# Statistical models for binary repeated measures and hierarchical data in veterinary science

By

Elmabrok A. M. Masaoud

A Thesis

Submitted to the Graduate Faculty
in Partial Fulfillment of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

Department of Health Management
Faculty of Veterinary Medicine
University of Prince Edward Island

*Your file Votre référence*
*ISBN: 978-0-494-49860-6*
*Our file Notre référence*
*ISBN: 978-0-494-49860-6*

# Condition for the use of the thesis

Chairman of the Department of Health Management

Faculty of Veterinary Medicine

University of Prince Edward Island

Charlottetown, P. E. I.

Canada C1A 4P3

## Permission to use postgraduate thesis

Title of thesis: Statistical Models for Binary Repeated Measures and Hierarchical Data in Veterinary Science

Name of Author: Elmabrok A. M. Masaoud

Department: Health Management

Degree: Doctor of Philosophy Year: 2009

In presenting this thesis in partial fulfillment of the requirements for a postgraduate degree from the University of Prince Edward Island, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the professor or professors who supervised my thesis work, or, in their absence, by the Chairman of the Department or the Dean of the Faculty in which my thesis work was done. It is understood any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Prince Edward Island in any scholarly use which may be made of any material in my thesis.

Signature:

Address: Department of Health Management

Faculty of Veterinary Medicine

University of Prince Edward Island

550 University Avenue

Charlottetown, PE C1A 4P3

CANADA.

Date: March 25, 2009.

SIGNATURE PAGES

iv

REMOVED

*Dedicated to:*

*Family and parents*

# Abstract

The objective of the thesis was to assess the performance of statistical procedures for the analysis of binary longitudinal data in veterinary science, specifically, to describe and quantify their performance in terms of statistical properties such as unbiasedness, confidence interval coverage and efficiency. The focus was on marginal and random effects procedures including: ordinary logistic regression (OLR), alternating logistic regression (ALR), generalized estimating equations (GEE), marginal quasi likelihood (MQL), penalized quasi likelihood (PQL), pseudo likelihood (REPL), maximum likelihood (ML) and Bayesian Markov chain Monte Carlo (MCMC). The marginal and random effects procedures handle the within-subject dependence differently, and they offer different interpretations of regression estimates for binary longitudinal data. Several simulation studies covered a wide range of data structures and designs including a two-level balanced longitudinal design, a three-level balanced setting of binary repeated measures data, and repeated measures data with missing values. A statistical simulation approach was used as the tool for the assessment.

The first study involved a two-level setting of binary repeated measures data. Results for the marginal model data showed the autoregressive GEE showed to be highly efficient when treatment was within subjects, even with strongly correlated responses. For treatment between subjects, random effects methods also performed well in some situations; however, a small number of subjects with short time series proved a challenge for both marginal and random effects methods. Results for the random effects model data showed bias in estimates from random effects methods while the marginal model produced estimates close to the marginal parameters.

The second study involved binary repeated measures data with an additional hierarchical structure. Results indicate that in data generated by random intercept models, the ML and MCMC procedures performed well and had fairly similar estimation errors. The PQL regression estimates were attenuated while the variance estimates were less accurate than ML and MCMC, but the direction of the bias depended on whether binomial or extra-binomial dispersion was assumed. In datasets with autocorrelation, random effects estimates procedures gave downward biased estimates, while marginal estimates were little affected by the presence of autocorrelation. The results also indicate that in addition to ALR, a GEE procedure that accounts for clustering at the highest hierarchical level is sufficient. The REPL procedure performed poorly and produced unsatisfactory estimates regardless of autocorrelation values.

The third study involved binary repeated measures data with an additional hierarchical structure and missing values, where five different scenarios of simulated incomplete datasets were considered. The first scenario corresponded to a combination of three types of missingness patterns present in a real (scc40) dataset: delayed entry and drop-outs as well as intermittent missing values. 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 first 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, GEE 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 level were relatively unaffected. The presence of autocorrelation in the data did not substantially alter the impact of missing values although the shrinkage of random effects estimates was marginally less pronounced than in the full datasets.

