferable to use cluster as a factor in computing the inverse intensity weights, as it is less computationally demanding.

The Liang method also performed well with 10 and 20 clusters. The choice between the clustered Liang or inverse-intensity weighted methods depends on whether, within clusters, the observation and outcome processes are conditionally independent given covariates. Liang et al.[5] note that this can be tested by examining the bootstrap distribution of the estimated regression coefficient θ, since it is this parameter that estimates the dependence between the outcome-level random effects and the visit process random effects. Table 8 provides a side-by-side comparison of the methods to help analysts choose an appropriate approach for their data. Code for implementing the clustered Liang model is included in Appendix 1, together with code for calculating the inverse-intensity weights using either a factor or frailty approach. Computationally, the inverse-intensity weighted methods are faster to compute than the clustered Liang model. Computation time for the Liang model for a single iteration of the simulation with a sample size of 500 was 24 s on an Intel i7-3667U core @ 2 GHz; obtaining a bootstrap standard error with 100 bootstrap samples would thus require 40 min.

Similar to other work, we found that neither of the methods worked well for four clusters.[43] In the GEEs, this was due to the sandwich variance corrections either under-estimating (KC) or over-estimating (MdR, MBN) the true variance. Since the estimating equations are aggregated up to the level of cluster, it is not surprising that it is difficult to estimate a variance over four clusters reliably. The issue with the clustered Liang model is similar: it requires an estimate of the variance of the cluster-level frailties, which is difficult to obtain with so few clusters. In cases with few clusters, clustering can be accounted for by including clusters as fixed effects in the outcome model. This is of course not possible when clusters are completely confounded with the covariate of interest, as in the case of a cluster-randomised trial. However, in this scenario it could be argued that four clusters are too small to allow for generalisability.[44]

The example shows that accounting for clustering can result in qualitatively different results, compared to ignoring clustering. There were also qualitative differences between the two approaches to accounting for clustering (inverse weighting vs. the semi-parametric joint model). In this example, neither model is correct; the Liang model ignores the impact of previous outdoor play on visit frequency while the inverse-intensity weighted GEE incorrectly assumes that the visit and outcome processes are conditionally independent within clusters given the auxiliary covariates. This dependence may explain the striking difference in estimated mean minutes of outdoor play for December vs. January: visits were most frequent in the month following the child's birthday, but were also less frequent in December (564 visits in December vs. >850 in every other month); it may be that children who visited in December also had different outdoor play patterns (for example, due to travel or not going to school).

The example also highlights two important limitations of the clustered Liang approach: first, the inability to account for a time-dependent covariate in the visit process when the initial visit time is stochastic, and second, the need for time-dependent covariates in the outcome model to be known over the entire time-frame of interest rather than just at visit times. In a scenario like our example where it is suspected that each model is capturing important, but different associations between the visit and outcome processes, sensitivity analysis using both models may be the best option.

There are many aspects of inference with clustered longitudinal data that remain to be explored. For example, in the inverse-intensity weighted GEEs we considered only the between-within degrees of freedom correction. Others have been proposed: Pan[20] and Fay and Graubard[17] consider a correction based on the variance of the sandwich estimator. Our motivation for using the between-within approach was that the Pan, and Fay and Graubard corrections were developed in the case of a single level of clustering, whereas we have two levels. Simulation work provides support for using the between-within method in the context of generalised linear mixed models.[29] We showed that our methods work well for 10 and 20 clusters but not for 4, leaving the question of what would happen with 5–9 clusters. Moreover, we used equal cluster sizes, where previous work with sandwich variance corrections showed that the performance of the estimators changed as cluster size varied.[24] Our simulation for IIW-GEEs considered only a linear link function; we note that the theory shows that the method should provide asymptotically unbiased inferences for a general link function. Many semi-parametric joint models have been proposed; we extended just one of these. It would be interesting to extend the Sun model[7] as this has a multiplicative model for the mean and would be useful for Poisson outcomes. Throughout this paper, we have treated clustering as a nuisance; however, there are scenarios in which understanding reasons for clustering may be useful. For example, in the simulation there is between-cluster variation in the efficacy of the intervention, in addition to between-cluster variation in the outcome in the control group. In practice, identifying reasons for cluster-to-cluster variation in intervention efficacy can be an important part of successful implementation of the intervention. Individual- and cluster-level moderators and mediators can be explored in a causal analysis, see Hong[45] for a detailed review.

