Journal of Statistical Research *https://doi.org/10.3329/jsr.v57i12.72968*
2023, Vol. 57, No. 1-2, pp. 35-67 *ISSN 0256 - 422 X*

A SIMULATION STUDY TO ASSESS THE IMPACT OF MISSING VALUES ON THE PERFORMANCE OF DIFFERENT STATISTICAL METHODS FOR ANALYSIS OF BINARY REPEATED MEASURES DATA WITH AN ADDITIONAL HIERARCHICAL STRUCTURE

Elmabrok Masaoud*
School of Epidemiology and Public Health, University of Ottawa
451 Smyth Road, Ottawa, Ontario, K1H 8M5, Canada
Email: emasaoud@uottawa.ca

Henrik Stryhn
Department of Health Management, Atlantic Veterinary College
University of Prince Edward Island, Charlottetown, PEI C1A 4P3, Canada
Email: hstryhn@upei.ca

SUMMARY

The primary objective of the study was to assess the impact of missing values on the analysis of binary repeated measures data with an additional hierarchical structure. One motivating example for the present study was records of high somatic cell counts in milk samples obtained by approximately monthly sampling throughout the lactations of cows in dairy herds. Random effects models with autocorrelated ($\rho = 1, 0.9$ or 0.5) subject-level random effects were behind the simulated data. In general, the settings of the simulation were chosen to reflect a real somatic cell count dataset (scc40), except that the within-cow time series length was set to 8-time points for each cow. The estimation procedures considered were: Ordinary Logistic Regression (OLR), Alternating Logistic Regression (ALR), Weighted Generalized Estimating Equations (WGEE), Penalized Quasi Likelihood (PQL), Maximum likelihood via numerical integration (ML) and Bayesian Markov chain Monte Carlo (MCMC). Multiple scenarios of simulated incomplete datasets were considered and include: a scenario corresponded to a combination of missingness patterns present in the scc40 dataset (scc40 scenario) The remaining scenarios involved only drop-outs, and corresponded to either moderate or high percentages of values either missing at random (MAR) or not missing at random (NMAR), respectively. In the scc40 scenario, all estimation procedures except OLR performed well and produced estimates with small relative bias (generally less than 5%) for levels of missingness that roughly corresponded to the scc40 data. In MAR missingness scenarios, some biases were found for ALR, WGEE and PQL procedures, whereas the likelihood-based procedures were largely unaffected by the missing values. In NMAR scenarios, all procedures experienced similar and strong biases in the time coefficient; however, fixed effects estimates at the subject and cluster levels were relatively unaffected.

Keywords and phrases: Missing values; Hierarchical structure; Binary repeated measures data; Simulation.

* Corresponding author

1 Introduction

Missing values in binary repeated measures data with an additional hierarchical structure refers to data with incomplete records over time on the same subjects (e.g., patients, animals or farms), which in addition are nested within some (physical) clusters (e.g., hospitals, herds, provinces). Missing data usually arise when some subjects are not available for certain measurements. Subjects may leave the study at some point in time before completing their measurements (drop-outs), subjects may miss some measurements and reappear again for later measurements (intermittent missing values), or subjects may join the study at different times (delayed entry). Our motivating example was a literature scc40 dataset of (Dohoo *et al*., 2009, Chapter 31) on incomplete records of presence or absence of high somatic cell counts in milk samples from cows housed in multiple herds. Thus, the hierarchical structure is the clustering of cows in herds, the repeated measures are the monthly test records based on the milk samples, and the missing values are the incomplete records on each cow. A previous study by (Masaoud and Stryhn, 2020) targeted the added complexity of the additional hierarchical structure in a balanced full datasets setting, whereas the present study is focused on the missing values part. Generally, missingness in longitudinal data presents a potential source of bias. In part, the bias could be due to the change in data structure from being balanced to unbalanced, which in turn may raise technical difficulties, especially for those statistical methods that can only cope with balanced data. If the process of the observations being missing (the missingness mechanism) varies from subject to subject, the distribution of the observed outcome values may not be the same as for the full dataset.

Despite the large body of literature on missing data (Little and Rubin, 2002; Laird, 1988; Diggle and Kenward, 1994; Fitzmaurice, 2003; Little, 1995; Heyting *et al*., 1992; Hogan *et al*., 2004), listed in order of relevance to the present study, most authors agree that handling missing values is not a trivial task and that in many instances there is a need for sensitivity analysis (Kenward *et al*., 2001). Thus, additional information about the missingness mechanism is required. Missing data mechanisms have been classified into different categories (Little and Rubin, 2002) according to their randomness process. They include, missing completely at random where the probability of an observation being missing does not depend on the prior observed nor the future unobserved values of the outcome; missing at random where the probability of an observation being missing depends only on the prior observed outcome; and not missing at random where the probability of an observation being missing depends directly on the unobserved outcome(s).

Several procedures (models) have been proposed for the analysis of binary repeated measures data; a basic distinction is between marginal (population–averaged, or PA) and random effects (subject-specific, or SS) models (Neuhaus, 1992; Diggle *et al*., 2002, Chapters 8-9). Many articles have discussed the choice between the two models (PA vs. SS) (e.g., (Diggle *et al*., 2002, Chapters 8-9) or for balanced data (Masaoud and Stryhn, 2010, 2020)). However, the presence of missing values poses problems for procedures of both types, and to our knowledge the performance of statistical procedures for the analysis of binary repeated measures data with additional hierarchical structure in the presence of missing values has not yet been described.

Previous studies on missing values include assessments of the impact of drop-out missing data on different statistical methods (Ali and Talukder, 2005; Touloumi *et al*., 2001). Fitzmaurice (2004) recommends performing analysis of incomplete data using methods to handle various types of missing data mechanisms, in order to obtain insight into the actual type of missing data present. This approach may be difficult to employ and justify if there is a combination of different types of missing values within the same dataset. The analytical approach taken for the present study was simulation. Simulation studies can be targeted towards a specific data structure by incorporating as much of that structure as possible in the simulated datasets (Stryhn *et al*., 2000). This idea can be extended to incomplete data by matching also the missing data values.

In order to realistically reflect the choice an applied researcher faces when it comes to data analysis, only estimation procedures implemented in broadly accessible statistical software were considered for the study. Specifically, the following procedures previously studied for hierarchically structured binary repeated measures data (Masaoud and Stryhn, 2020) were included: maximum likelihood via numerical integration (ML), Bayesian Markov chain Monte Carlo (MCMC), penalized quasi-likelihood with binomial dispersion (PQL) and an extra-binomial dispersion (PQLx), ordinary logistic regression (OLR), alternating logistic regression (ALR), and weighted generalized estimating equations (WGEE). The adapted ALR macro for 3-level of clustering (Kunthel *et al*., 2014) is recently available when estimation of the association structure is of primary interest, though was not included in the present simulation study.

The primary objective of this study was to assess the impact of missing values on the performance of different statistical estimation procedures for the analysis of binary repeated measures data with an additional hierarchical structure. A secondary goal of this study was to demonstrate a simple simulation approach to assess the impact of missing values in an actual dataset.

2 Missing Values

Within the context of binary repeated measures data, let y_{ij} refer to complete binary records on each of n subjects ($i = 1, \dots, n$) at t time points ($j = 1, \dots, t$). Furthermore, let r_{ij} be the indicator of y_{ij} being missing. In this notation, a subject i drops out from the study at time d, if $r_{id-1} = 0$ and $r_{ij} = 1$ for all $j \geq d$. Little and Rubin (2002) (for a longitudinal data context, see e.g., (Laird, 1988)) classified missingness mechanisms in terms of the conditional distribution of (r_{ij}) given (y_{ij}). Note that in the following we also use r_{ij} as an indicator for a missing value of a particular type, which should be evident from the context.

2.1 Classification of missing data

Missing completely at random (MCAR) (Little and Rubin, 2002; Laird, 1988) refers to a missingness mechanism (or missing data process) that does not depend on prior observed outcome values or on intended measurement values of the outcome (unobserved outcome values), but may depend on covariates such as time. Little and Rubin (2002) showed that in the presence of an MCAR process, the estimated parameters are not biased by the absence of data, thus the missing data can be ignored. Diggle and Kenward (1994) introduced a completely random drop-out (CRD) process that assumes MCAR. One implication of the MCAR assumption is that the distribution of the prior observed outcome values at time j is the same regardless of whether a subject drops out or remains in the study after that particular time point. Also, the distribution of the unobserved outcome values is unaffected by the drop-out. Missing at random (MAR) (Little and Rubin, 2002; Laird, 1988) and random drop-out (RD) (Diggle and Kenward, 1994) refer to a missing data (drop-out) process that depends on the prior observed values of the outcome only. Not missing at random (NMAR) (Little and Rubin, 2002; Laird, 1988) and informative drop-out (ID) (Diggle and Kenward, 1994) refer to a missingness mechanism that depends on the unobserved outcome (current or future unobserved values).

2.2 Approaches to handle missing data

Several approaches have been proposed to assess and account for missing values (Fitzmaurice, 2003), including the complete case method (also termed "listwise deletion" (McKnight *et al*., 2007, Chapter 5)). By this method, subjects with at least one missing value are dropped from the analysis. Fitzmaurice (2003) and Little and Rubin (2002) showed that this method is valid only under the MCAR missing data process. Another approach is based on the observed data and called the available case method (also termed "pairwise deletion" (Little and Rubin, 2002; Fitzmaurice, 2003; McKnight *et al*., 2007, Chapter 5)). Fitzmaurice (2003) argued that WGEE falls under this approach. Kim and Curry (1977) showed that for an MCAR process, methods based on the available cases are considered more efficient than complete case methods, as one would expect because all the available data is used. Little (1988) and Little and Rubin (2002) explained that these methods assume the strong MCAR assumption. Little and Rubin (2002) argued that neither the complete case method nor the available case method are generally satisfactory. Little and Rubin (2002) showed that an MAR process can be ignored when using likelihood-based inference. Robins *et al*. (1995) showed that ordinary GEE does not allow an MAR process to be ignored, and outlined a weighting scheme (WGEE) to achieve valid inference under the MAR assumption. Its implementation for drop-out missing data is detailed by Jansen *et al*. (2006). Hogan *et al*. (2004) defined ignorability as the situation where "the missing data model can be left unspecified or ignored". For NMAR processes, both likelihood and GEE approaches can be extended to model the missing data (Molenberghs and Verbeke, 2005, Chapter 27). However, these approaches (Roy, 2003) fall beyond the present scope of this study.

2.3 Assessing the impact of missing data by simulation

A theoretical knowledge of which procedures under certain assumptions would provide biased or unbiased estimates is valuable, but does not give the analyst a quantitative sense of the impact of missing data in an actual dataset. The question posed is what biases might arise from the missing data under different assumptions about the missingness mechanisms. Here the impact of missing data means the difference between results for the incomplete dataset and those for the corresponding full dataset. Given an actual (incomplete) dataset this approach is counterfactual because the full dataset is not available. However, it lends itself to simulation if realistic models for the full dataset as well as the missingness mechanism can be established. We outline briefly how the MCAR and the MAR processes may be adapted to an actual dataset.

A first step is to discriminate between drop-outs, intermittent missing data and any other types of missing data. For each type of missing data, a binary matrix of indicators of missing values (termed a "shadow matrix" (Cook and Swayne, 2007)) with rows corresponding to subjects and columns corresponding to possible instances of "events" of missing values is created. For example, each row in the shadow matrix for drop-outs consists of a series of zeros until either the occurrence of a drop-out (represented by a 1 and followed by missing values) or the last time point in the series. This structure is similar to that of discrete time single event data (Singer and Willett, 1993). For intermittent missing values, each subject could have multiple events corresponding to a standard two-level (repeated measures) data structure.

Under an MCAR assumption, shadow matrix data would most naturally be analyzed by logistic regression models that may incorporate covariates such as subject characteristics or time. Parameter estimates from the actual dataset are then used to generate missing data values for the simulation. Under an MAR assumption, the logistic regression models may be extended to include outcomes at one or several previous time points, for example the model proposed by Diggle and Kenward (1994)

$$\text{logit}(\Pr(r_{ij} = 1)) = \beta_0 + \beta_1 time_j + \beta_2 y_{ij-1}. \tag{2.1}$$

Thus, the probability that subject i drops out at time j given that it was observed at time $j - 1$ is modelled as a function of the time and the previous measurement through the logit link function.

2.4 Hierarchically structured data

The presence of missing values in multilevel data structures has been discussed in the literature (Gibson and Olejnik, 2003). In multilevel datasets, it is possible to have data missing at more than one level (Gibson and Olejnik, 2003). However, it is more problematic for data analysis, when a unit is missing at a higher level, because it implies that the data at lower level is also missing. Snijders and Bosker (1993) argued that even a small proportion of missing values at a higher level may lead to a loss of a lot of information on individuals at the lower level. Gibson and Olejnik (2003) added that methods for treating these missing data could alleviate the problem. Although the focus here is on missing values for the repeated measures data structure and less on missing data at higher levels, the basic definitions are unaffected by subjects being attributed to clusters. Models for missing data such as (2.1) can be extended to clustered data by adding random effects to represent heterogeneity between clusters.

3 Example: Somatic Cell Count Data

The scc40 dataset of Dohoo *et al.* (2009, Chapter 31) is a small subset of a large mastitis dataset collected by Jens Agger and the Danish Cattle Organization in 1993-94. It contains 13,487 non-missing observations at the first 10 time points (of the lactation) for 2,172 cows from 40 herds. Milk samples from each lactating cow were collected approximately monthly within the regular milk control scheme. Only records from a single lactation for each cow were included, and when the study period spanned parts of two lactations for a cow, the longer period of the two was selected. A binary indicator of intra-mammary infection or mastitis was obtained by dichotomizing the somatic cell counts at 200 000 cells/ml.

The scc40 dataset contains three types of missingness pattern: delayed entry, drop-outs and intermittent missing values. In general, a delayed entry occurs if a subject enters the study or becomes under observation after the start time of the study. For example, if time is measured relative to a fixed time point, subjects physically arriving after that point to an open study cohort (Dohoo *et al.*, 2009, Chapter 8) are delayed in their entry. For the scc40 data, each cow's time refers to the days since calving ("days in risk"). In this situation, a delayed entry occurs if the calving event took place outside (before) the study period, and the time points within a cow prior to study onset were considered as missing values. A drop-out occurs when a cow exited from the study before ending its intended measurements, whereas, intermittent missing values are occasions where a cow missed some measurements but reappeared again for later measurements in the study.

3.1 Analysis of the missing data in the scc40 dataset

In the context of the scc40 dataset, let y_{ijk} refer to complete binary records on each of n cows ($i = 1, \dots, n$) distributed on m herds ($k = 1, \dots, m$) at t time points ($j = 1, \dots, t$). Furthermore, let r_{ijk} be the indicator of y_{ijk} being missing. A shadow matrix was constructed for the corresponding full dataset, and the distribution of the missing values was explored. The total percentage of missing values in the constructed shadow matrix was about 31%, distributed as 17% delayed entry, 14% drop-out and 0.3% intermittent missing values. We will now detail the modelling for each type of missing values.

3.1.1 Missing values caused by drop-outs

A matrix of binary indicators of drop-outs was constructed according to the approach described earlier (2.3). Subjects with delayed entry were included only from their point of entry. Conditional on herd random effects, the probability that cow i in herd k drops out at time j was modelled by the random effects extension of Equation (2.1) based on an MAR process

$$\text{logit}(\Pr(r_{ijk} = 1|v_k)) = \beta_0 + \beta_1 time_j + \beta_2 y_{ij-1k} + v_k, \tag{3.1}$$

where $(v_1, \dots, v_m)$ are normally distributed independent random variables, say $v_k \sim N(0, \sigma_h^2)$ where σ_h^2 represents the heterogeneity (variance) between herds. Inclusion of a second order time lag (y_{ij-2k}) as well as a quadratic term for the effect of time were explored, but not considered of significance for the modelling.

3.1.2 Missing values caused by delayed entry

A matrix of binary indicators of missing values prior to entry was constructed from the shadow matrix. Each row consists of a series of 1's until the subject is observed (represented by a 0) for the first time in the study. Subsequent observations for the subject are not included. This data structure is similar to the structure for drop-outs, except that 0's and 1's are reversed.