Additionally, a hierarchical data structure arising in an aquaculture vaccine trial on Infectious Salmon Anaemia Virus (ISAV), where multiple treatment groups of fish in the same tanks were observed over time, was studied. The focus was to assess and account for neighbour treatment effects. By neighbour treatment effects in an incomplete block design setting, we mean that treatments present in the same block (tank) may affect each other in their performance. Two statistical models were proposed to assess and account for neighbour treatment effects. The first approach was based on a non-linear model, and the second involved cross-classified and multiple membership models. The performance of the models was evaluated by simulation.

Results demonstrated that both proposed models show promise in capturing neighbour treatment effects of the type assumed, whenever such neighbour effects are of at least moderate magnitude. Analyses of the ISAV trial data by both models did not provide any evidence of substantial neighbour effects.

# Acknowledgements

The completion of this thesis would not have been possible without the help and support of many wonderful people to whom I am indebted.

First, I would like to express my deepest gratitude to my supervisor Dr. Stryhn. His guidance, encouragement, patience, goodwill and endless support will remain in my memory forever, also and most importantly, the friendship of Henrik and his family, Jette, Ida and Emma beyond academic work.

I thank General People's Committee of Higher Education (GPCHE) in Libya, University of Seventh April, for my personal support and funding this research project.

I am deeply grateful to Dr. Dohoo, for his guidance and teachings, and endless advice. His availability and patience in reviewing manuscripts, and his critical comments have contributed so much for this thesis.

Special thanks are due to Dr. Browne, for his comments, ideas and always great advice. I am very thankful to Dr. Sanchez, who was willing to step in when I needed him, also, for letting me do part of my simulations on his computer.

Many thanks are due to the members of my supervisory committee: Dr. Greenwood and Dr. Hammell for their advice and critical discussions of my thesis.

Thanks to the staff at Department of Health Management and fellow graduate students: Tim, Ahmed, Khalid, Fiaz and Pascale for the friendly environment. Thanks are also extended to Dr. Whyte.

Special thanks to my Mom and Dad who are always proud of me despite the distance. I express my sincere thanks to my brothers (Elmahdi, Masaoud, Ibrahim, Mustfa, Mohammed and Ali), to my sisters (Aisha, Omsaad, and the late Shala) and their respective families.

Finally, and most importantly, my wonderful family deserves some sort of medal. My sons, Alaa and Diya. They constantly did well making it easy for me to keep focused on my thesis. My wife, Zainab, thank you for being patient and supportive.

# List of abbreviations

| | |
|---|---|
| ALR | Alternating logistic regression |
| ANOVA | Analysis of variance |
| AR | Autoregressive process |
| AR(1) | First order autoregressive correlation structure |
| BIBD | Balanced incomplete block design |
| BS | Between-subject design |
| CAR | Conditional autoregressive modelling |
| CC | Cross-classified model |
| CRD | Completely random drop-out |
| CI | Confidence Interval |
| DIC | Deviance Information Criterion |
| ELISA | Enzyme-linked immunosorbant assay |
| Exch | Exchangeable correlation structure |
| GEE | Generalized estimating equations |
| GEEci | GEE with independence correlation at cluster level |
| GEEce | GEE with exchangeable correlation at cluster level |
| GEEf | GEE with fixed effects for cluster level and autoregressive correlation at subject level |
| GEEs | GEE with autoregressive correlation at subject level |
| GLM | Generalized linear model |
| GLMs | Generalized Linear Models |
| GLMM | Generalized linear mixed model |
| GLMMs | Generalized linear mixed models |
| ICC | Intra-class correlation |
| ID | Informative drop-out |
| IGLS | Iterative weighted least squares procedure |
| ISAV | Infectious Salmon Anaemia Virus |
| LOCF | Last observation carried forward |
| LMMs | Linear mixed models |
| MAR | Missing at random |
| MARL | Moderate percentage of missing values at random |
| MARH | High percentage of missing values at random |
| MCAR | Missing completely at random |
| MCMC | Markov chain monte carlo |