The availability of electronic health records makes data collection faster and more efficient, and so clustered longitudinal data with irregular follow-up is likely to become more common. We have shown that ignoring the clustering in the analysis can lead to biased inferences. Our proposed methods correct this bias, thus facilitating correct analyses of these valuable data.

Declaration of conflicting interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: EMP received operating funds from the Natural Sciences and Engineering Research Council, and is supported by a New Investigator Award from the Canadian Institutes of Health Research.

ORCID iD

Eleanor M Pullenayegum https://orcid.org/0000-0003-4265-1330

References

1. Gulliford M, van Staa T, McDermott L, et al. Cluster randomized trials utilizing primary care electronic health records: methodological issues in design, conduct, and analysis (eCRT study). *Trials* 2014; **15**: 220, https://www.ncbi.nlm.nih.gov/pubmed/24919485
2. Annual physical examination practices by province/territory in Canada. Technical report, The College of Family Physicians of Canada, https://www.cfpc.ca/uploadedFiles/Health_Policy/CFPC_Policy_Papers_and_Endorsements/CFPC_Policy_Papers/CFPC%20PT%20Annual%20Exam.pdf (2013, accessed 7 January 2021).
3. Lin H, Scharfstein D and Rosenheck R. Analysis of longitudinal data with irregular, outcome-dependent follow-up. *J Royal Stat Soc Ser B* 2004; **66**: 791–813.
4. Buzkova P and Lumley T. Longitudinal data analysis for generalized linear models with follow-up dependent on outcome-related variables. *Can J Stat* 2007; **35**: 485–500.
5. Liang Y, Lu W and Ying Z. Joint modeling and analysis of longitudinal data with informative observation times. *Biometrics* 2009; **65**: 377–384.
6. Sun L, Song X and Zhou J. Regression analysis of longitudinal data with time-dependent covariates in the presence of informative observation and censoring times. *J Stat Plan Inference* 2011; **141**: 2902–2919.
7. Sun L, Mu X, Sun Z, et al. Semiparameteric analysis of longitudinal data with informative observation times. *Acta Mathematicae Applicatae Sinica – English Series* 2011; **27**: 29–42.
8. Sun L, Song X, Zhou J, et al. Joint analysis of longitudinal data with informative observation times and a dependent terminal event. *J Am Stat Assoc* 2012; **107**: 688–700.
9. Song X, Mu X and Sun L. Regression analysis of longitudinal data with time-dependent covariates and informative observation times. *Scand J Stat* 2012; **39**: 248–258.
10. Cai N, Lu W and Zhang H. Time-varying latent effect model for longitudinal data with informative observation times. *Biometrics* 2012; **68**: 1093–1102.
11. Pullenayegum E and Lim L. Longitudinal data subject to irregular observation: a review of methods with a focus on visit processes, assumptions, and study design. *Stat Meth Med Res* 2016; **25**: 2992–3014.
12. Chen Y, Ning J and CC. Regression analysis of longitudinal data with irregular and informative observation times. *Biostatistics* 2015; **16**: 727–739. DOI:10.1093/biostatistics/kxv008.
13. Carsley S, Borkhoff C, Maguire J, et al. Cohort profile: the applied research group for kids (target kids!). *Int J Epidemiol* 2015; **44**: 776–88.