This type of missing values is most likely a result of issues not related to the observed (or unobserved) values. Therefore it was modelled by an MCAR process. Then, the conditional probabilities were modelled by a random effects logistic regression model incorporating only time effects (by linear and quadratic terms)

$$\text{logit}(\Pr(r_{ijk} = 1|v_k)) = \beta_0 + \beta_1 time_j + \beta_{12} time_j^2 + v_k, \qquad (3.2)$$

with similar random effects assumptions as above. Note that the fixed and random terms in model (3.2) are different from those in model (3.1) as well as the forthcoming model (3.3); for simplicity of notation we retain the same symbols.

3.1.3 Intermittent missing values

The times of the first and last observation for each subject were excluded in the data for intermittent missing values. Each subject could have multiple missing values, either following each other or at isolated time points. Therefore, the MAR process model in Equation (3.1) was further extended to include cow random effects. In addition, the observed value at the previous time point could legitimately be missing, leading to the inclusion of an extra parameter in the model. In summary, the conditional probability that cow i in herd k has an intermittent missing value at time j given the cow and herd random effects (u_{ik}) and (v_k), respectively, was modeled by a random effects logistic regression model of the form

$$\text{logit}(\Pr(r_{ijk} = 1|v_k, u_{ijk})) = \beta_0 + \beta_1 time_j + \beta_2 y_{ij-1k} + \beta_3 r_{ij-1k} + u_{ijk} + v_k, \qquad (3.3)$$

for independent random variables $u_{ijk} \sim N(0, \sigma_c^2)$ and $v_k \sim N(0, \sigma_h^2)$ with the variances σ_c^2 and σ_h^2 representing the heterogeneity (variance) between cows and herds, respectively.

4 Statistical Methods

4.1 Estimation procedures

Random effects and marginal estimation procedures were selected based on their performance in the full and balanced simulated datasets (Masaoud and Stryhn, 2020). Random effects estimation procedures included several approximation algorithms, aimed at producing estimates close to the global ML estimate without actually computing the likelihood function (Breslow, 2003). These algorithms carry a number of different names and acronyms typically involving “weighted least squares” and “quasi”- or “pseudo-likelihood”.

Estimation in the forthcoming model (4.1) by numerical approximation most commonly employs the Gauss-Hermite quadrature procedure. Adaptive quadrature (Rabe-Hasketh *et al.*, 2002) is preferable for normally distributed random effects. In adaptive quadrature, the quadrature points are rescaled and shifted to the shape of the log likelihood function. In model (4.1), however, the added random effects at the cluster level pose some challenges for the direct maximization of the log likelihood (ML) and the integration becomes difficult (Diggle *et al.*, 2002) and may substantially increase computation time.

Estimation by Markov chain Monte Carlo (MCMC) techniques in a Bayesian framework, may be viewed as a numerical approach to avoid the computational difficulties of the log likelihood. In this study MCMC techniques are used as an estimation algorithm for the frequentist model rather than for exploring the genuine Bayesian models with informative prior distributions. The MCMC approach has been shown to perform well across a range of settings (Browne and Draper, 2006) and (Masaoud and Stryhn, 2010). Breslow and Clayton (1993) presented two estimation procedures based on quasi-likelihood function called penalized quasi-likelihood (PQL) and marginal quasi-likelihood (MQL). Both estimation procedures allow for an extra binomial dispersion parameter (ϕ), see (Skrondal and Rabe-Hesketh, 2007) for discussion of ϕ in PQL. MQL estimates are derived under random effects model assumptions (Goldstein, 1991). Both procedures iteratively employ linear mixed model estimation to an "adjusted" variate obtained by Taylor approximation of the outcome around its current estimated mean, until convergence, using either maximum likelihood (ML) or restricted maximum likelihood (REML), thus results in iterative generalized least squares (IGLS) or restricted iterative generalized least squares (RIGLS), respectively. One major difference between the two algorithms is that MQL does not incorporate the random effects u_i in the linearization of the mean (Molenberghs and Verbeke, 2005, Chapter 9) whereas PQL does. It has been also suggested to refine the approximations by including a second-order term in the Taylor expansions, usually denoted as second order PQL and MQL procedures (Goldstein and Rasbash, 1996) and (Rodriguez and Goldman, 1995). It is well-known that caution should be exercised in using these algorithms because under certain conditions they are prone to bias towards the null (e.g., (Rodriguez and Goldman, 1995) and (Rodriguez and Goldman, 2001)). The random effects procedures used the first order MQL and second order PQL procedures, with REML (RIGLS) option and implemented in the `MLwiN` software (version 2.02), and additionally we included version of PQL procedures with an extra binomial dispersion parameter, denoted as PQLx. The Bayesian estimation procedures were implemented in `WinBUGS` version 1.4 called from `the R software` using `the R2WinBUGS package` (Sturtz *et al.*, 2005). Vague ("non-informative") prior distributions (i.e. N$(0, 10^6)$) were used for all fixed effects parameters. The uniform distribution for inverse variances, or precisions ($\tau \sim$ uniform(0,100)) was used (Lambert *et al.*, 2005) and (Gelman, 2006). The Markov chains were run with 500 burn-in samples (Browne and Draper, 2006), and the subsequent estimates (posterior distribution medians) were based on 2000 samples. These burn-in and estimation sample sizes were arrived at after inspecting MCMC diagnostics for selected datasets.

Marginal estimation procedures included GEE, generalized estimating equations, and some of its variants; for more details, see (Masaoud and Stryhn, 2020). For missing data scenarios involving drop-outs by an MAR process, a weighted generalized estimating equation (WGEE) procedure was employed to account for the bias induced by the MAR mechanism. A GEE procedure may allow an MAR process to be ignored if the working correlation structure is specified correctly (Liang and Zeger, 1986; Jansen *et al.*, 2006); see however (Preisser *et al.*, 2002) for examples where this does not hold. The GEE procedure was set up with either an independence or exchangeable working correlation structure at the cluster (herd) level; results from (Masaoud and Stryhn, 2020) showed that GEE with these correlations at the cluster level performed well for balanced repeated measures data with an additional hierarchical structure. The calculations involved in the weighting scheme have been detailed elsewhere (Jansen *et al.*, 2006; Molenberghs and Verbeke, 2005, Chapter 27). In brief, the weight for each subject was calculated by fitting a marginal logistic regression for the binary indicators of drop-outs similar to (3.1). The differences were: time being modelled as a categorical predictor instead of a linear term, all fixed effects predictors being included, and the random effects being replaced by an exchangeable GEE working correlation structure. The predicted values from this model were used to compute weights for each subject and time point for the actual WGEE analysis, as the inverse probabilities of not dropping out up to the current time point. The weighting procedure and analysis were implemented using `SAS` software, by modifying the `SAS` code of Jansen *et al.* (2006) to facilitate looping across the simulated datasets. An ordinary logistic regression (OLR) with robust ("sandwich") variance estimates is included, that has been reported to work well for data comprising at least 30 subjects (Ziegler *et al.*, 1998). An alternative variant of the GEE procedure is alternating logistic regression (ALR). It uses the same estimating equation for the fixed effects as GEE, but differs from GEE by modeling the association among responses (e.g., within subjects) in terms of odds ratios. ALR is numerically more efficient than GEE for large clusters (Carey *et al.*, 1993). The ALR procedure has the advantage of providing standard errors for the association parameters. Furthermore, ALR allows one to distinguish between odds-ratios within clusters and within subclusters (in the current case subjects); however, the within–subject correlation must be modelled as exchangeable. An adapted ALR macro for 3-level of clustering (Kunthel *et al.*, 2014) is available when estimation of the association structure is of primary interest, though it was not included in our simulation study. For two-level binary repeated measures data, both GEE with an exchangeable correlation structure and ALR yield asymptotically unbiased estimates, which can be nearly efficient relative to GEE with a correctly specified working correlation structure (Masaoud and Stryhn, 2010) and to maximum-likelihood estimates in a fully and correctly specified model (Diggle *et al.*, 2002, Chapter 8).

4.2 Simulation procedures

In this simulation approach, the balanced full datasets were generated first. Then the desired missing data values were generated from a specified model, and the actual outcome values were replaced by their counterpart missing values. The whole process was repeated $N = 1000$ times. All full datasets were balanced with 8 time points, 20 subjects per cluster and 30 clusters. A total of five scenarios of missingness datasets were included. The scc40 scenario included all types of missing values present in the scc40 dataset. As described previously, about half of the missing values were due to delayed entry which could be argued to be assumed missing completely at random. In order to study the impact of scenarios with higher proportions of values missing that were not as a result of an MCAR process, missing values consisting exclusively of drop-outs were constructed. The drop-out missing values were modelled by either MAR or NMAR processes and were adjusted to either low (L) (approx. 31%) or high (H) (approx. 52%) proportions of missing values (designated as MARL/MARH and NMARL/NMARH).

4.2.1 Simulated balanced full datasets

The following random effects true models with autocorrelated ($\rho = 1$, 0.9 or 0.5) subject-specific random effects were used to generate the balanced full datasets.

$$\text{logit}(\Pr(y_{ijk} = 1|v_k, u_{ijk})) = \beta_0 + \beta_1 x_{1ijk} + \ldots + \beta_p x_{pijk} + u_{ijk} + v_k, \tag{4.1}$$

where y_{ijk} is a binary records on each of n subjects ($i = 1, \ldots, n$) distributed on m clusters ($k = 1, \ldots, m$) at t time points ($j = 1, \ldots, t$), as well as a set $x_1, \ldots, x_p$ of explanatory variables at different hierarchical levels recorded at every time point. The $(v_1, \ldots, v_m)$ are independent random variables with the same distribution and $(u_{i1k}, \ldots, u_{itk})$ are a series of autocorrelated random effects with $\rho(u_{ijk}, u_{ij'k}) = \rho^{|j-j'|}$. The most commonly assumed distribution is the Gaussian (normal), say $u_{ijk} \sim N(0, \sigma_2^2)$ where σ_2^2 represents the heterogeneity (variance) between subjects and $v_k \sim N(0, \sigma_3^2)$ where σ_3^2 represents the heterogeneity (variance) between clusters. Model (4.1) is for the conditional probability of an "event" given the random effects v_k and u_{ijk} of the kth cluster and of the ith subject at jth time point, respectively.

The simulation settings of the balanced full datasets were motivated by the scc40 dataset of (Dohoo *et al.*, 2009, Chapter 8) for repeated measures of binary records of intra-mammary infection or mastitis in milk samples from cows housed in multiple herds. In the scc40 context, predictors of interest existed at both the herd and cow levels; thus, the simulation design included binary covariates at the cluster and subject levels. Including also (for simplicity) a linear time effect but no interactions with time, the linear predictor included the following parameters set at the indicated true values: the intercept centered at first time point $(\beta_0) = -1$; the slope for time $= 0, \ldots, t-1$ $(\beta_1) = 0.15$; the coefficient for subject level covariates $(\beta_2) = -1$ and the coefficient for cluster level covariate $(\beta_3) = 1$.

The random part of the model included normally distributed subject and cluster level random effects with standard deviations set at $\sigma_2 = 1.5$ and $\sigma_3 = 0.75$, respectively. These values approximated the estimates in a random intercept model for a binary outcome in the scc40 dataset, as described in section 3.

By the latent variable approximation to the variance partition coefficient (Goldstein *et al.*, 2002), this corresponds to 37% and 9% of the unexplained variance residing at the subject and cluster levels, respectively. Simulated datasets were generated for highly and moderately autocorrelated subject-specific random effects ($\rho = 0.9$ and $\rho = 0.5$) as well as for a random intercept model ($\rho = 1$). Note that the correlation between binary outcomes is different than the correlation between the random effects. In particular, the latent variable approximation with an observation-level variance component of $\pi^2/3$ (Snijders and Bosker, 2012, Chapter 14) yields an intra-class correlation of $\sigma^2/(\sigma^2 + \pi^2/3) = 0.46$, where $\sigma^2 = \sigma_2^2 + \sigma_3^2$, and a first-order correlation of $\rho\sigma^2/(\sigma^2 + \pi^2/3)$, and the values 0.42 and 0.23 for $\rho = 0.9, 0.5$, respectively.

The autocorrelated random effects of each subject were generated by multiplying a vector of t independent variables by the upper triangular factor of the Cholesky decomposition of the correlation matrix (as described in Congdon (2003, Chapter 1)). Generation of the binary outcomes then followed the usual scheme for random effects logistic regrssion models (Stryhn *et al.*, 2000). A comprehensive and detailed analysis of the full datasets appeared in (Masaoud and Stryhn, 2020).

4.2.2 Missing values: scc40 scenario

The three types of missing values were simulated in the following order: delayed entry based on model (3.2), drop-outs based on model (3.1), and intermittent missing values based on model (3.3). The parameter estimates of these models for the scc40 data (Table 1) were taken as true values for the simulations of the missing value patterns.

4.2.3 Missing at random scenarios: MARL and MARH

The scc40 regression estimates (Table 1) for the drop-out coefficients in model (3.1) were retained except that a stronger dependence on the previous value was imposed. Specifically, we used $\beta_0 = -4.7$, $\beta_1 = 0.35$ and $\sigma_h = 0.068$, and the coefficient for the previous value was set at either $\beta_2 = 2$ (MARL) or $\beta_2 = 4$ (MARH). Overall, this produced expected percentages of missing values of approximately 31% (about the same overall level as the scc40 data) and 52%, respectively. The expected percentages of missing values ranged from 6% and 19% at the second time point to 70% and 85% at the last time point, for MARL and MARH respectively.

4.2.4 Not missing at random scenarios: NMARL and NMARH

Although this study does not include methods to estimate NMAR models, data could be generated from a NMAR scenario by directly allowing the probability of a missing value to depend on the actual value from the full dataset. For simplicity, we used model (3.1) with the previous outcome replaced by the current outcome and the same parameters as for the MAR scenarios. This resulted in overall percentages of missing values of 31% and 52% and similar ranges of percentages at individual time points as for MAR.

4.3 Analysis of results for simulated data

The estimates of marginal or random effects estimation procedures under different scenarios were compared both to the true values of the simulation and to the estimates obtained from the full simulated datasets. The latter comparison was of interest for studying the impact of missing data on the performance of the estimation procedures, whereas the former comparison would be used for an overall assessment of each procedure under specific scenarios. The comparison of estimates to the true values used the same formulae and methods as the analysis of the balanced full data (Masaoud and Stryhn, 2020). In brief, the relative bias was defined as difference between the average estimate among simulations ($\hat{\beta}$) and the, marginal or subject-specific, true value (β), divided by the true value,

$$\text{relative bias to true value (RBT)} = \frac{\hat{\beta}_M - \beta}{\beta} \times 100\%. \tag{4.2}$$

Note that $\hat{\beta}_M$ refers to the estimate based on the incomplete data. The scaling by the true value was useful because the parameters were not standardized to a uniform scale. In a similar fashion, the relative bias to the average estimate based on the full data ($\hat{\beta}_F$) was defined as

$$\text{relative bias to full data (RBF)} = \frac{\hat{\beta}_M - \hat{\beta}_F}{\beta} \times 100\%. \qquad (4.3)$$

One could also use $\hat{\beta}_F$ in the dominator of (4.3); one advantage of our simpler form is that the RBF is obtained as the difference of the RBTs for the full and incomplete data. Only datasets where valid estimates were obtained by both full and incomplete data were included. For any of the estimates (of both fixed effects and variance parameters), the presence of statistically significant bias compared with the full data was assessed by a t-test based on the differences between estimates obtained from the full and incomplete datasets among the simulations.