14. Rondeau V, Gonzalez J, Mazroui Y, et al. *Package" frailtypack"*, 2.7.2 ed, 2014, http://cran.r-project.org/web/packages/frailtypack/frailtypack.pdf
15. Chesher A and Jewitt I. The bias of a heteroscedasticity consistent covariance matrix estimator. *Economietrica* 1987; **55**: 1217–1222.
16. Kauermann G and Carroll R. A note on the efficiency of sandwich covariance matrix estimation. *J Am Stat Assoc* 2001; **96**: 1387–1396, https://www.ncbi.nlm.nih.gov/pubmed/24919485
17. Fay M and Graubard B. Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics* 2001; **57**: 1198. DOI:10.1111/j.0006-341X.2001.01198.x.
18. Morel J, Bokossa M and Neerchal N. Small sample correction for the variance of gee estimators. *Biometric J* 2003; **45**: 395–409.
19. Mancl L and DeRouen T. A covariance estimator for gee with improved small-sample properties. *Biometrics* 2001; **57**: 126–134.
20. Pan W and Wall M. Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. *Stat Med* 2002; **21**: 1429–41.
21. Gosho M, Sato Y and Takeuchi H. Robust covariance estimator for small-sample adjustment in the generalized estimating equations: a simulation study. *Sci J Appl Math Stat* 2014; **2**: 20–25.
22. Wang M and Long Q. Modified robust variance estimator for generalized estimating equations with improved small-sample performance. *Stat Med* 2011; **30**: 1278–1291. DOI:10.1002/sim.4150.
23. Westgate P and Burchett W. Improving power in small-sample longitudinal studies when using generalized estimating equations. *Stat Med* 2016; **35**: 3733–3744. DOI:10.1002/sim.6967.
24. Li P and Redden D. Small sample performance of bias-corrected sandwich estimators for cluster-randomized trials with binary outcomes. *Stat Med* 2015; **34**: 281–296. DOI:10.1002/sim.6344.
25. Lu B, Preisser J, Qaqish B, et al. A comparison of two bias-corrected covariance estimators for generalized estimating equations. *Biometrics* 2007; **63**: 935–941.
26. Fan C, Zhang D and Zhang C. A comparison of bias-corrected covariance estimators for generalized estimating equations. *J Biopharmaceut Stat* 2013; **23**: 1172–1187.
27. Ford W and Westgate P. A comparison of bias-corrected empirical covariance estimators with generalized estimating equations in small-sample longitudinal study settings. *Stat Med* 2018; **37**: 4318–4329.
28. Schluchter M and Elashoff J. Small-sample adjustments to tests with unbalanced repeated measures assuming several covariance structures. *J Stat Computat Simulat* 1990; **37**: 69–87.
29. Li P and Redden D. Comparing denominator degrees of freedom approximations for the generalized linear mixed model in analyzing binary outcome in small sample cluster-randomized trials. *BMC Med Res Methodol* 2015; **15**: 38. DOI: 10.1186/s12874-015-0026-x.
30. Pullenayegum E. Multiple outputation for the analysis of longitudinal data subject to irregular observation. *Stat Med* 2016; **35**: 1800–1818, https://www.ncbi.nlm.nih.gov/pubmed/26661690
31. Hoffman E, Sen P and Weinberg C. Within-cluster resampling. *Biometrika* 2001; **88**: 1121–1134.
32. Follmann D, Proschan M and Leifer E. Multiple outputation: inference for complex clustered data by averaging analyses from independent data. *Biometrics* 2003; **59**: 420–429.
33. Rosenheck R, Kasprow W, Frismn L, et al. Cost-effectiveness of supported housing for homeless persons with mental illness. *Arch Gen Psychiatr* 2003; **60**: 940–951.
34. Daley D, Vere-Jones D and Smirnov B. *An introduction to the theory of point processes: elementary theory and methods*. New York, NY: Springer, 2002.
35. Roberts K, Yao X, Carson V, et al. Meeting the Canadian 24-hour movement guidelines for children and youth. *Health Rep* 2017; **28**: 3–7, https://www.ncbi.nlm.nih.gov/pubmed/29044440
36. Copeland K, Khoury J and Kalkwarf H. Child care center characteristics associated with preschoolers' physical activity. *Am J Prev Med* 2016; **50**: 470–479, https://www.ncbi.nlm.nih.gov/pubmed/26585052
37. Tandon P, Saelens B, Zhou C, et al. A comparison of preschoolers' physical activity indoors versus outdoors at child care. *Int J Environ Res Public Health* 2018; **15**, https://www.ncbi.nlm.nih.gov/pubmed/30400603
38. Razak L, Yoong S, Wiggers J, et al. Impact of scheduling multiple outdoor free-play periods in childcare on child moderate-to-vigorous physical activity: a cluster randomised trial. *Int J Behav Nutr Phys Act* 2018; **15**: 34, https://www.ncbi.nlm.nih.gov/pubmed/29615061
39. Ministry of the environment conservation and parks, Government of Ontario. *AQHI categories and health messages*, http://www.airqualityontario.com/aqhi/health_messages.php#HealthMessages (2020, accessed 13 January 2020).
40. Stieb D, Burnett R, Smith-Doiron M, et al. A new multipollutant, no-threshold air quality health index based on short-term associations observed in daily time-series analyses. *J Air Waste Manag Assoc* 2008; **58**: 435–50, https://www.ncbi.nlm.nih.gov/pubmed/18376646
41. Ontario Agency for Health Protection and Promotion (Public Health Ontario), Chen H and Copes R. *Review of air quality index and air quality health index*. Toronto, ON: Queen's Printer for Ontario, 2013.
42. van Buuren S and Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. *J Stat Software* 2011; **41**: 1–67.