5 Results

After a brief review of the parameter estimates (Table 1) obtained from analyses of the three different types of missing values (see section 3.1) in the scc40 dataset, the results are presented subdivided by the true model data (random intercept or autocorrelated random effects model) and the missing value scenarios. As the main interest is in the impact of the missing values, the focus here is on the relative bias to the full data (RBF) in Tables 2–5, and we defer relative biases and standard errors to the true values (RBT) to an appendix (Appendix A, Tables A1–A5). The coverages of confidence intervals are shown in Figures 1–3; these must necessarily refer to the true values. The performance of estimation procedures for the corresponding full datasets was discussed previously (Masaoud and Stryhn, 2020) and includes, briefly, minor attenuation of variance estimates at the cluster level for random effects procedures in random intercept model data and strong downwards biases for all random effects procedures in autocorrelated data, as well as a small negative relative bias by marginal estimation procedures in both data settings

5.1 Missingness types for scc40 data

The strongest effects on patterns of missingness in the scc40 data were found for drop-outs (Table 1). The likelihood of a subject dropping out increased significantly both with time (OR $= 1.42$ per month) and with the previous value being an event (OR $= 1.25$). The estimated probabilities of a subject with no events to drop out increased from 1.5% at the second time point to 15% at the last time point ($t = 9$). There was little between-herd variation in the occurrence of drop-outs.

The probability of a delayed entry also depended strongly on time, but in a non-linear fashion (Table 1). The negative quadratic term ensures the likelihood of a delayed entry missing value to decrease as time progresses; in the data, all missing value series eventually stop because otherwise the subject would not be part of the dataset. The estimated proportion of non-delayed subjects (with $r_{i1k} = 0$) was 46.6%, slightly above the 42.4% in the scc40 data. The herd-level variation in delayed entries was very small, but statistically significant.

The probability of intermittent values declined with time (OR $= 0.82$ per month) and depended on the previous observation being an event (OR $= 0.50$); both of these associations were quite uncertain (Table 1), a consequence of the small number (0.3%) of intermittent values in the scc40 dataset. Variances at the cow and herd levels were estimated at moderate values but were however not statistically significant.

Table 1: Random effects logistic regression estimates of fixed effects and variances, with standard errors, from analyses for three different types of missing values in the scc40 dataset; interpretation of parameters: β_0 = intercept, β_1 = time coefficient, β_{12} = quadratic term for time coefficient, β_2 = previous outcome, β_3 = previous outcome missing, σ_h^2 = herd-level variance, σ_c^2 = cow-level variance.

	Type of missing values					
	Delayed entry		Drop-out		Intermittent	
Parameter	Est.	SE	Est.	SE	Est.	SE
β_0	−0.444	0.083	−4.850	0.143	−4.582	0.604
β_1	0.666	0.055	0.350	0.019	−0.196	0.075
β_{12}	−0.084	0.007				
β_2			0.224	0.072	−0.698	0.347
β_3					1.421	0.999
σ_h^2	0.017	0.011	0.068	0.026	0.295	0.257
σ_c^2					0.938	1.008

5.2 Random intercept model data ($\rho = 1$)

5.2.1 Missing values: scc40 scenario

All estimation procedures gave estimates in fairly close agreement with those of the full datasets (Table 2). Small but significant negative biases for the time coefficient (β_1) were found for OLR and PQL. The variance estimates from PQL, PQLx and MCMC showed some minor negative and positive biases that in all cases were in the same direction as the bias in the estimates of the full data (Appendix A, Table A1)

Table 2: Relative bias of estimates to the full data (RBF) with a significance indication and standard error in parenthesis, based on analyses of 1000 simulated datasets generated by random intercept model ($\rho = \mathbf{1}$) in five simulated scenarios of missing values: scc40 (missing values as in scc40 dataset), MARL, MARH (low (31%) and high (52%) proportion of missing values at random due to drop-outs), NMARL, NMARH (low (31%) and high (52%) proportion of missing values not at random). Parameters: β_0 (intercept), β_1 (time coefficient), β_2 (subject level factor), β_3 (cluster level factor), σ_2^2 (variance at subject level), σ_3^2 (variance at cluster level), ϕ (extra-binomial dispersion). Estimation procedures: OLR (ordinary logistic regression), ALR (alternating logistic regression), PQL (2nd order penalized quasi-likelihood), PQLx (2nd order penalized quasi-likelihood with extra-binomial dispersion), ML (maximum likelihood), MCMC (Bayesian Markov chain Monte Carlo).

Scen-	Param-	Statistical Methods					
ario	eter	OLR	ALR	PQL	PQLx	ML	MCMC
scc40	β_0	-1.7 (1.2)	-0.9 (1.2)	-1.4 (1.2)	0.0 (1.3)	-0.8 (1.3)	-0.6 (1.3)
	β_1	$-2.9^{\ddagger}$(0.4)	-0.6 (0.4)	$-1.7^{\ddagger}$(0.6)	0.0 (0.6)	-0.8 (0.6)	-0.8 (0.6)
	β_2	-0.6 (0.7)	-0.6 (0.6)	-1.1 (0.6)	-0.2 (0.7)	-0.5 (0.7)	-0.4 (0.7)
	β_3	0.1 (1.4)	0.3 (1.3)	-0.2 (1.4)	0.9 (1.4)	0.5 (1.4)	0.7 (1.4)
	σ_2^2			$-2.2^{\ddagger}$(0.4)	$4.8^{\ddagger}$(0.5)	1.0 (0.5)	$1.5^{\dagger}$(0.5)
	σ_3^2			$-3.8^{\ddagger}$(1.4)	-2.2 (1.5)	-2.9 (1.5)	-3.3 (1.7)
	ϕ				$-2.9^{\ddagger}$(0.1)		
MARL	β_0	$-8.0^{\ddagger}$(1.3)	-0.3 (1.3)	-1.0 (1.3)	1.9 (1.3)	0.2 (1.3)	2.5 (1.7)
	β_1	$-52.1^{\ddagger}$(0.5)	0.1 (0.4)	$-2.9^{\ddagger}$(0.6)	$10.1^{\ddagger}$(0.7)	0.1 (0.6)	1.2 (0.7)
	β_2	$-3.4^{\ddagger}$(0.7)	-0.7 (0.7)	$-1.4^{\dagger}$(0.7)	0.6 (0.7)	-1.0 (0.7)	-0.4 (0.8)
	β_3	$-3.6^{\ddagger}$(1.4)	-0.9 (1.4)	-1.3 (1.4)	0.8 (1.5)	-0.8 (1.4)	1.2 (1.7)
	σ_2^2			$-3.5^{\ddagger}$(0.4)	$8.0^{\ddagger}$(0.5)	0.8 (0.5)	$2.6^{\ddagger}$(0.6)
	σ_3^2			-2.4 (1.4)	1.5 (1.5)	-1.4 (1.5)	-1.4 (2.0)
	ϕ				$-5.0^{\ddagger}$(0.1)		
MARH	β_0	-1.5 (1.2)	$-5.3^{\ddagger}$(1.3)	$-3.3^{\ddagger}$(1.3)	$10.9^{\ddagger}$(1.5)	0.4 (1.4)	1.0 (1.4)
	β_1	$-140.1^{\ddagger}$(0.6)	$23.0^{\ddagger}$(0.5)	$-22.6^{\ddagger}$(0.8)	$89.2^{\ddagger}$(1.2)	0.5 (0.9)	$3.5^{\ddagger}$(0.9)
	β_2	$-11.6^{\ddagger}$(0.7)	0.0 (0.7)	$-4.1^{\ddagger}$(0.7)	$11.6^{\ddagger}$(0.8)	-1.0 (0.7)	-0.4 (0.8)
	β_3	$-11.4^{\ddagger}$(1.3)	0.0 (1.4)	$-3.6^{\ddagger}$(1.4)	$12.4^{\ddagger}$(1.6)	-0.6 (1.5)	0.3 (1.5)
	σ_2^2			$-17.4^{\ddagger}$(0.5)	$64.0^{\ddagger}$(1.0)	$1.5^{\dagger}$(0.7)	$4.0^{\ddagger}$(0.7)
	σ_3^2			$-7.5^{\ddagger}$(1.4)	$22.8^{\ddagger}$(1.8)	-1.9 (1.6)	-1.6 (1.7)
	ϕ				$-20.4^{\ddagger}$(0.2)		
NMARL	β_0	$-7.7^{\ddagger}$(1.3)	-0.5 (1.3)	$-2.6^{\dagger}$(1.3)	-0.1 (1.3)	-0.7 (1.3)	-0.1 (1.4)
	β_1	$-88.2^{\ddagger}$(0.5)	$-53.0^{\ddagger}$(0.4)	$-79.6^{\ddagger}$(0.6)	$-75.0^{\ddagger}$(0.6)	$-78.4^{\ddagger}$(0.6)	$-78.2^{\ddagger}$(0.6)
	β_2	$-2.0^{\ddagger}$(0.7)	-0.4 (0.7)	$-1.8^{\ddagger}$(0.7)	0.0 (0.7)	$-1.5^{\dagger}$(0.7)	-1.4 (0.7)
	β_3	-2.1 (1.4)	-0.5 (1.4)	-1.4 (1.4)	0.5 (1.5)	-1.2 (1.5)	-0.6 (1.5)
	σ_2^2			$-2.6^{\ddagger}$(0.4)	$8.5^{\ddagger}$(0.5)	0.3 (0.5)	0.9 (0.5)
	σ_3^2			-2.4 (1.4)	1.0 (1.5)	-2.2 (1.5)	-2.5 (1.7)
	ϕ				$-6.5^{\ddagger}$(0.1)		
NMARH	β_0	$11.0^{\ddagger}$(1.3)	$16.1^{\ddagger}$(1.3)	$11.5^{\ddagger}$(1.3)	$23.5^{\ddagger}$(1.5)	$13.8^{\ddagger}$(1.4)	$14.5^{\ddagger}$(1.4)
	β_1	$-317.8^{\ddagger}$(0.9)	$-223.5^{\ddagger}$(0.9)	$-318.3^{\ddagger}$(1.1)	$-300.2^{\ddagger}$(1.2)	$-318.6^{\ddagger}$(1.1)	$-319.0^{\ddagger}$(1.1)
	β_2	$-4.3^{\ddagger}$(0.8)	0.9 (0.8)	$-5.6^{\ddagger}$(0.7)	$5.0^{\ddagger}$(0.8)	$-6.2^{\ddagger}$(0.8)	$-5.7^{\ddagger}$(0.8)
	β_3	$-4.9^{\ddagger}$(1.4)	0.4 (1.5)	$-5.4^{\ddagger}$(1.4)	$5.7^{\ddagger}$(1.6)	$-6.4^{\ddagger}$(1.5)	$-5.7^{\ddagger}$(1.5)
	σ_2^2			$-14.4^{\ddagger}$(0.6)	$50.4^{\ddagger}$(1.0)	$-11.2^{\ddagger}$(0.7)	$-9.6^{\ddagger}$(0.7)
	σ_3^2			$-8.3^{\ddagger}$(1.4)	$12.5^{\ddagger}$(1.7)	$-10.1^{\ddagger}$(1.5)	$-11.1^{\ddagger}$(1.7)
	ϕ				$-21.5^{\ddagger}$(0.3)		

† significant bias at $P < 0.05$; ‡ significant bias at $P < 0.01$

Table 3: Relative bias of estimates to the full data (RBF) with a significance indication and standard error in parenthesis, based on analyses of 1000 simulated datasets generated by autoregressive random effects model with ($\rho = \mathbf{0.9}$) in five simulated scenarios of missing values: scc40 (missing values as in scc40 dataset), MARL, MARH (low (31%) and high (52%) proportion of missing values at random due to drop-outs), NMARL, NMARH (low (31%) and high (52%) proportion of missing values not at random). Parameters: β_0 (intercept), β_1 (time coefficient), β_2 (subject level factor), β_3 (cluster level factor), σ_2^2 (variance at subject level), σ_3^2 (variance at cluster level), ϕ (extra-binomial dispersion). See Table 2 for coding of estimation procedures.

Scen-	Param-	Statistical Methods					
ario	eter	OLR	ALR	PQL	PQLx	ML	MCMC
scc40	β_0	$-2.2^{\dagger}$(1.2)	-1.5 (1.2)	$3.3^{\dagger}$(1.2)	$4.9^{\ddagger}$(1.2)	$4.2^{\ddagger}$(1.2)	$4.1^{\ddagger}$(1.2)
	β_1	$-3.2^{\ddagger}$(0.5)	$-1.0^{\dagger}$(0.4)	$3.5^{\ddagger}$(0.6)	$5.5^{\ddagger}$(0.6)	$4.3^{\ddagger}$(0.6)	$4.4^{\ddagger}$(0.6)
	β_2	0.0 (0.7)	-0.1 (0.6)	$4.0^{\ddagger}$(0.6)	$-3.0^{\ddagger}$(0.6)	$5.5^{\ddagger}$(0.6)	$5.7^{\ddagger}$(0.6)
	β_3	-0.9 (1.3)	-1.0 (1.4)	$3.2^{\dagger}$(1.3)	$4.5^{\ddagger}$(1.3)	$4.6^{\ddagger}$(1.3)	$4.5^{\ddagger}$(1.3)
	σ_2^2			$17.8^{\ddagger}$(0.3)	$25.9^{\ddagger}$(0.4)	$22.8^{\ddagger}$(0.4)	$23.4^{\ddagger}$(0.4)
	σ_3^2			$5.4^{\ddagger}$(1.3)	$7.8^{\ddagger}$(1.3)	$8.0^{\ddagger}$(1.3)	$8.6^{\ddagger}$(1.5)
	ϕ				$-6.9^{\ddagger}$(0.1)		
MARL	β_0	$-8.6^{\ddagger}$(1.2)	-1.4 (1.2)	$2.9^{\dagger}$(1.2)	$5.9^{\ddagger}$(1.2)	$4.3^{\ddagger}$(1.2)	$4.0^{\ddagger}$(1.2)
	β_1	$-47.5^{\ddagger}$(0.5)	0.0 (0.4)	$2.5^{\ddagger}$(0.6)	$14.4^{\ddagger}$(0.7)	$5.2^{\ddagger}$(0.6)	$5.7^{\ddagger}$(0.6)
	β_2	$-2.2^{\ddagger}$(0.6)	0.0 (0.6)	$3.9^{\ddagger}$(0.6)	$-2.2^{\ddagger}$(0.6)	$5.3^{\ddagger}$(0.6)	$5.6^{\ddagger}$(0.6)
	β_3	$-4.0^{\ddagger}$(1.3)	-1.7 (1.4)	2.3 (1.3)	$4.6^{\ddagger}$(1.3)	$3.8^{\ddagger}$(1.3)	$3.5^{\ddagger}$(1.3)
	σ_2^2			$16.4^{\ddagger}$(0.4)	$28.2^{\ddagger}$(0.4)	$22.4^{\ddagger}$(0.4)	$23.1^{\ddagger}$(0.4)
	σ_3^2			$6.1^{\ddagger}$(1.2)	$10.3^{\ddagger}$(1.3)	$8.8^{\ddagger}$(1.3)	$9.4^{\ddagger}$(1.4)
	ϕ				$-8.8^{\ddagger}$(0.1)		
MARH	β_0	-0.4 (1.2)	$-3.3^{\ddagger}$(1.2)	1.0 (1.2)	$11.6^{\ddagger}$(1.3)	$3.8^{\ddagger}$(1.2)	$3.8^{\ddagger}$(1.2)
	β_1	$-113.1^{\ddagger}$(0.6)	$32.8^{\ddagger}$(0.6)	$-6.4^{\ddagger}$(0.9)	$84.8^{\ddagger}$(1.3)	$13.9^{\ddagger}$(1.0)	$16.6^{\ddagger}$(1.0)
	β_2	$-8.8^{\ddagger}$(0.6)	$1.7^{\ddagger}$(0.6)	0.0 (0.6)	$6.0^{\ddagger}$(0.7)	$3.4^{\ddagger}$(0.6)	$3.9^{\ddagger}$(0.6)
	β_3	$-10.3^{\ddagger}$(1.3)	0.0 (1.4)	-1.5 (1.3)	$12.7^{\ddagger}$(1.4)	1.8 (1.3)	1.8 (1.3)
	σ_2^2			$-3.4^{\ddagger}$(0.4)	$54.0^{\ddagger}$(0.9)	$11.6^{\ddagger}$(0.6)	$13.4^{\ddagger}$(0.6)
	σ_3^2			-1.5 (1.2)	$24.1^{\ddagger}$(1.5)	$4.2^{\ddagger}$(1.3)	$4.8^{\ddagger}$(1.4)
	ϕ				$-18.1^{\ddagger}$(0.2)		
NMARL	β_0	$-7.0^{\ddagger}$(1.2)	-2.1 (1.2)	-2.1 (1.1)	-0.2 (1.2)	-1.2 (1.1)	-1.3 (1.2)
	β_1	$-80.6^{\ddagger}$(0.5)	$-56.0^{\ddagger}$(0.5)	$-75.2^{\ddagger}$(0.6)	$-72.4^{\ddagger}$(0.7)	$-74.1^{\ddagger}$(0.6)	$-74.0^{\ddagger}$(0.6)
	β_2	-0.6 (0.6)	0.5 (0.6)	0.4 (0.6)	$-6.4^{\ddagger}$(0.6)	0.7 (0.6)	0.9 (0.6)
	β_3	-2.4 (1.3)	-1.3 (1.3)	-1.1 (1.3)	0.5 (1.3)	-0.9 (1.3)	-1.1 (1.3)
	σ_2^2			$2.1^{\ddagger}$(0.3)	$9.9^{\ddagger}$(0.4)	$4.6^{\ddagger}$(0.4)	$4.9^{\ddagger}$(0.4)
	σ_3^2			0.5 (1.2)	$3.1^{\dagger}$(1.2)	0.9 (1.2)	0.9 (1.3)
	ϕ				$-6.1^{\ddagger}$(0.1)		
NMARH	β_0	$13.0^{\ddagger}$(1.3)	$16.0^{\ddagger}$(1.3)	$14.9^{\ddagger}$(1.2)	$25.7^{\ddagger}$(1.3)	$16.3^{\ddagger}$(1.2)	$16.4^{\ddagger}$(1.2)
	β_1	$-299.6^{\ddagger}$(0.9)	$-232.1^{\ddagger}$(0.9)	$-301.1^{\ddagger}$(1.0)	$-291.2^{\ddagger}$(1.1)	$-300.7^{\ddagger}$(1.0)	$-301.0^{\ddagger}$(1.0)
	β_2	$-3.3^{\ddagger}$(0.7)	1.0 (0.7)	$-2.5^{\ddagger}$(0.6)	-1.3 (0.7)	$-2.0^{\ddagger}$(0.6)	$-1.6^{\dagger}$(0.6)
	β_3	$-4.4^{\ddagger}$(1.4)	-0.3 (1.4)	$-3.2^{\ddagger}$(1.3)	$6.6^{\ddagger}$(1.4)	$-3.1^{\dagger}$(1.3)	$-3.2^{\dagger}$(1.3)
	σ_2^2			$-1.4^{\ddagger}$(0.5)	$46.1^{\ddagger}$(0.8)	$3.0^{\ddagger}$(0.5)	$3.9^{\ddagger}$(0.5)
	σ_3^2			-1.9 (1.2)	$15.6^{\ddagger}$(1.5)	-1.6 (1.2)	-1.9 (1.4)
	ϕ				$-19.6^{\ddagger}$(0.3)		