43. McNeish D and Stapleton L. Modeling clustered data with very few clusters. *Multivariate Behavio Res* 2016; **51**: 495–518.
44. Taljaard M, Teerenstra S, Ivers N, et al. Substantial risks associated with few clusters in cluster randomized and stepped wedge designs. *Clin Trials* 2016; **13**: 459–63, https://www.ncbi.nlm.nih.gov/pubmed/26940696
45. Hong G. Causality in *a social world: moderation, meditation and spill-over*. Hoboken, NJ: Wiley, 2002.

Appendix 1. Conditional expectation of random effects

To obtain estimates of $E(\eta_{Vsij}|m_{ij}, C_{ij}, i = 1, \ldots, n_j)$ and $E(\eta_{Vcj}|m_{ij}, C_{ij}, i = 1, \ldots, n_j)$ and hence $\hat{B}^s_{ij}$, $\hat{B}^c_j$, we require a parametric distribution for the random effects $\eta_{V\ sij}$ and $\eta_{V\ cj}$. Following Liang et al.,[5] assume that $\eta_{Vsij} \sim \Gamma(1/\sigma_s^2, 1/\sigma_s^2)$, $\eta_{Vcj} \sim \Gamma(1/\sigma_c^2, 1/\sigma_c^2)$. Since

$$m_{ij}|C_{ij}, \eta_{Vsij}, \eta_{Vcj} \sim Poisson(\eta_{Vcj}\eta_{Vsij}\Lambda(C_{ij}))$$

it follows that

$$\eta_{Vsij}|\eta_{Vcj}, m_{ij}, C_{ij} \sim \Gamma(m_{ij} + 1/\sigma_s^2, \eta_{Vcj}\Lambda_0(C_{ij})\exp(Z_{ij}\gamma) + 1/\sigma_s^2)$$

and so

$$E(\eta_{Vsij}|\eta_{Vcj}, m_{ij}, C_{ij}, i = 1, \ldots, n_j) = \frac{m_{ij}\sigma_s^2 + 1}{\eta_{Vcj}\Lambda_0(C_{ij})\exp(Z_{ij}\gamma)\sigma_s^2 + 1} \tag{11}$$

If there are sufficient observations per cluster, we can define $\alpha_j = \log(\eta_{Vcj})$ so that $\Lambda_i(C_{ij}) = \eta_{Vsij}$ $\Lambda_0(C_{ij})\exp(Z_{ij}\gamma + \alpha_j)$, and so

$$E(\eta_{Vsij}|m_{ij}, C_{ij}, i = 1, \ldots, n_j) = \frac{m_{ij}\sigma_s^2 + 1}{\exp(Z_{ij}\gamma + \alpha_j)\Lambda_0(C_{ij})\sigma_s^2 + 1}$$

Alternatively, the expectation can be obtained by numerical integration over the distribution of $\eta_{Vcj}|m_{ij}, C_{ij}, i = 1, \ldots, n_j$.
Similarly for B^c, if $\eta_{Vcj} \sim \Gamma(1/\sigma_c^2, 1/\sigma_c^2)$, then

$$m_{+j} = \sum_j m_{ij}|C_{ij}, \eta_{Vcj}, \eta_{Vsij}, i = 1, \ldots, n_j \sim Poisson(\sum_i \eta_{Vcj}\eta_{Vsij}\Lambda_0(C_{ij})\exp(Z_{ij}\gamma))$$