† significant bias at $P < 0.05$; ‡ significant bias at $P < 0.01$

Table 4: Relative bias of estimates to the full data (RBF) with a significance indication and standard error in parenthesis, based on analyses of 1000 simulated datasets generated by autoregressive random effects model with (ρ = **0.5**) in five simulated scenarios of missing values: scc40 (missing values as in scc40 dataset), MARL, MARH (low (31%) and high (52%) proportion of missing values at random due to drop-outs), NMARL, NMARH (low (31%) and high (52%) proportion of missing values not at random). Parameters: β_0 (intercept), β_1 (time coefficient), β_2 (subject level factor), β_3 (cluster level factor), σ_2^2 (variance at subject level), σ_3^2 (variance at cluster level), ϕ (extra-binomial dispersion). See Table 2 for coding of estimation procedures.

Scen-	Param-	Statistical Methods					
ario	eter	OLR	ALR	PQL	PQLx	ML	MCMC
scc40	β_0	-1.3 (1.2)	-0.8 (1.2)	$5.9^{\ddagger}$(1.0)	$7.2^{\ddagger}$(1.0)	$6.0^{\ddagger}$(1.0)	$6.1^{\ddagger}$(1.0)
	β_1	$-1.9^{\ddagger}$(0.5)	-0.5 (0.5)	$6.1^{\ddagger}$(0.5)	$7.5^{\ddagger}$(0.5)	$6.2^{\ddagger}$(0.5)	$6.4^{\ddagger}$(0.5)
	β_2	0.6 (0.5)	0.7 (0.5)	$7.0^{\ddagger}$(0.4)	$8.1^{\ddagger}$(0.4)	$7.4^{\ddagger}$(0.4)	$7.5^{\ddagger}$(0.4)
	β_3	-0.5 (1.4)	-0.4 (1.3)	$6.0^{\ddagger}$(1.1)	$7.2^{\ddagger}$(1.1)	$6.4^{\ddagger}$(1.1)	$6.4^{\ddagger}$(1.1)
	σ_2^2			$19.9^{\ddagger}$(0.2)	$24.5^{\ddagger}$(0.2)	$21.9^{\ddagger}$(0.2)	$22.2^{\ddagger}$(0.2)
	σ_3^2			$9.2^{\ddagger}$(0.9)	$11.0^{\ddagger}$(0.9)	$9.6^{\ddagger}$(0.9)	$10.8^{\ddagger}$(1.0)
	ϕ				$-8.5^{\ddagger}$(0.1)		
MARL	β_0	$-6.0^{\ddagger}$(1.2)	-1.0 (1.2)	$5.8^{\ddagger}$(1.0)	$7.7^{\ddagger}$(1.0)	$6.1^{\ddagger}$(1.0)	$6.2^{\ddagger}$(1.0)
	β_1	$-31.6^{\ddagger}$(0.5)	-0.5 (0.5)	$5.3^{\ddagger}$(0.6)	$11.1^{\ddagger}$(0.6)	$5.7^{\ddagger}$(0.6)	$6.0^{\ddagger}$(0.6)
	β_2	-1.0 (0.5)	0.6 (0.5)	$7.1^{\ddagger}$(0.4)	$8.5^{\ddagger}$(0.4)	$7.3^{\ddagger}$(0.4)	$7.4^{\ddagger}$(0.4)
	β_3	-1.0 (1.3)	0.6 (1.3)	$7.2^{\ddagger}$(1.1)	$8.6^{\ddagger}$(1.1)	$7.3^{\ddagger}$(1.1)	$7.4^{\ddagger}$(1.1)
	σ_2^2			$23.0^{\ddagger}$(0.2)	$25.6^{\ddagger}$(0.2)	$22.1^{\ddagger}$(0.2)	$22.5^{\ddagger}$(0.2)
	σ_3^2			$10.5^{\ddagger}$(0.9)	$13.0^{\ddagger}$(0.9)	$10.8^{\ddagger}$(0.9)	$12.1^{\ddagger}$(1.0)
	ϕ				$-9.7^{\ddagger}$(0.1)		
MARH	β_0	1.5 (1.2)	2.2 (1.2)	$4.2^{\ddagger}$(1.0)	$9.4^{\ddagger}$(1.0)	$5.4^{\ddagger}$(1.0)	$5.4^{\ddagger}$(1.0)
	β_1	$-56.9^{\ddagger}$(0.7)	$40.1^{\ddagger}$(0.7)	$4.1^{\ddagger}$(0.8)	$42.9^{\ddagger}$(1.2)	$14.0^{\ddagger}$(0.9)	$12.8^{\ddagger}$(0.9)
	β_2	$-4.5^{\ddagger}$(0.5)	$2.8^{\ddagger}$(0.6)	$1.6^{\ddagger}$(0.4)	$8.1^{\ddagger}$(0.5)	$3.2^{\ddagger}$(0.5)	$3.0^{\ddagger}$(0.5)
	β_3	$-4.3^{\ddagger}$(1.3)	$2.9^{\dagger}$(1.3)	1.7 (1.1)	$8.3^{\ddagger}$(1.1)	$3.2^{\ddagger}$(1.1)	$3.2^{\ddagger}$(1.1)
	σ_2^2			$2.8^{\ddagger}$(0.2)	$17.4^{\ddagger}$(0.4)	$6.9^{\ddagger}$(0.3)	$6.2^{\ddagger}$(0.3)
	σ_3^2			$2.2^{\ddagger}$(0.8)	$12.4^{\ddagger}$(1.0)	$4.6^{\ddagger}$(0.9)	$4.9^{\ddagger}$(1.0)
	ϕ				$-10.3^{\ddagger}$(0.2)		
NMARL	β_0	$-2.8^{\dagger}$(1.2)	-1.0 (1.2)	-0.1 (0.9)	0.7 (1.0)	0.1 (1.0)	0.1 (0.9)
	β_1	$-65.7^{\ddagger}$(0.5)	$-55.4^{\ddagger}$(0.5)	$-62.6^{\ddagger}$(0.6)	$-62.2^{\ddagger}$(0.6)	$-62.1^{\ddagger}$(0.6)	$-62.0^{\ddagger}$(0.6)
	β_2	0.4 (0.5)	0.8 (0.5)	$1.2^{\ddagger}$(0.4)	$1.9^{\ddagger}$(0.4)	$1.3^{\ddagger}$(0.4)	$1.4^{\ddagger}$(0.4)
	β_3	0.6 (1.3)	1.0 (1.3)	1.3 (1.0)	2.1 (1.1)	1.4 (1.1)	1.5 (1.1)
	σ_2^2			$2.2^{\ddagger}$(0.1)	$4.4^{\ddagger}$(0.2)	$2.9^{\ddagger}$(0.2)	$2.8^{\ddagger}$(0.2)
	σ_3^2			$1.7^{\ddagger}$(0.8)	$2.8^{\ddagger}$(0.8)	$1.9^{\dagger}$(0.8)	$2.2^{\dagger}$(0.9)
	ϕ				$-3.4^{\ddagger}$(0.1)		
NMARH	β_0	$18.7^{\ddagger}$(1.2)	$19.8^{\ddagger}$(1.2)	$17.8^{\ddagger}$(1.0)	$23.3^{\ddagger}$(1.0)	$18.0^{\ddagger}$(1.0)	$18.0^{\ddagger}$(1.0)
	β_1	$-254.2^{\ddagger}$(0.9)	$-225.5^{\ddagger}$(0.9)	$-251.9^{\ddagger}$(1.0)	$-253.8^{\ddagger}$(1.0)	$-251.3^{\ddagger}$(1.0)	$-251.5^{\ddagger}$(1.0)
	β_2	$-1.2^{\dagger}$(0.6)	0.7 (0.6)	0.1 (0.5)	$4.2^{\ddagger}$(0.5)	0.6 (0.5)	0.5 (0.5)
	β_3	-1.4 (1.3)	1.0 (1.3)	0.3 (1.1)	$4.7^{\ddagger}$(1.1)	0.7 (1.1)	0.7 (1.1)
	σ_2^2			$4.2^{\ddagger}$(0.2)	$17.8^{\ddagger}$(0.4)	$5.6^{\ddagger}$(0.3)	$4.4^{\ddagger}$(0.3)
	σ_3^2			0.9 (0.9)	$7.4^{\ddagger}$(1.0)	1.5 (0.9)	1.7 (1.0)
	ϕ				$-11.4^{\ddagger}$(0.2)		

† significant bias at $P < 0.05$;‡ significant bias at $P < 0.01$

Table 5: Relative bias of estimates to the full data (RBF) with a significance indication and standard error in parenthesis, based on analyses of 1000 simulated datasets generated by autoregressive random effects model with ($\rho = 1, 0.9, 0.5$) in five simulated scenarios of missing values: scc40 (missing values as in scc40 dataset), MARL, MARH (low (31%) and high (52%) proportion of missing values at random due to drop-outs), NMARL, NMARH (low (31%) and high (52%) proportion of missing values not at random). Parameters: β_0 (intercept), β_1 (time coefficient), β_2 (subject level factor), β_3 (cluster level factor). Estimation procedures: WGEEci (weighted generalized estimating equations (WGEE) with independence correlation at cluster level), WGEEce (WGEE with exchangeable correlation at cluster level).

Scen-	Param-	correlation	$\rho = 1$		$\rho = 0.9$		$\rho = 0.5$	
ario	eter	procedure	WGEEci	WGEEce	WGEEci	WGEEce	WGEEci	WGEEce
MARL	β_0		0.1 (1.3)	7.1‡(1.5)	−1.5 (1.3)	6.4‡(1.5)	−0.8 (1.3)	14.4‡(1.4)
	β_1		1.7‡(0.6)	3.8‡(0.5)	1.4†(0.6)	3.0‡(0.6)	0.3 (0.6)	1.5†(0.6)
	β_2		−0.1 (0.8)	0.4 (0.8)	0.9 (0.7)	1.4†(0.7)	1.0 (0.6)	2.1‡(0.6)
	β_3		−0.6 (1.4)	1.3 (1.8)	−1.3 (1.4)	0.0 (1.7)	0.8 (1.4)	1.0 (1.6)
MARH	β_0		13.5‡(2.7)	−6.8†(3.3)	7.8‡(2.9)	−3.4 (3.3)	7.7‡(2.4)	8.5‡(2.7)
	β_1		−24.0‡(2.2)	−20.3‡(2.1)	−22.9‡(2.4)	−18.2‡(2.3)	−13.7‡(2.3)	−7.6‡(2.3)
	β_2		−0.3 (2.4)	−3.9†(1.7)	1.7 (2.3)	−3.0 (1.7)	2.9 (1.8)	1.5 (1.4)
	β_3		−1.7 (2.7)	−4.6 (3.9)	−1.6 (2.6)	−1.5 (3.5)	2.5 (2.2)	2.6 (3.0)

† significant bias at $P < 0.05$;‡ significant bias at $P < 0.01$

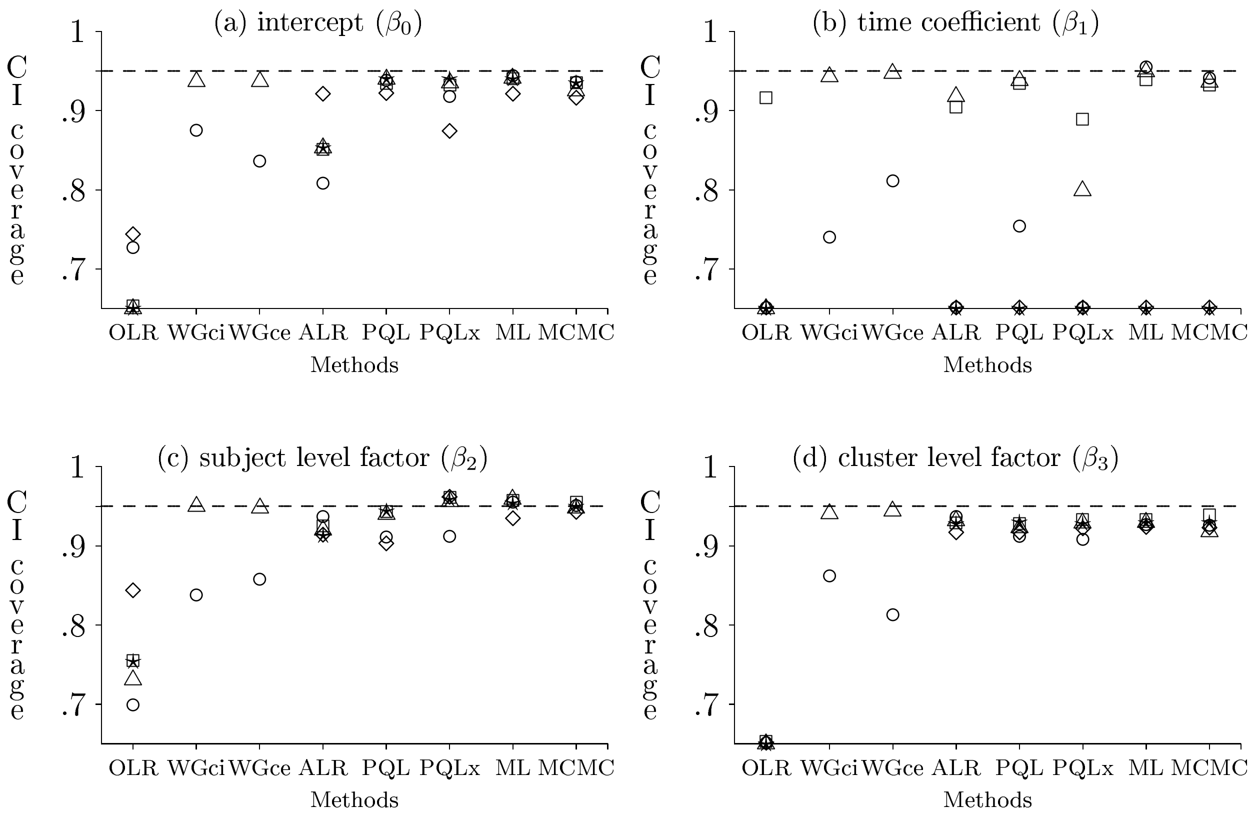

Figure 1: Confidence interval coverage for estimates of fixed effects parameters of eight estimation procedures, based on 1000 simulated datasets with missing values generated by random intercept model ($\rho = \mathbf{1}$) in five simulated scenarios of missing values: scc40 (missing values as in scc40 dataset), MARL, MARH (low (31%) and high (52%) proportion of missing values at random due to drop-outs), NMARL, NMARH (low (31%) and high (52%) proportion of missing values not at random) $\sim$ ($\square, \triangle, \circ, \star, \diamond$). Parameters: β_0 (intercept), β_1 (time coefficient), β_2 (subject level factor), β_3 (cluster level factor). Estimation procedures: OLR (ordinary logistic regression), WGci (weighted generalized estimating equations (WGEE) with independence correlation at cluster level), WGce (WGEE with exchangeable correlation at cluster level) ALR (alternating logistic regression), PQL (2nd order penalized quasi-likelihood), PQLx (2nd order penalized quasi-likelihood with extra-binomial dispersion), ML (maximum likelihood), MCMC (Bayesian Markov chain Monte Carlo).