It follows that

$$\eta_{Vcj}|m_{+j}, C_{ij}, \eta_{Vsij}, i = 1, \ldots, n_j \sim \Gamma(m_{+j} + 1/\sigma_c^2, \sum_i \eta_{Vsij}\Lambda_0(C_{ij})\exp(Z_{ij}\gamma) + 1/\sigma_c^2)$$

and so

$$E(\eta_{Vcj}|m_{+j}, C_{ij}, \eta_{Vsij}, i = 1, \ldots, n_j) = \frac{m_{+j}\sigma_c^2 + 1}{\sum_i \eta_{Vsij}\Lambda_0(C_{ij})\exp(Z_{ij}\gamma)\sigma_c^2 + 1}$$

and $E(\eta_{Vcj}|m_{ij}, C_{ij}, i = 1, \ldots, n_j)$ can be estimated by Monte Carlo simulation (i.e. repeatedly simulating $\eta_{Vsij} \sim Gamma(1/\sigma_s^2, 1/\sigma_s^2)$).

A.1 Solution with cluster-invariant censoring and covariates

In the special case that $C_{ij} = C_j$ for all i and j, and all the visit process covariates are at the cluster level (i.e. $Z_{ij} = Z_j$), the expectation simplifies as follows. Without loss of generality, suppose we are interested in

$E(\eta_{Vs1j}|m_{ij,}i=1,\ldots,n_j)$, and note that this is the same as $E(\eta_{Vs1j}|m_{1j},m_{+j})$. Define $\Lambda_j=\Lambda_0(C_j)\exp(Z_j\gamma)$. Then the joint distribution of the random effects given the event counts can be written as

$$\begin{aligned}
&f(\eta_{Vs1j},\eta_{Vs+j},\eta_{Vcj}|m_{1j},m_{+j})\\
&=f(\eta_{Vs1j},\eta_{Vs+j}-\eta_{Vs1j},\eta_{Vcj}|m_{1j},m_{+j}-m_{1j})\\
&=f(\eta_{Vs1j})f(\eta_{Vs+j}-\eta_{Vs1j})f(\eta_{Vcj})f(m_{1j},m_{+j}-m_{1j}|\eta_{Vs1j},\eta_{Vs+j}-\eta_{Vs1j},\eta_{Vcj})/\\
&\quad f(m_{1j},m_{+j}-m_{1j})\\
&\quad\propto\exp(-(\eta_{Vs1j}/\sigma_s^2+(\eta_{Vs+j}-\eta_{Vs1j})/\sigma_s^2+\eta_{Vcj}/\sigma_c^2+\eta_{Vs1j}\eta_{Vcj}\Lambda_j\\
&\quad+(\eta_{Vs+j}-\eta_{Vs1j})\eta_{Vcj}\Lambda_j))\times\eta_{Vs1j}^{m_{1j}+1/\sigma_s^2-1}(\eta_{Vs+j}-\eta_{Vs1j})^{m_{+j}-m_{1j}+(n_j-1)/\sigma_s^2-1}\eta_{Vcj}^{m_{+j}+1/\sigma_c^2-1}
\end{aligned}$$

Now consider a change of variable, with $w=\eta_{Vs+j}$ and $p=\eta_{Vs1j}/w$. We then have

$$\begin{aligned}
f(w,p,\eta_{Vcj}|m_{1j},m_{+j})&\propto\exp(-(w/\sigma_s^2+\eta_{Vcj}/\sigma_c^2+w\eta_{Vcj}\Lambda_j))p^{m_{1j}+1/\sigma_s^2-1}\\
&\times(1-p)^{m_{+j}-m_{1j}+(n_j-1)/\sigma_s^2-1}w^{m_{+j}+n_j/\sigma_s^2-1}\eta_{Vcj}^{m_{+j}+1/\sigma_c^2-1}
\end{aligned}\tag{12}$$