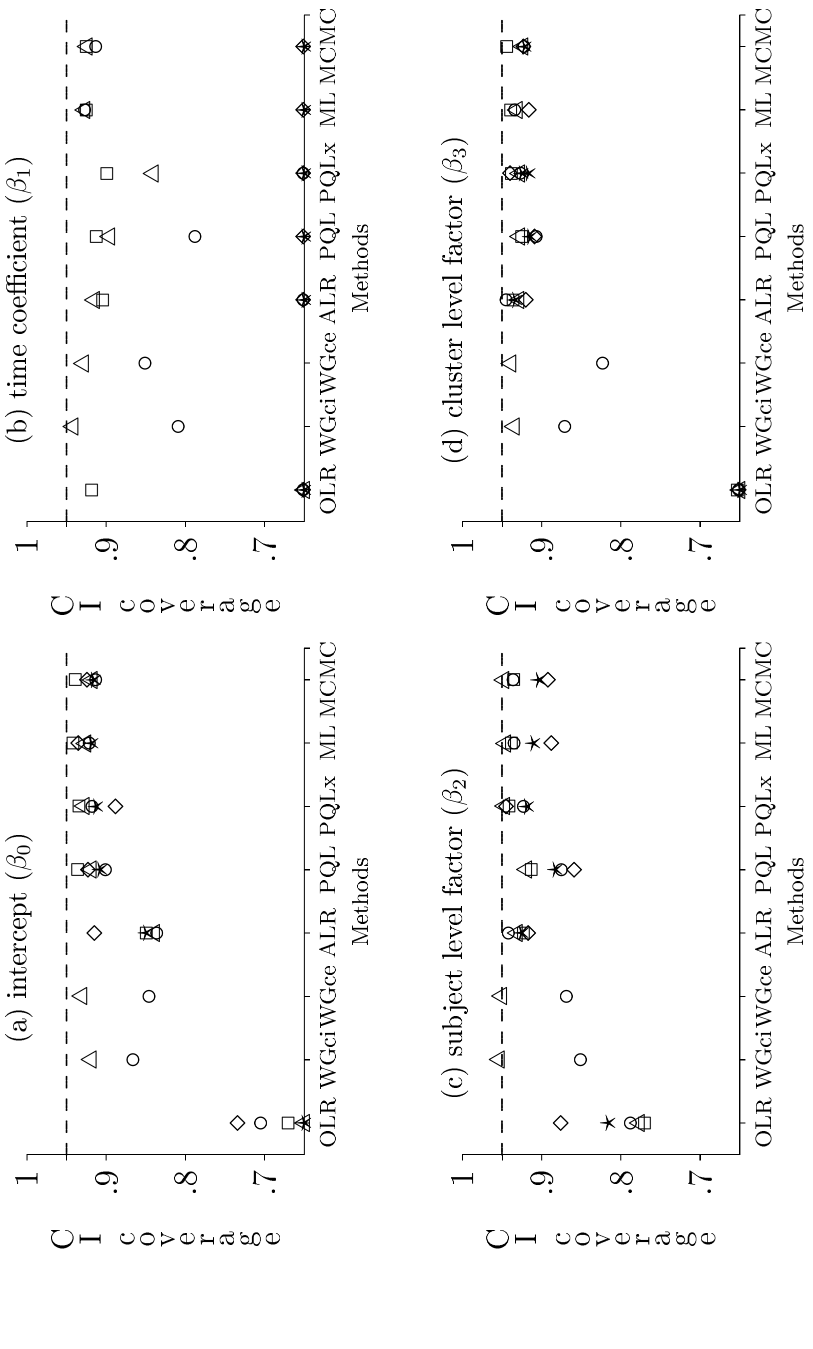

Figure 2: Confidence interval coverage for estimates of fixed effects parameters of eight estimation procedures, based on 1000 simulated datasets with missing values generated by random effects model ($\rho = \mathbf{0.9}$) in five simulated scenarios of missing values: scc40 (missing values as in scc40 dataset), MARL, MARH (low (31%) and high (52%) proportion of missing values at random due to drop-outs), NMARL, NMARH (low (31%) and high (52%) proportion of missing values not at random) $\sim$ ($\square, \triangle, \circ, \star, \diamond$). Parameters: β_0 (intercept), β_1 (time coefficient), β_2 (subject level factor), β_3 (cluster level factor). See Figure 1 for coding of estimation procedures.

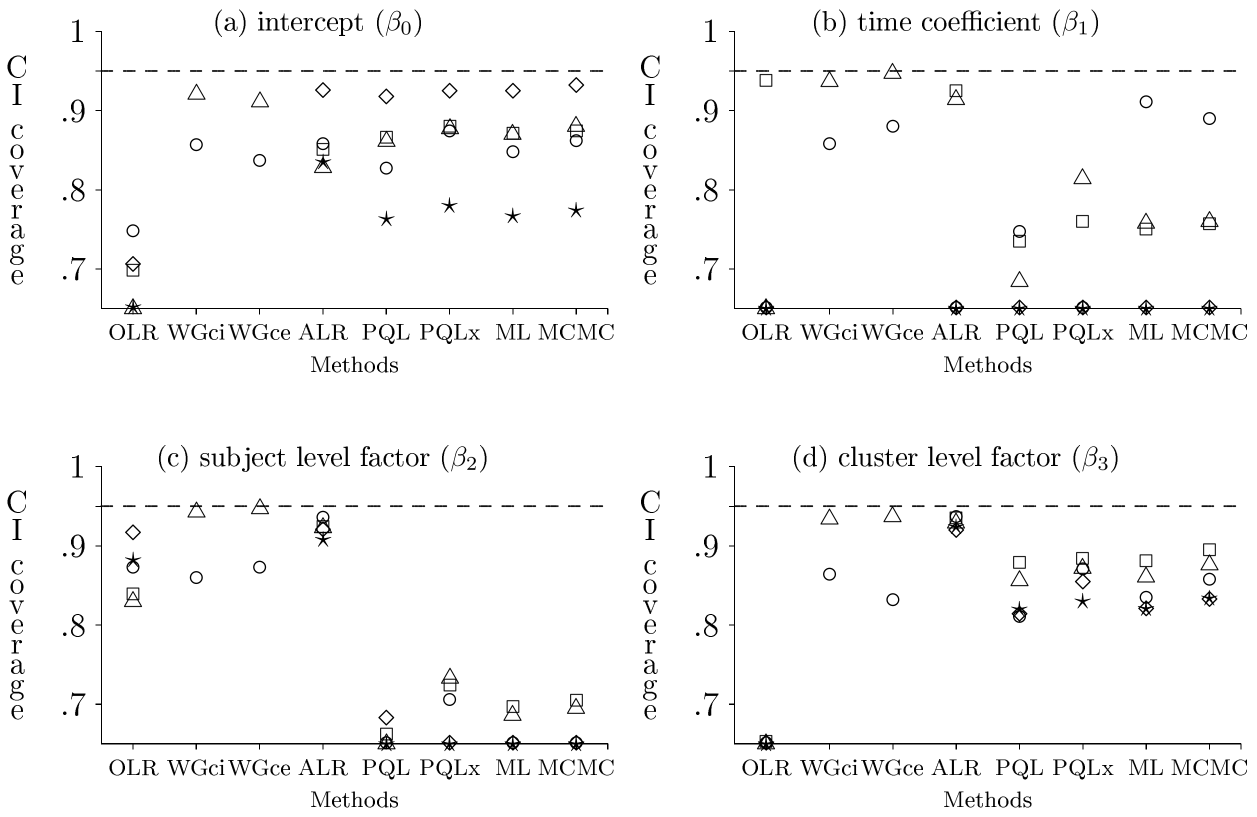

Figure 3: Confidence interval coverage for estimates of fixed effects parameters of eight estimation procedures, based on 1000 simulated datasets with missing values generated by random effects model ($\rho = \mathbf{0.5}$) in five simulated scenarios of missing values: scc40 (missing values as in scc40 dataset), MARL, MARH (low (31%) and high (52%) proportion of missing values at random due to drop-outs), NMARL, NMARH (low (31%) and high (52%) proportion of missing values not at random) $\sim$ ($\square, \triangle, \circ, \star, \diamond$). Parameters: β_0 (intercept), β_1 (time coefficient), β_2 (subject level factor), β_3 (cluster level factor). See Figure 1 for coding of estimation procedures.

5.2.2 Missing values: MAR scenarios

The positive dependence of the drop-out probability on a preceding event resulted in datasets with fewer events at the end of the time series than in the full dataset. For example, at $t = 8$ the full and MARH datasets had a proportion of events of 53% and 11%, respectively. Consequently, the strongest impact of the missing values for the simple OLR analysis was a negative bias for β_1, ranging down below -100%, and thus amounting to a sign switch in the coefficient (Table 2). The other coefficients showed a negative bias as well, and the confidence interval (CI) coverage was far below nominal (Figure 1).

The two likelihood-based procedures (ML, MCMC) were only slightly affected by the missing values, the only consistent significant changes being some increased estimates for σ_2^2 (Table 2). Overall, the proportion of missing values had no impact, except that the MARH scenario produced an additional small positive bias for β_1 for MCMC. CI coverages were close to but mostly below nominal (Figure 1).

The PQL procedure showed some negative biases, in particular for the time coefficient and variances parameters, and increasing with the severity of missing values. The bias of the time coefficient was substantial ($\approx$ 20%) and in the same direction as for OLR but less pronounced. Addition of an extra-binomial dispersion parameter (PQLx) altered the performance of the procedure dramatically. Biases for all parameters (except the dispersion parameter) were positive and of a larger magnitude (up to approx. 90% for β_1) than for PQL (Table 2). The extra-binomial parameter estimates of PQLx were centered at 0.72 for MARH setting only. However, except for β_1, the coverage of fixed effects CIs was fairly close to nominal for both PQL procedures (Figure 1).

The ALR procedure performed well in the MARL scenario, but produced a substantially inflated estimate of β_1 for MARH. The two weighted GEE (WGEE) procedures showed minor biases for MARL and substantial biases for MARH, in particular in the estimates of β_1 (Table 5). The direction of the biases varied across the two WGEE versions and the two data settings. The exchangeable correlation structure produced biases away from zero for MARL and towards zero for MARH. For MARH, all estimates from both versions of WGEE were associated with too small standard errors relative to the true values (Appendix A, Table A5), leading to substantial to strong undercoverage of CIs (Figure 1).

5.2.3 Missing values: NMAR scenarios

All estimation procedures included in the NMAR scenarios showed strong, negative relative biases (RBF range 53 - 320%) for the time coefficient (Table 2). Estimation of subject- and cluster-level fixed effects was relatively unaffected, with only minor biases (up to 6.4%) of which only few were significant for NMARL, but all except ALR were significant for NMARH. All significant biases were negative, except for PQLx. Subject- and cluster level variances showed similar patterns, with RBF values up to 14.4% (except for 50.4% for σ_2^2 and PQLx). Confidence intervals were strongly affected for β_1 and OLR but otherwise had coverages fairly close to nominal (Figure 1).

5.3 Autocorrelated data ($\rho < 1$)

Generally, the impact of the missing values was more affected by the amount of autocorrelation present in the data for random effects than marginal procedures. This finding is plausibly linked to the strong direct impact of the autocorrelation on the random effects estimates in the balanced full data (Masaoud and Stryhn, 2020). Specifically, when autocorrelation was present, estimates from random effects procedures tended to be less shrunk towards zero (i.e., inflated) in datasets with missing values than in the full data. Thus, the missing values to some extent counteracted the shrinkage caused by the autocorrelation.

5.3.1 Missing values: scc40 scenario

All random effects estimation procedures showed inflated estimates across almost all parameters relative to the estimates from the full data (Tables 3–4). The extra-binomial dispersion parameter for PQLx was downwards biased away from nominal dispersion ($\phi = 1$). The inflation was in most cases more pronounced at $\rho = 0.5$ than $\rho = 0.9$, except for the subject-level variance. Despite the inflation, the estimates were still clearly attenuated towards zero, although less so than in the full data (Appendix A, Tables A2–A3, and the CIs suffered from strong undercoverage for some parameters, in particular for $\rho = 0.5$ (Figures 2–3). For the marginal procedures (OLR and ALR), the impact of the missing values was still minor and almost unchanged from the random intercept model data.

5.3.2 Missing values: MAR and NMAR scenarios

For random effects procedures, the impacts of missing data were similar to those described above for the scc40 scenario. Some notable exceptions were that the extra binomial dispersion parameter for PQLx moved towards 1 in the MARH scenario, and some fixed effects estimates for ML and MCMC were similar at $\rho = 0.9$ and $\rho = 0.5$, or even closer to zero at the latter.

The marginal procedures showed different bias patterns with decreasing values of ρ (Table 5). For example, OLR biases generally decreased, whereas ALR biases were stable around zero for MARL, but for MARH the previously observed positive bias for β_1 increased in magnitude. In MARL data, the two weighted GEE procedures performed roughly on par with the random intercept data. Some bias reduction could be seen for MARH with decreasing ρ, but the bias in standard errors and resulting poor coverage of CIs remained (Table 4 and Figures 2–3).

In NMAR scenarios, the introduction of autocorrelation had similar impacts on the biases of the different estimation procedures as in the MAR scenarios. However, from a practical point of view it did not alter the magnitude and severity of the biases described for the random intercept model data substantially (Tables 3–5). The CI coverages for random effects procedures dropped substantially below nominal with decreasing ρ (Figures 2–3), but this was attributable to the autocorrelation itself and not a result of the missing values.

6 Discussion

6.1 Modelling of missing values in a dataset

When an (applied) researcher is confronted with a dataset containing missing values, they face a crucial decision (among many others) regarding the analysis: whether to ignore or model the missing values. A quick glance through scientific journals publishing studies involving statistical analyses will show that in most cases the missingness is ignored, despite the good statistical understanding of procedures to model missing data (e.g., Little and Rubin, 2002; McKnight *et al.*, 2007, Chapter 5). Among the reasons for this apparent negligence in the statistical analysis would be beliefs that (i) the statistical methods actually used were robust to missing values, and (ii) statistical methods to deal with missing values would be difficult to employ and assess. While focusing on the quantification of assumption (i), the present paper also puts forward the idea of modelling the *occurrence* of missing values by simple models, in order to gain insight into the types of missing values in a dataset before deciding whether the missing values should be modelled or not.