Integrating over η_{Vcj}

$$\begin{aligned}
f(w,p|m_{1j},m_{+j})&\propto\exp(-w/\sigma_s^2)w^{m_{+j}+n_j/\sigma_s^2-1}(1/\sigma_c^2+w\Lambda_j)^{-(m_{+j}+1/\sigma_c^2)}\\
&\times p^{m_{1j}+1/\sigma_s^2-1}(1-p)^{m_{+j}-m_{1j}+(n_j-1)/\sigma_s^2-1}
\end{aligned}$$

It follows that $W\perp\!\!\!\perp P|m_{1j},m_{+j}$ and $P\sim Beta(m_{1j}+1/\sigma_s^2,m_{+j}-m_{1j}+(n_j-1)/\sigma_s^2)$. Therefore $E(\eta_{Vs1j}|m_{1j},m_{+j})=E(WP|m_{1j},m_{+j})=E(W|m_{1j},m_{+j})\frac{m_{1j}+1/\sigma_s^2}{m_{+j}+n_j/\sigma_s^2}$, with

$$E(W|m_{1j},m_{+j})=\frac{\int_0^\infty(w\Lambda_j+1/\sigma_c^2)^{-(m_{+j}+1/\sigma_c^2)}w^{m_{+j}+n_j/\sigma_s^2}\exp(-w/\sigma_s^2)\,\mathrm{d}w}{\int_0^\infty(w\Lambda_j+1/\sigma_c^2)^{-(m_{+j}+1/\sigma_c^2)}w^{m_{+j}+n_j/\sigma_s^2-1}\exp(-w/\sigma_s^2)\,\mathrm{d}w}$$

Similarly, $E(\eta_{Vcj}|m_{1j},m_{+j})$ can be derived by integrating equation (12) over p and then w

$$\begin{aligned}
&f(w,\eta_{Vcj}|m_{1j},m_{+j})\propto\exp(-(w(\eta_{Vcj}\Lambda_j+1/\sigma_s^2)+\eta_{Vcj}/\sigma_c^2))w^{m_{+j}+n_j/\sigma_s^2-1}\eta_{Vcj}^{m_{+j}+1/\sigma_c^2-1}\\
&f\eta_{Vcj}|m_{1j},m_{+j})\propto\exp(-\eta_{Vcj}/\sigma_c^2)(\eta_{Vcj}\Lambda_j+1/\sigma_s^2)^{-(m_{+j}+n_j/\sigma_s^2)}\eta_{Vcj}^{m_{+j}+1/\sigma_c^2-1}
\end{aligned}$$

It follows that

$$E(\eta_{Vcj}|m_{1j},m_{+j})=\frac{\int_0^\infty(1/\sigma_s^2+\eta_{Vcj}\Lambda_j)^{-(m_{+j}+n_j/\sigma_s^2)}\eta_{Vcj}^{m_{+j}+1/\sigma_c^2}\exp(-\eta_{Vcj}/\sigma_c^2)\,\mathrm{d}\eta_{Vcj}}{\int_0^\infty(1/\sigma_s^2+\eta_{Vcj}\Lambda_j)^{-(m_{+j}+n_j/\sigma_s^2)}\eta_{Vcj}^{m_{+j}+1/\sigma_c^2-1}\exp(-\eta_{Vcj}/\sigma_c^2)\,\mathrm{d}\eta_{Vcj}}$$

A.2 Closed-form solution with cluster-invariant censoring and covariates and large clusters

Furthermore, if the number n_j of individuals per cluster is large in relation to σ_s, then it may be reasonable to take $W=n_j$. When this is the case, it follows from equation (12) that $\eta_{V\ cj}$ and P are independently distributed with

$$\begin{aligned}
&\eta_{Vcj}\sim\Gamma(m_{+j}+1/\sigma_c^2,n_j\Lambda_j+1/\sigma_c^2)\\
&P\sim Beta(m_{1j}+1/\sigma_s^2,m_{+j}-m_{1j}+(n_j-1)/\sigma_s^2)
\end{aligned}$$