Our example dataset (scc40) contained a total of 31% missing values relative to a dataset with complete series on all subjects, intuitively a relatively large proportion. However, more than half of the missing values were due to a type of missing values (delayed entry) that could reasonably be assumed to have arisen by the least serious missing value process (MCAR). Delayed entry can be thought of as a left truncation of the time series on a subject, whereas a drop-out can be thought of as a right truncation of the series. Little attention seems to have been paid in the literature to delayed entry as a source of missing values, but in our view it may occur commonly for data collected retrospectively from databases.

It is critically important to model missing values in a single dataset appropriately. We modelled the different types of missing values by variants of the logistic regression model proposed by Diggle and Kenward (1994). Possible extensions of the approach can easily be suggested. For data including treatment factors of key interest, it would be natural to include these as fixed effects in the models. Also, if NMAR processes are suspected for some of the missingness types, one could consider specific NMAR models, such as pattern-mixture models (Molenberghs *et al.*, 2001), even though they may be more difficult to fit to the missingness values. We considered intermittent missing values as the type most likely to involve NMAR missingness, and by the very low proportion of such missing values in the data, NMAR modelling was considered unnecessary in our example.

The simulation results for the scc40 scenario showed almost no impact of the combination of missing patterns on the estimation procedures. Obvious reasons for this perhaps somewhat surprising finding, given the relatively large proportion of missing values, are that delayed entry accounted for a substantial part of the missing values, and that the missing value mechanisms studied did not include NMAR.

6.2 Impact of missing values

Evidently, the impact of missing values in a dataset depends on the types and probabilistic mechanisms of the missing values as well as their proportions. Our simulation studies gave a sense of the required level of missingness needed to substantially affect results (of different procedures), and the extent to which individual parameters were affected. As discussed above, estimation in the scc40 data seemed hardly affected at all despite a sizeable proportion of missing values. With the most severe missingness mechanism (NMAR) at the same level of missing values, the picture changed completely. The strong biases for the time coefficient across all procedures agrees with findings reported by Little and Rubin (2002); Laird (1988) that ignoring the NMAR missing process leads to biased estimates, even when only a small proportion of the sample is missing (Choi and Lu, 1995). It is notable that subject- and cluster-level parameters could be relatively little affected even in the most extreme scenarios, indicating that without a direct link to the missingness mechanism results could be relatively robust. Specific comments for some of the procedures follow.

6.2.1 Weighted generalized estimating equations (WGEE)

The GEE procedures of interest for the present 3-level structure involved either independent or exchangeable correlations at the cluster level. As these structures ignore the within-subject correlations, they seem unlikely to capture the true correlation structure. The strong biases for OLR in MAR scenarios, whose estimates may be interpreted as of an unweighted GEE with independent correlation structure, confirmed our suspicion.

The WGEE procedures performed fairly well relative to the full data for MARL regardless of the correlation structure in the data, in agreement with findings reported by Jansen *et al*. (2006) and Molenberghs and Verbeke (2005, Chapter 27). Small biases have also been reported (Preisser *et al*., 2002), which could substantiate the small bias we found for the time coefficient. For MARH, the same parameter exhibited substantial biases which seem to contradict its theoretical (asymptotic) properties (Robins *et al*., 1995), but has also been reported previously for two-level data (Preisser *et al*., 2002). One possible source of the bias is fluctuations in estimating the weights as the number of measurements per subject becomes small, if not very small.

6.2.2 Alternating logistic regression (ALR)

Overall, we found ALR estimates to be in close agreement with those of the full data (except for the time coefficient in MARH and NMAR scenarios), regardless of the correlation structure in the data. The bias in the MARH data was somewhat surprisingly in the opposite direction of biases for OLR and WGEE. As ALR is based on similar estimating equations as GEE, one may speculate that a weighting scheme akin to WGEE could be developed for ALR processes; in any case, the properties of ALR under MAR processes warrant further study.

6.2.3 Penalized quasi-likelihood procedures (PQL, PQLx)

Drawbacks and caveats of iterative reweighting algorithms such as PQL for estimation in random effects models have been discussed extensively in the literature (Breslow, 2003). However, we are not aware of published work discussing any inferior performance of quasi-likelihood procedures under MAR processes. Our results for PQL demonstrated a bias in the time coefficient that we think is not attributable to the well-known attenuation of variance parameters in certain settings, because it does not affect all fixed effects parameters equally nor has the same direction as for OLR. As for ALR, a suitable weighting scheme for PQL under MAR processes could be hypothesized. Allowing for an extra-binomial dispersion (PQLx) produced stronger biases and in the opposite direction, adding to the evidence from previous work (Masaoud and Stryhn, 2020) that the inclusion of the extra-binomial parameter has more profound impacts on the performance of the procedure than one might intuitively expect. Based on our findings, the inclusion of the extra-binomial parameter in the presence of substantial missing data is not to be recommended.

6.2.4 Likelihood-based procedures (ML, MCMC)

Strictly speaking, both ML and MCMC are based on likelihood approximations, either by quadrature or MCMC sampling. From this perspective, our results for these procedures demonstrated that the accuracy of the approximations were sufficient to, by and large, ensure the ignorability of MCAR and MAR processes predicted from theory (Little, 1988). However, slight increases in MCMC estimates for the time coefficient and cluster level variance remained unexplained. On the other hand, NMAR processes affected the likelihood-based procedures to roughly the same extent as the other procedures, so their advantage in this context is essentially linked to the MAR assumption.

References

Ali, M. W., Talukder, E. (2005), “Analysis of longitudinal binary data with missing data due to dropouts,” *Journal of Biopharmaceutical Statistics*, 15, 993–1007.

Breslow, N.E. (2003), “Whither PQL?,” *University of Washington Biostatistics Working Paper Series*, 192.

Breslow, N. E. and Clayton, D.G. (1993), “Approximate inference in generalized linear mixed models,” *Journal of the American Statistical Association*, 88, 9–25.

Browne, W. J. and Draper, D. (2006), “A comparison of Bayesian and likelihood-based methods for fitting multilevel models,” *Bayesian Analysis*, 3, 473–514.

Carey, V., Zeger, S.L. and Diggle, P. (1993), “Modeling multivariate binary data with alternating logistic regressions,” *Biometrika*, 80, 517–526.

Choi, S., Lu, I. L. (1995), “Effect of non-random missing data mechanisms in clinical trials,” *Statistics in Medicine*, 14, 2675–2684.

Cook, D., Swayne, D. F. (2007), *Interactive and Dynamic Graphics for Data Analysis With R and GGobi*. Springer, New York.

Congdon, P. (2003) *Applied Bayesian Modelling*. Wiley, New York.

Diggle P. J., Kenward M. G. (1994), “Informative dropout in longitudinal data analysis,” *Applied Statistics*, 43, 49–93.

Diggle P. J., Heagerty, P., Liang, K. Y., and Zeger, S. L. (2002), *Analysis of Longitudinal Data (2nd ed.)*, Oxford University Press, Oxford.

Dohoo, I .R., Martin, S. W. and Stryhn, H. (2009), *Veterinary Epidemiologic Research (2nd ed.)*. AVC Inc., Charlottetown, Canada.

Fitzmaurice G. (2004), *Applied Logitudinal analysis*. Wiley-Interscience, Hoboken, NJ.

Fitzmaurice, G. M. (2003), "Methods for handling dropouts in longitudinal clinical trials," *Statistica Neerlandica*, 57, 75–99.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (Comments on article by Browne and Draper). *Bayesian Analysis* **3**, 515–534.

Goldstein, H. (1991), "Nonlinear multilevel models with an application to discrete response data," *Biometrika*, 78, 45–51.

Goldstein, H. and Rasbash, J. (1996), "Improved approximations for multilevel models with binary responses," *Journal of the Royal Statistical Society, Series A*, 159, 505–513.

Gibson, N. M., Olejnik, S. (2003), "Treatment of missing data at the second level of hierarchical linear models," *Educational and Psychological Measurement*, 63, 204–238.

Goldstein, H., Browne, W. J. and Rasbash, J. (2002), "Partitioning variation in multilevel models," *Understanding Statistics*, 1, 223–231.

Heyting, A., Tolboom, J. T., Essers, J. G. (1992), "Statistical handling of drop-outs in longitudinal clinical trials," *Statistics in Medicine*, 11, 2043–2061.

Hogan, J. W., Roy, J., Korkontzelou, C. (2004), "Biostatistics tutorial: Handling dropout in longitudinal data," *Statistics in Medicine*, 23, 1455–1497.

Jansen, I., Beunckens, C., Molenberghs, G., Verbeke, G., Mallinckrodt, C. (2006), "Analyzing incomplete discrete longitudinal clinical trial data," *Statistical Science*, 21, 52–69.

Kenward, M. G., Goetghebeur, J. T., Molenberghs, G. (2001), "Sensitivity analysis for incomplete categorical data," *Statistical Modelling*, 1, 31–48.

Kim, J. O., Curry, J. (1977), "The treatment of missing data in multivariate analysis," *Sociological Methods and Analysis*, 6, 215–240.

Kunthel Bya, Bahjat F. Qaqisha, John S. Preisser, Jamie Perinb, Richard C. Zinkc. (2014), "ORTH: R and SAS software for regression models of correlated binary data based on orthogonalized residuals and alternating logistic regressions," *Computer Methods and Programs in Biomedicine*, 113, 557–568

Laird, N. M. (1988), "Missing data in longitudinal studies," *Statistics in Medicine*, 7, 305–315.

Liang, K. Y. and Zeger, S. L. (1986), "Longitudinal Data-Analysis Using Generalized Linear-Models," *Biometrika*, 73, 13–22.

Little, R. J. A. (1995), "Modeling the drop-out mechanism in repeated-measures studies," *Journal of the American Statistical Association*, 90, 1112–1121.

Little, R. J. A. (1988), "Robust estimation of the mean and covariance matrix from data with missing values," *Applied Statistics*, 37, 23–38.

Little, R. J. A., Rubin D. B. (2002), *Statistical Analysis With Missing Data (2nd ed.)*, 2nd ed., Wiley-Interscience, Hoboken, New Jersey.

Masaoud, E. A. M. (2009), “Statistical models for binary repeated measures and hierarchical data in veterinary science. PhD Thesis, Department of Health Management, Atlantic Veterinary College, Charlottetown, Canada.

Masaoud, E. and Stryhn, H. (2010), “A simulation study to assess statistical methods for binary repeated measures data,” *Preventive Veterinary Medicine*, 93, 81–97.

Masaoud, E. and Stryhn, H. (2020), “A comparison of statistical methods for the analysis of binary repeated measures data with additional hierarchical structure,” *Journal of Statistical Research*, 54, 1–25..

McKnight, P. E., McKnight, K. M., Sidani, S., Figueredo, A. (2007), *Missing Data: A Gentle Introduction*. Guilford Press, New York.

Molenberghs, G., Kenward, M. G., Goetghebeur, E. (2001), “Sensitivity analysis for incomplete contingency tables: the Slovenian plebiscite case,” *Applied Statistics*, 50, 15–29.

Molenberghs, G., Verbeke, G. (2005),” *Models for Discrete Longitudinal Data*. Springer, New York.

Neuhaus, J. M. (1992), “Statistical methods for longitudinal and clustered design with binary responses,” *Statistical Methods in Medical Research*, 1, 249–273.

Preisser, J. S., Lohman, K. K., Rathouz, P. J. (2002), “Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random,” *Statistics in Medicine*, 21, 3035–3054.

Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R. and Jones, D. R. (2005). How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine* **24**, 2401–2428.

Rabe-Hasketh, S., Skrondal, A. and Pickles, A. (2002), “Reliable estimation of generalised linear mixed models using adaptive quadrature,” *The Stata Journal* , 2, 1–21.

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

Roy, J. (2003), “Modeling longitudinal data with non-ignorable dropouts using a latent dropout class model,” *Biometrics*, 59, 829–836.

Rodriguez G, Goldman N. (1995), “An assessment of estimation procedures for multilevel models with binary responses,” *Journal of the Royal Statistical Society, Series A*, 158, 73–89.

Rodriguez G, Goldman N. (2001), “Improved estimation procedures for multilevel models with binary response: a case-study,” *Journal of the Royal Statistical Society, Series A*, 164, 339–355.

Skrondal, A. and Rabe-Hesketh, S. (2007), “Redundant overdispersion parameters in multilevel models for categorical responses,” *Journal of educational and behavioral statistics*, 32, 419–430.

Singer, J., Willett, J. (1993), “It’s about time: Using discrete-time survival analysis to study duration and the timing of events,” *Journal of Educational Statistics*, 18, 155–195.

Snijders, T. A. B., Bosker, R. J. (1993), “Standard errors and sample sizes for two-level research,” *Journal of Educational Statistics*, 18, 237–259.

Snijders, T. A. B. and Bosker, R. J. (2012)," *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modelling (2nd edition.)*. Sage Publishers, London.

Stryhn, H., Dohoo, I. R., Tillard, E. and Hagedorn-Olsen, T. (2000), "Simulation as a tool of validation in hierarchical generalised linear models. IX*th* International Conference of Veterinary Epidemiology and Economics, Breckenridge, Colorado.

Sturtz, S., Ligges, U. and Gelman, A. (2005), "R2WinBUGS: A package for running WinBUGS from R," *Journal of Statistical Software*, 12, 1–16.

Touloumi, G., Babiker, A. G., Pocock, S. J., Darbyshire, J. H. (2001), "Impact of missing data due to drop-outs on estimators for rates of change in longitudinal studies: a simulation study," *Statistics in Medicine* , 20, 3715–28.

Ziegler, A., Kastner, C. and Blettner, M. (1998), "The generalized estimating equations: an annotated bibliography," *Biometrical Journal*, 40, 115–139.

Received: June 15, 2023

Accepted: January 26, 2024

A Appendix

Table A1: Relative bias of estimates and standard errors to the true values with a significance indication, based on analyses of 1000 simulated datasets generated by random intercept model ($\rho = \mathbf{1}$) in five simulated scenarios of missing values: scc40 (missing values as in scc40 dataset), MARL, MARH (low (31%) and high (52%) proportion of missing values at random due to drop-outs), MNARL, MNARH (low (31%) and high (52%) proportion of missing values not at random). Parameters: β_0 (intercept), β_1 (time coefficient), β_2 (subject level factor), β_3 (cluster level factor), σ_2^2 (variance at subject level), σ_3^2 (variance at cluster level), ϕ (extra-binomial dispersion). Estimation procedures: PQL (2nd order penalized quasi-likelihood), PQLx (2nd order penalized quasi-likelihood with extra-binomial dispersion), ML (maximum likelihood), MCMC (Bayesian Markov chain Monte Carlo).

		PQL		PQLx		ML		MCMC	
Scenario	Parameter	Est.	SE	Est.	SE	Est.	SE	Est.	SE
scc40	β_0	-1.3	-4.9^{*}	$2.7^{\ddagger}$	-5.9^{*}	0.1	-4.3^{*}	0.0	-3.1
	β_1	$-0.7^{\ddagger}$	-8.7^{*}	$4.0^{\ddagger}$	-16.7^{*}	0.0	-4.3	0.0	-4.3
	β_2	$-4.2^{\dagger}$	-0.6	-4.2	-0.6	-0.5	0.0	-0.1	0.0
	β_3	-2.0	-4.3^{*}	1.5	-4.1	1.5	-4.1	1.7	-1.3
	σ_2^2	$-10.8^{\ddagger}$	-15.1^{*}	$7.5^{\ddagger}$	-23.2^{*}	$0.8^{\dagger}$	-2.8	$2.5^{\ddagger}$	-2.0
	σ_3^2	$-15.4^{\ddagger}$	-5.1^{*}	$-9.2^{\ddagger}$	-5.2^{*}	$-9.2^{\ddagger}$	-3.1	0.1	8.8^{*}
	ϕ			$-17.8^{\ddagger}$	37.5^{*}				
MARL	β_0	-1.0	-2.4	$4.6^{\ddagger}$	-3.4	1.1	-1.1	$3.2^{\dagger}$	-5.2^{*}
	β_1	$-2.1^{\ddagger}$	-5.0^{*}	$13.8^{\ddagger}$	-17.0^{*}	0.8	0.1	$2.1^{\ddagger}$	-1.7
	β_2	$-4.5^{\ddagger}$	-2.0	0.0	-1.2	-1.0	0.1	-0.1	-1.0
	β_3	$-3.2^{\ddagger}$	-5.2^{*}	1.4	-5.3^{*}	0.2	-4.4^{*}	2.2	-3.4
	σ_2^2	$-12.1^{\ddagger}$	-12.4^{*}	$10.7^{\ddagger}$	-23.2^{*}	0.7	2.4	$3.5^{\ddagger}$	-0.7
	σ_3^2	$-14.0^{\ddagger}$	-6.4^{*}	$-5.5^{\ddagger}$	-6.2^{*}	$-7.6^{\ddagger}$	-3.8^{*}	2.0	11.0^{*}
	ϕ			$-19.0^{\ddagger}$	25.6^{*}				
MARH	β_0	$-3.2^{\ddagger}$	-5.3^{*}	$13.6^{\ddagger}$	-6.3^{*}	1.3	-2.1	1.6	-0.4
	β_1	$-21.7^{\ddagger}$	-15.7^{*}	$92.9^{\ddagger}$	-45.6^{*}	1.2	2.0	$4.4^{\ddagger}$	-0.3
	β_2	$-7.1^{\ddagger}$	-7.7^{*}	$11.0^{\ddagger}$	-3.1	-1.0	-0.6	-0.2	-0.7
	β_3	$-5.5^{\ddagger}$	-8.0^{*}	$12.9^{\ddagger}$	-7.3^{*}	0.5	-5.4^{*}	1.3	-3.5
	σ_2^2	$-26.0^{\ddagger}$	-26.7^{*}	$66.7^{\ddagger}$	-49.3^{*}	$1.4^{\dagger}$	3.6	$5.0^{\ddagger}$	1.3
	σ_3^2	$-19.0^{\ddagger}$	-9.5^{*}	$15.8^{\ddagger}$	-6.0^{*}	$-8.1^{\ddagger}$	-3.4	1.7	8.4^{*}
	ϕ			$-27.6^{\ddagger}$	-26.7^{*}				
NMARL	β_0	$-2.6^{\ddagger}$	-1.6	$2.7^{\ddagger}$	-2.5	0.2	-0.7	0.6	-1.8
	β_1	$-78.8^{\ddagger}$	-3.8	$-71.3^{\ddagger}$	-14.7^{*}	$-77.7^{\ddagger}$	-1.8	$-77.4^{\ddagger}$	-2.4
	β_2	$-4.8^{\ddagger}$	-0.4	-0.6	0.4	$-1.5^{\ddagger}$	0.7	$-1.1^{\dagger}$	-0.4
	β_3	$-3.3^{\ddagger}$	-4.6^{*}	1.1	-4.6^{*}	-0.2	-4.1	0.4	-3.2
	σ_2^2	$-11.2^{\ddagger}$	-11.9^{*}	$11.2^{\ddagger}$	-22.0^{*}	0.2	1.3	$1.9^{\ddagger}$	1.1
	σ_3^2	$-13.9^{\ddagger}$	-6.5^{*}	$-6.0^{\ddagger}$	-6.5^{*}	$-8.4^{\ddagger}$	-4.8^{*}	0.8	7.8^{*}
	ϕ			$-19.8^{\ddagger}$	16.4^{*}				
NMARH	β_0	$11.5^{\ddagger}$	-6.3^{*}	$26.2^{\ddagger}$	-5.7^{*}	$14.7^{\ddagger}$	-3.1	$15.1^{\ddagger}$	-1.0
	β_1	$-317.4^{\ddagger}$	-5.7^{*}	$-296.5^{\ddagger}$	-22.8^{*}	$-318.0^{\ddagger}$	-10.3^{*}	$-318.1^{\ddagger}$	-10.9^{*}
	β_2	$-8.7^{\ddagger}$	-4.0	$4.4^{\ddagger}$	1.4	$-6.2^{\ddagger}$	-2.2	$-5.5^{\ddagger}$	-2.2
	β_3	$-7.3^{\ddagger}$	-7.7^{*}	$6.3^{\ddagger}$	-6.2^{*}	$-5.3^{\ddagger}$	-6.5^{*}	$-4.7^{\ddagger}$	-4.3^{*}
	σ_2^2	$-23.0^{\ddagger}$	-24.8^{*}	$53.1^{\ddagger}$	-40.9^{*}	$-11.3^{\ddagger}$	0.8	$-8.7^{\ddagger}$	0.3
	σ_3^2	$-19.8^{\ddagger}$	-7.9^{*}	$5.5^{\ddagger}$	-3.6	$-16.3^{\ddagger}$	-4.3^{*}	$-7.8^{\ddagger}$	7.0^{*}
	ϕ			$-28.3^{\ddagger}$	-47.3^{*}				

† significant bias in estimate at $P < 0.05$; ‡ significant bias in estimate at $P < 0.01$; * significant bias in standard error at $P < 0.05$

Table A2: Relative bias of estimates and standard errors to the true values with a significance indication, based on analyses of 1000 simulated datasets generated by autoregressive random effects model with ($\rho = \mathbf{0.9}$) in five simulated scenarios of missing values: scc40 (missing values as in scc40 dataset), MARL, MARH (low (31%) and high (52%) proportion of missing values at random due to drop-outs), MNARL, MNARH (low (31%) and high (52%) proportion of missing values not at random). Parameters: β_0 (intercept), β_1 (time coefficient), β_2 (subject level factor), β_3 (cluster level factor), σ_2^2 (variance at subject level), σ_3^2 (variance at cluster level), ϕ (extra-binomial dispersion). see Table A1 for coding of estimation procedures.

		Statistical Methods							
Scen-	Parm-	PQL		PQLx		ML		MCMC	
ario	eter	Est.	SE	Est.	SE	Est.	SE	Est.	SE
scc40	β_0	$-6.1^{\ddagger}$	-1.3	$-2.7^{\ddagger}$	-2.1	$-5.3^{\ddagger}$	-0.9	$-5.1^{\ddagger}$	-1.3
	β_1	$-6.5^{\ddagger}$	-9.1^{*}	$-1.3^{\ddagger}$	-13.6^{*}	$-4.7^{\ddagger}$	-4.5^{*}	$-4.7^{\ddagger}$	-4.5^{*}
	β_2	$-6.4^{\ddagger}$	-1.4	$-3.4^{\ddagger}$	-0.7	$-3.9^{\ddagger}$	0.7	$-3.7^{\ddagger}$	0.0
	β_3	$-6.5^{\ddagger}$	-3.8	$-3.5^{\dagger}$	-3.7	$-4.2^{\dagger}$	-3.4	$-3.9^{\ddagger}$	-2.0
	σ_2^2	$-25.2^{\ddagger}$	-16.6^{*}	$-10.3^{\ddagger}$	-24.7^{*}	$-16.8^{\ddagger}$	-2.9	$-15.6^{\ddagger}$	-2.8
	σ_3^2	$-20.7^{\ddagger}$	-12.5^{*}	$-15.4^{\ddagger}$	-12.2^{*}	$-16.6^{\ddagger}$	-9.7^{*}	$-8.3^{\ddagger}$	2.4
	ϕ			$-16.0^{\ddagger}$	46.7^{*}				
MARL	β_0	$-6.5^{\ddagger}$	-5.7^{*}	-1.7	-6.5^{*}	$-5.2^{\ddagger}$	-4.1	$-5.2^{\ddagger}$	-3.7
	β_1	$-6.1^{\ddagger}$	-11.1^{*}	$7.7^{\ddagger}$	-21.6^{*}	$-3.8^{\ddagger}$	-5.9^{*}	$-3.2^{\ddagger}$	-6.2^{*}
	β_2	$-6.5^{\ddagger}$	-0.4	$-2.5^{\ddagger}$	-0.5	$-4.1^{\ddagger}$	2.3	$-3.8^{\ddagger}$	2.2
	β_3	$-7.4^{\ddagger}$	-5.4^{*}	$-3.4^{\ddagger}$	-5.4^{*}	$-5.1^{\ddagger}$	-4.1	$-5.0^{\ddagger}$	-2.6
	σ_2^2	$-26.7^{\ddagger}$	-20.2^{*}	$-8.0^{\ddagger}$	-29.8^{*}	$-17.2^{\ddagger}$	-4.7^{*}	$-15.9^{\ddagger}$	-4.6^{*}
	σ_3^2	$-20.0^{\ddagger}$	-5.7^{*}	$-12.8^{\ddagger}$	-6.0^{*}	$-15.9^{\ddagger}$	-2.7	$-7.4^{\ddagger}$	10.8^{*}
	ϕ			$-17.1^{\ddagger}$	40.7^{*}				
MARH	β_0	$-8.4^{\ddagger}$	-8.4^{*}	$4.0^{\ddagger}$	-9.7^{*}	$-5.7^{\ddagger}$	-4.9^{*}	$-5.4^{\ddagger}$	-3.5
	β_1	$-15.0^{\ddagger}$	-25.8^{*}	$78.1^{\ddagger}$	-50.7^{*}	$4.9^{\ddagger}$	-9.0^{*}	$7.7^{\ddagger}$	-10.6^{*}
	β_2	$-10.4^{\ddagger}$	-4.1	$5.6^{\ddagger}$	-3.2	$-6.1^{\ddagger}$	2.8	$-5.5^{\ddagger}$	2.8
	β_3	$-11.2^{\ddagger}$	-6.0^{*}	$4.7^{\ddagger}$	-6.2^{*}	$-7.1^{\ddagger}$	-3.3	$-6.6^{\ddagger}$	-1.6
	σ_2^2	$-46.5^{\ddagger}$	-35.0^{*}	$17.8^{\ddagger}$	-57.7^{*}	$-27.9^{\ddagger}$	-8.7^{*}	$-25.6^{\ddagger}$	-10.0^{*}
	σ_3^2	$-27.6^{\ddagger}$	-9.3^{*}	1.0	-9.2^{*}	$-20.4^{\ddagger}$	-2.5	$-12.0^{\ddagger}$	10.1^{*}
	ϕ			$-22.3^{\ddagger}$	-8.6^{*}				
NMARL	β_0	$-11.5^{\ddagger}$	-4.9^{*}	$-7.8^{\ddagger}$	-5.8^{*}	$-10.7^{\ddagger}$	-3.5	$-10.5^{\ddagger}$	-2.4
	β_1	$-83.8^{\ddagger}$	-12.2^{*}	$-79.1^{\ddagger}$	-20.3^{*}	$-83.2^{\ddagger}$	-9.8^{*}	$-82.9^{\ddagger}$	-10.3^{*}
	β_2	$-10.0^{\ddagger}$	0.0	$-6.7^{\ddagger}$	0.5	$-8.7^{\ddagger}$	3.0	$-8.4^{\ddagger}$	2.0
	β_3	$-10.8^{\ddagger}$	-3.5	$-7.5^{\ddagger}$	-3.6	$-9.7^{\ddagger}$	-2.9	$-9.5^{\ddagger}$	-1.0
	σ_2^2	$-40.9^{\ddagger}$	-20.4^{*}	$-26.3^{\ddagger}$	-29.3^{*}	$-35.0^{\ddagger}$	-5.0	$-34.1^{\ddagger}$	-5.0^{*}
	σ_3^2	$-25.7^{\ddagger}$	-6.6^{*}	$-20.0^{\ddagger}$	-7.0^{*}	$-23.8^{\ddagger}$	-3.3	$-15.9^{\ddagger}$	10.0^{*}
	ϕ			$-15.6^{\ddagger}$	49.5^{*}				
NMARH	β_0	$5.5^{\ddagger}$	-6.8^{*}	$18.1^{\ddagger}$	-7.9^{*}	$6.8^{\ddagger}$	-3.5	$7.2^{\ddagger}$	-0.8
	β_1	$-309.6^{\ddagger}$	-6.9^{*}	$-297.9^{\ddagger}$	-20.9^{*}	$-309.8^{\ddagger}$	-10.6^{*}	$-309.9^{\ddagger}$	-10.9
	β_2	$-12.9^{\ddagger}$	-2.4	$-1.6^{\ddagger}$	-1.0	$-11.5^{\ddagger}$	0.4	$-11.0^{\ddagger}$	0.2
	β_3	$-12.9^{\ddagger}$	-3.5	-1.4	-3.1^{*}	$-12.0^{\ddagger}$	-2.3	$-11.6^{\ddagger}$	0.8
	σ_2^2	$-44.5^{\ddagger}$	-26.8^{*}	$10.0^{\ddagger}$	-48.0^{*}	$-36.5^{\ddagger}$	3.7	$-35.1^{\ddagger}$	-5.4^{*}
	σ_3^2	$-28.0^{\ddagger}$	-6.4^{*}	$-7.5^{\ddagger}$	-5.4^{*}	$-26.3^{\ddagger}$	-1.8	$-18.8^{\ddagger}$	10.4^{*}
	ϕ			$-23.1^{\ddagger}$	-39.6^{*}				

† significant bias in estimate at $P < 0.05$; ‡ significant bias in estimate at $P < 0.01$; * significant bias in standard error at $P < 0.05$

Table A3: Relative bias of estimates and standard errors to the true values with a significance indication, based on analyses of 1000 simulated datasets generated by autoregressive random effects model with ($\rho = \mathbf{0.5}$) in five simulated scenarios of missing values: scc40 (missing values as in scc40 dataset), MARL, MARH (low (31%) and high (52%) proportion of missing values at random due to drop-outs), MNARL, MNARH (low (31%) and high (52%) proportion of missing values not at random). Parameters: β_0 (intercept), β_1 (time coefficient), β_2 (subject level factor), β_3 (cluster level factor), σ_2^2 (variance at subject level), σ_3^2 (variance at cluster level), ϕ (extra-binomial dispersion). see Table A1 for coding of estimation procedures.