Therefore

$$E(\eta_{Vcj}|m_{1j}, m_{+j}) = \frac{m_{+j} + 1/\sigma_c^2}{n_j\Lambda_j + 1/\sigma_c^2}$$
$$E(\eta_{Vs1j}|m_{1j}, m_{+j}) = n_j\frac{m_{1j} + 1/\sigma_s^2}{m_{+j} + n_j/\sigma_s^2}$$

A.3 Closed-form solution when subject and cluster-level frailty variances are small

In the case where both σ_s and σ_c are small, an approximate closed-form solution may be derived using a Taylor expansion of the conditional expectations in A.1.

$$\begin{aligned}E(\eta_{Vsij}|\eta_{Vcj}, m_{ij}, C_{ij}, i = 1, \ldots, n_j) &= \frac{m_{ij}\sigma_s^2 + 1}{\eta_{Vcj}\Lambda_0(C_{ij})\exp(Z_{ij}\gamma)\sigma_s^2 + 1}\\ &= (m_{ij}\sigma_s^2 + 1)(1 - \Lambda_0(C_{ij})\exp(Z_{ij}\gamma)\sigma_s^2 + O((\eta_{Vcj}\Lambda_0(C_{ij})\exp(Z_{ij}\gamma)\sigma_s^2)^2))\end{aligned}$$

Importantly, this only works if $\eta_{Vcj}\Lambda_0(C_{ij})\exp(Z_{ij}\gamma)\sigma_s^2 < 1$. Since η_{Vcj} is latent, there is no way to know this definitively, but the upper 95th and 99th percentiles of the Gamma$(1/\sigma_c^2, 1/\sigma_c^2)$ distribution could be examined.

A similar procedure can be used for the conditional expectation of the cluster-level frailties:

$$\begin{aligned}&E(\eta_{Vcj}|m_{+j}, C_{ij}, \eta_{Vsij}, i = 1, \ldots, n_j)\\ &= \frac{m_{+j}\sigma_c^2 + 1}{\sum_i \eta_{Vsij}\Lambda_0(C_{ij})\exp(Z_{ij}\gamma)\sigma_c^2 + 1}\\ &= (m_{+j}\sigma_c^2 + 1)\left(1 - \sigma_c^2\sum_i \Lambda_0(C_{ij})\exp(Z_{ij}\gamma)\right) + O\left(\left(\sum_i \eta_{Vsij}\Lambda_0(C_{ij})\exp(Z_{ij}\gamma)\sigma_c^2\right)^2\right)\end{aligned}$$

This approximation relies on $\sum_i \eta_{Vsij}\Lambda_0(C_{ij})\exp(Z_{ij}\gamma)\sigma_c^2 < 1$, and hence depends not only on the variances of the subject and cluster-level frailties beng small, but also the number of subjects per cluster being small enough in relation to the hazards and variances for the sum to remain smaller than 1. Provided the sum is smaller than 1, higher order expansions are possible in order to improve precision.

A.4 Estimates of random effect variances

These procedures require estimates of σ_s^2, and σ_c^2 which, using a method of moments, are as follows

$$\hat{\sigma}_c^2 = \frac{\sum_{ij} m_{ij}^2 - \sum_j m_{+j}^2}{\sum_j \Lambda 2_{+j} - \Lambda 2_{++}} - 1$$
$$\hat{\sigma}_s^2 = \frac{\sum_{ij} m_{ij}^2 - \sum_{ij} m_{ij}}{\sum_{ij} m_{ij}^2 - \sum_j m_{+j}^2}\frac{\sum_j \Lambda 2_{+j} - \Lambda 2_{++}}{\sum_j \Lambda 2_{+j}} - 1, \text{ with}$$

$$\Lambda_{+j} = \sum_i \hat{\Lambda}_0(C_{ij})\exp(Z_{ij}\hat{\gamma})$$
$$\Lambda_{++} = \sum_{ij} \hat{\Lambda}_0(C_{ij})\exp(Z_{ij}\hat{\gamma})$$
$$\Lambda 2_{+j} = \sum_i \hat{\Lambda}_0(^{C_{ij})2}\exp(2Z_{ij}\hat{\gamma})$$
$$\Lambda 2_{++} = \sum_j \Lambda_{+j}^2$$