		Statistical Methods							
Scen-	Parm-	PQL		PQLx		ML		MCMC	
ario	eter	Est.	SE	Est.	SE	Est.	SE	Est.	SE
scc40	β_0	$-16.7^{\ddagger}$	-2.5	$-14.7^{\ddagger}$	-3.5	$-16.8^{\ddagger}$	-1.0	$-17.1^{\ddagger}$	0.0
	β_1	$-16.7^{\ddagger}$	0.0	$-14.7^{\ddagger}$	-5.3^{*}	$-16.7^{\ddagger}$	0.0	$-16.7^{\ddagger}$	0.0
	β_2	$-16.7^{\ddagger}$	-2.7	$-14.9^{\ddagger}$	-3.5	$-16.5^{\ddagger}$	0.9	$-16.3^{\ddagger}$	0.0
	β_3	$-16.8^{\ddagger}$	-2.4	$-15.0^{\ddagger}$	-2.4	$-16.7^{\ddagger}$	-1.6	$-16.7^{\ddagger}$	1.6
	σ_2^2	$-68.4^{\ddagger}$	-14.2^{*}	$-61.9^{\ddagger}$	-21.6^{*}	$-65.5^{\ddagger}$	-1.6	$-65.2^{\ddagger}$	0.0
	σ_3^2	$-36.4^{\ddagger}$	-10.2^{*}	$-33.5^{\ddagger}$	-11.1^{*}	$-36.2^{\ddagger}$	-7.0^{*}	$-29.6^{\ddagger}$	6.3^{*}
	ϕ			$-10.2^{\ddagger}$	140.0^{*}				
MARL	β_0	$-16.9^{\ddagger}$	-5.5^{*}	$-14.2^{\ddagger}$	-6.4^{*}	$-16.8^{\ddagger}$	-3.9	$-16.9^{\ddagger}$	-0.9
	β_1	$-18.5^{\ddagger}$	-8.9^{*}	$-10.8^{\ddagger}$	-15.8^{*}	$-17.3^{\ddagger}$	-2.7	$-17.0^{\ddagger}$	-3.6
	β_2	$-16.9^{\ddagger}$	-4.9^{*}	$-14.5^{\ddagger}$	-4.8^{*}	$-16.6^{\ddagger}$	0.9	$-16.4^{\ddagger}$	-1.0
	β_3	$-16.0^{\ddagger}$	-7.2^{*}	$-13.6^{\ddagger}$	-7.5^{*}	$-15.7^{\ddagger}$	-6.3^{*}	$-15.7^{\ddagger}$	-2.7
	σ_2^2	$-68.7^{\ddagger}$	-16.1^{*}	$-60.8^{\ddagger}$	-26.0^{*}	$-65.3^{\ddagger}$	1.2	$-65.0^{\ddagger}$	-0.2
	σ_3^2	$-35.4^{\ddagger}$	-7.6^{*}	$-31.6^{\ddagger}$	-8.1^{*}	$-35.0^{\ddagger}$	-3.4	$-28.3^{\ddagger}$	10.4^{*}
	ϕ			$-10.9^{\ddagger}$	111.7^{*}				
MARH	β_0	$-18.4^{\ddagger}$	-5.6^{*}	$-12.6^{\ddagger}$	-8.2^{*}	$-17.5^{\ddagger}$	-3.4	$-17.7^{\ddagger}$	-0.1
	β_1	$-18.5^{\ddagger}$	-24.5^{*}	$20.9^{\ddagger}$	-47.8^{*}	$-8.9^{\ddagger}$	-7.0^{*}	$-10.1^{\ddagger}$	-12.4^{*}
	β_2	$-22.0^{\ddagger}$	-8.0^{*}	$-14.9^{\ddagger}$	-13.1^{*}	$-20.6^{\ddagger}$	-1.9	$-20.8^{\ddagger}$	-3.1
	β_3	$-21.2^{\ddagger}$	-7.6^{*}	$-13.9^{\ddagger}$	-10.2^{*}	$-19.8^{\ddagger}$	-6.0^{*}	$-19.9^{\ddagger}$	-1.3
	σ_2^2	$-85.5^{\ddagger}$	-28.1^{*}	$-69.1^{\ddagger}$	-57.3^{*}	$-80.6^{\ddagger}$	-7.0^{*}	$-81.3^{\ddagger}$	-19.8^{*}
	σ_3^2	$-43.3^{\ddagger}$	-11.5^{*}	$-32.2^{\ddagger}$	-16.1^{*}	$-41.3^{\ddagger}$	-5.3^{*}	$-35.5^{\ddagger}$	8.0^{*}
	ϕ			$-11.2^{\ddagger}$	9.9^{*}				
NMARL	β_0	$-22.7^{\ddagger}$	-5.7^{*}	$-21.2^{\ddagger}$	-6.5^{*}	$-22.8^{\ddagger}$	-4.5^{*}	$-23.0^{\ddagger}$	-1.7
	β_1	$-85.2^{\ddagger}$	-12.8^{*}	$-84.1^{\ddagger}$	-16.6^{*}	$-85.0^{\ddagger}$	-10.9^{*}	$-85.0^{\ddagger}$	-11.1^{*}
	β_2	$-22.4^{\ddagger}$	-3.6	$-21.1^{\ddagger}$	-4.0	$-22.5^{\ddagger}$	-0.1	$-22.5^{\ddagger}$	-0.4
	β_3	$-21.5^{\ddagger}$	-7.1^{*}	$-20.1^{\ddagger}$	-7.3^{*}	$-21.6^{\ddagger}$	-6.3^{*}	$-21.7^{\ddagger}$	-2.4
	σ_2^2	$-86.1^{\ddagger}$	-13.3^{*}	$-82.0^{\ddagger}$	-24.6^{*}	$-84.5^{\ddagger}$	-1.2	$-84.7^{\ddagger}$	-6.3^{*}
	σ_3^2	$-43.8^{\ddagger}$	-7.0^{*}	$-41.8^{\ddagger}$	-7.6^{*}	$-44.0^{\ddagger}$	-3.2	$-38.2^{\ddagger}$	10.8^{*}
	ϕ			$-7.3^{\ddagger}$	107.1^{*}				
NMARH	β_0	$-4.8^{\ddagger}$	-6.7^{*}	$1.4^{\ddagger}$	-9.9^{*}	$-4.9^{\ddagger}$	-4.3^{*}	$-5.1^{\ddagger}$	-1.4
	β_1	$-274.6^{\ddagger}$	-12.5^{*}	$-275.7^{\ddagger}$	-18.2^{*}	$-274.3^{\ddagger}$	-13.4^{*}	$-274.4^{\ddagger}$	-13.6^{*}
	β_2	$-23.5^{\ddagger}$	0.4	$-18.8^{\ddagger}$	-2.2	$-23.2^{\ddagger}$	3.5	$-23.3^{\ddagger}$	2.6
	β_3	$-22.5^{\ddagger}$	-7.2^{*}	$-17.5^{\ddagger}$	-8.2^{*}	$-22.4^{\ddagger}$	-6.1^{*}	$-22.4^{\ddagger}$	-2.4
	σ_2^2	$-84.1^{\ddagger}$	-14.8^{*}	$-68.6^{\ddagger}$	-46.9^{*}	$-81.8^{\ddagger}$	-1.5	$-83.2^{\ddagger}$	-16.8^{*}
	σ_3^2	$-44.6^{\ddagger}$	-10.0^{*}	$-37.2^{\ddagger}$	-12.4^{*}	$-44.3^{\ddagger}$	-5.1^{*}	$-38.7^{\ddagger}$	7.4^{*}
	ϕ			$-11.8^{\ddagger}$	-21.4^{*}				

† significant bias in estimate at $P < 0.05$; ‡ significant bias in estimate at $P < 0.01$; * significant bias in standard error at $P < 0.05$

Table A4: Relative bias of estimates and standard errors to the marginal true values with a significance indication, based on analyses of 1000 simulated datasets generated by autoregressive random effects model with ($\rho = \mathbf{1, 0.9, 0.5}$) in five simulated scenarios of missing values: scc40 (missing values as in scc40 dataset), MARL, MARH (low (31%) and high (52%) proportion of missing values at random due to drop-outs), MNARL, MNARH (low (31%) and high (52%) proportion of missing values not at random). Parameters: β_0 (intercept), β_1 (time coefficient), β_2 (subject level factor), β_3 (cluster level factor). Estimation procedures: OLR (ordinary logistic regression), ALR (alternating logistic regression).

		correlation	$\rho = 1$				$\rho = 0.9$				$\rho = 0.5$			
Scen-	Parm-	procedure	OLR		ALR		OLR		ALR		OLR		ALR	
ario	eter		Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
scc40	β_0		$-5.3^{‡}$	-50.8^{*}	$-4.5^{‡}$	-4.7^{*}	$-6.4^{‡}$	-48.2^{*}	$-5.8^{‡}$	-1.2	$-5.0^{‡}$	-47.0^{*}	$-4.5^{‡}$	-3.7
	β_1		$-8.2^{‡}$	0.0	$-4.5^{‡}$	12.5^{*}	$-8.2^{‡}$	0.0	$-5.4^{‡}$	0.0	$-6.4^{‡}$	6.3^{*}	$-4.5^{‡}$	-6.3
	β_2		$-5.3^{‡}$	-36.3^{*}	$-5.3^{‡}$	-1.9	$-4.5^{‡}$	-34.5^{*}	$-4.6^{‡}$	-2.9	$-4.4^{‡}$	-22.6^{*}	$-4.4^{‡}$	-2.2
	β_3		$-3.5^{‡}$	-66.7^{*}	$-3.4^{‡}$	-3.3	$-4.6^{‡}$	-66.7^{*}	$-4.8^{‡}$	-2.9	$-4.6^{‡}$	-64.7^{*}	$-4.5^{‡}$	-2.0
MARL	β_0		$-11.5^{‡}$	-51.0^{*}	$-3.8^{‡}$	-1.8	$-12.8^{‡}$	-52.7^{*}	$-5.7^{‡}$	-5.7^{*}	$-9.7^{‡}$	-50.8^{*}	$-4.7^{‡}$	-5.1^{*}
	β_1		$-77.0^{‡}$	-6.0	$-3.7^{‡}$	-3.3	$-70.6^{‡}$	-9.7	$-3.8^{‡}$	-8.2^{*}	$-48.1^{‡}$	-6.9^{*}	$-4.4^{‡}$	-5.7^{*}
	β_2		$-8.2^{‡}$	-34.4^{*}	$-5.5^{‡}$	-2.9	$-6.7^{‡}$	-31.2^{*}	$-4.5^{‡}$	-0.3	$-5.9^{‡}$	-21.6^{*}	$-4.4^{‡}$	-3.2
	β_3		$-7.2^{‡}$	-65.3^{*}	$-4.6^{‡}$	-3.7	$-7.8^{‡}$	-65.3^{*}	$-5.6^{‡}$	-4.2^{*}	$-5.1^{‡}$	-65.1^{*}	$-3.5^{‡}$	-6.0^{*}
MARH	β_0		$-5.0^{‡}$	-42.4^{*}	$-8.8^{‡}$	-1.1	$-4.6^{‡}$	-44.2^{*}	$-7.5^{‡}$	-5.8^{*}	$-2.2^{‡}$	-42.4^{*}	$-1.5^{‡}$	-3.6
	β_1		$-200.7^{‡}$	-10.0^{*}	$28.4^{‡}$	-4.9^{*}	$-162.7^{‡}$	-13.1^{*}	$42.2^{‡}$	-9.6^{*}	$-83.7^{‡}$	-10.7^{*}	$52.6^{‡}$	-16.8^{*}
	β_2		$-16.3^{‡}$	-19.3^{*}	$-4.7^{‡}$	-1.6	$-13.3^{‡}$	11.8^{*}	$-2.8^{‡}$	2.9	$-9.4^{‡}$	-6.8^{*}	$-2.1^{‡}$	-1.7
	β_3		$-15.0^{‡}$	-54.2^{*}	$-3.6^{‡}$	-1.8	$-14.1^{‡}$	-53.9^{*}	$-3.9^{‡}$	-0.7	$-8.4^{‡}$	-55.2^{*}	$-1.2^{‡}$	-4.0
NMARL	β_0		$-11.2^{‡}$	-50.8^{*}	$-4.1^{‡}$	-1.9	$-11.2^{‡}$	-51.7^{*}	$-6.4^{‡}$	-4.8^{*}	$-6.5^{‡}$	-49.6^{*}	$-4.7^{‡}$	-4.7^{*}
	β_1		$-127.7^{‡}$	-5.9^{*}	$-78.3^{‡}$	-6.5^{*}	$-117.1^{‡}$	-11.4^{*}	$-82.5^{‡}$	-6.2^{*}	$-96.0^{‡}$	-11.7^{*}	$-81.6^{‡}$	-6.5^{*}
	β_2		$-6.7^{‡}$	-34.2^{*}	$-5.2^{‡}$	-4.0	$-5.1^{‡}$	-27.6^{*}	$-4.0^{‡}$	-1.5	$-4.5^{‡}$	-12.3^{*}	$-4.1^{‡}$	-3.7
	β_3		$-5.7^{‡}$	-65.2^{*}	$-4.1^{‡}$	-4.7^{*}	$-6.2^{‡}$	-64.1^{*}	$-5.1^{‡}$	-3.3	$-3.5^{‡}$	-64.1^{*}	$-3.1^{‡}$	-6.3^{*}
NMARH	β_0		$7.5^{‡}$	-41.2^{*}	$-12.6^{‡}$	-5.9^{*}	$8.8^{‡}$	-41.7^{*}	$11.8^{‡}$	-6.9^{*}	$-15.0^{‡}$	-40.0^{*}	$16.1^{‡}$	-5.8^{*}
	β_1		$-450.2^{‡}$	-13.7^{*}	$-317.7^{‡}$	-38.3^{*}	$-424.6^{‡}$	-12.7^{*}	$-329.7^{‡}$	-27.9^{*}	$-360.7^{‡}$	-13.9^{*}	$-320.4^{‡}$	-15.9^{*}
	β_2		$-9.1^{‡}$	-20.7^{*}	$-3.9^{‡}$	-8.7^{*}	$-7.8^{‡}$	-14.1^{*}	$-3.5^{‡}$	-5.8^{*}	$-6.2^{‡}$	-3.4	$-4.2^{‡}$	-1.7
	β_3		$-8.5^{‡}$	-53.1^{*}	$-3.2^{‡}$	-9.6^{*}	$-8.3^{‡}$	-51.0^{*}	$-4.1^{‡}$	-5.7^{*}	$-5.5^{‡}$	51.9^{*}	$-3.1^{‡}$	-8.0^{*}

† significant bias in estimate at $P < 0.05$; ‡ significant bias in estimate at $P < 0.01$; * significant bias in standard error at $P < 0.05$

Table A5: Relative bias of estimates and standard errors to the marginal true values with a significance indication, based on analyses of 1000 simulated datasets generated by autoregressive random effects model with ($\rho = \mathbf{1, 0.9, 0.5}$) in five simulated scenarios of missing values: scc40 (missing values as in scc40 dataset), MARL, MARH (low (31%) and high (52%) proportion of missing values at random due to drop-outs), MNARL, MNARH (low (31%) and high (52%) proportion of missing values not at random). Parameters: β_0 (intercept), β_1 (time coefficient), β_2 (subject level factor), β_3 (cluster level factor). Estimation procedures: WGEEci (weighted generalized estimating equations (WGEE) with independence correlation at cluster level), WGEEce (WGEE with exchangeable correlation at cluster level).

		correlation	$\rho = 1$				$\rho = 0.9$				$\rho = 0.5$			
Scen-	Parm-	procedure	WGEEci		WGEEce		WGEEci		WGEEce		WGEEci		WGEEce	
ario	eter		Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
MARL	β_0		$-3.4^{\ddagger}$	1.9	$3.2^{\ddagger}$	1.3	$-5.7^{\ddagger}$	-2.8	1.9	-2.8	$-4.5^{\ddagger}$	-4.2^{*}	$9.4^{\ddagger}$	-6.1^{*}
	β_1		$-1.5^{\dagger}$	2.3	1.4	3.3	$-1.9^{\ddagger}$	-3.5	0.4	-2.0^{*}	$-3.2^{\ddagger}$	-1.4	$-1.9^{\ddagger}$	1.3
	β_2		$-4.9^{\ddagger}$	3.8	$-4.3^{\ddagger}$	4.2	$-3.6^{\ddagger}$	5.1^{*}	$-3.1^{\ddagger}$	4.2	$-4.0^{\ddagger}$	3.0	$-3.1^{\ddagger}$	2.9
	β_3		$-4.2^{\ddagger}$	-0.6	-2.5	-3.4	$-5.2^{\ddagger}$	-1.6	$-4.0^{\ddagger}$	-3.6	$-3.3^{\ddagger}$	-4.4^{*}	$-3.5^{\ddagger}$	-3.9
MARH	β_0		$10.0^{\ddagger}$	-32.9^{*}	$-10.8^{\ddagger}$	-38.2^{*}	$3.6^{\ddagger}$	-38.2^{*}	$-7.9^{\dagger}$	-39.2^{*}	$4.0^{\ddagger}$	-31.5^{*}	3.5	-34.7^{*}
	β_1		$-37.6^{\ddagger}$	-39.1^{*}	$-32.4^{\ddagger}$	-34.6^{*}	$-36.0^{\ddagger}$	-36.6^{*}	$-29.4^{\ddagger}$	-33.4^{*}	$-22.9^{\ddagger}$	-32.3^{*}	$-14.7^{\ddagger}$	-29.3^{*}
	β_2		$-5.0^{\dagger}$	-39.3^{*}	$-8.7^{\ddagger}$	-26.0^{*}	$-2.8^{\ddagger}$	-40.4^{*}	$-7.6^{\ddagger}$	-25.7^{*}	$-2.1^{\ddagger}$	-35.6^{*}	$-3.7^{\ddagger}$	-27.2^{*}
	β_3		$-5.2^{\dagger}$	-35.5^{*}	$-8.4^{\dagger}$	-43.1^{*}	$-5.4^{\ddagger}$	-34.6^{*}	-5.6	-41.1^{*}	$-1.6^{\ddagger}$	-32.9^{*}	-1.8	-41.2^{*}

† significant bias in estimate at $P < 0.05$; ‡ significant bias in estimate at $P < 0.01$; * significant bias in standard error at $P < 0.05$