| | |
|---|---|
| ML | Maximum likelihood |
| MLE | Maximum likelihood estimate |
| MQL | Marginal quasi-likelihood |
| MQLx | Marginal quasi-likelihood with extra binomial dispersion |
| MM | Multiple membership model |
| MMCP | Multiple membership model with correlated pairs |
| MMI | Multiple membership model with independent pairs |
| NMAR | Not missing at random |
| NLM | Non-linear mixed model |
| NMARL | Moderate percentage of missing values not at random |
| NMARH | High percentage of missing values not at random |
| OLR | Ordinary logistic regression |
| PA | Population average |
| PL | Pseudo-likelihood |
| PQL | Second order penalized quasi-likelihood |
| PQLx | Second order penalized quasi-likelihood with extra binomial dispersion |
| RD | Random drop-out |
| RBF | Relative bias to full data |
| RBT | Relative bias to true value |
| REML | Restricted maximum likelihood |
| REPL | Restricted pseudo-likelihood |
| RIGLS | Restricted iterative weighted least squares procedure |
| SS | Subject-specific |
| VPC | Variance partition coefficient |
| WGEE | Weighted GEE with independence correlation at cluster level |
| WS | Within-subject design |

# Table of Contents

# List of Tables

# List of Figures

# An overview of statistical models for binary repeated measures and hierarchical data in veterinary science

## 1.1 Introduction

Repeated measures data are data with multiple records on the same subjects (e.g., animals or farms). In multi-level terminology [77, Chapter 12], this may be termed a two-level data structure, with observations ("measures") corresponding to level one (such as tests) and subjects to level two. However, measures on the same subject are usually ordered (e.g., by time) which make such data more challenging than a two-level structure with no ordering of units within clusters (e.g., animals in farms). Such data are commonly encountered in both experimental and observational studies.

Binary repeated measures data are encountered across a wide range of applications in veterinary science and veterinary epidemiology. The most evident examples of two-level data are records of presence or absence of disease conditions over time. Disease conditions may be detected clinically (e.g., mastitis) or by a test such as bacterial culture [66], faecal egg counts [2] or antibody determination for parasites [72]. Other examples are success of fertilization (e.g., in repeated reproduction cycles [3]), occurrence of certain behaviours in animal welfare studies [35, 81], or of treatment side effects in clinical trials (e.g., treatments for diabetes in dogs [41]). If the binary outcome is created by thresholding a quantitative outcome at a predefined cut-off value (e.g., ELISA for the diagnosis of Johne's disease; [78]) a substantial loss of information is implied but the dichotomous outcome may be of greater interest than the quantitative measurement. Another range of applications occur in the context of farm-level monitoring of product quality (e.g., milk [69]).

An extension of the two-level structure arises if subjects in addition are nested within some (physical) clusters (e.g., hospitals, herds, provinces). Such structure may be termed three-level repeated measures data, where clusters correspond to level three.

Binary repeated measures data with an additional hierarchical level has formed the basis of many studies of mastitis and dairy management factors (e.g., [30, 66]). Some examples from human preventive medicine include the effects of air pollution on school absences in the southern California Children's Health study [82] and the sickness episodes for workers over time [65].

This introductory chapter is intended to give the reader an overview of the current state of knowledge on statistical theory for modeling binary repeated measures (longitudinal, time series) data with/without additional hierarchical structure. Emphasis will be placed on reviewing and assessing the existing statistical models and estimation procedures that are implemented in broadly accessible statistical software.

## 1.2 Efficiency, data structure and experimental design

This section briefly introduces the efficiency, data structure and experimental study designs that will be discussed in this chapter and throughout the thesis.

### 1.2.1 Efficiency

By efficiency we mean the ability of a statistical procedure to produce a smaller variance estimate of the effect of interest in comparison with alternative methods. Efficiency is often expressed numerically as the ratio of estimated variances for the reference ("best") method and the procedure under study. The same terminology can be applied for study designs where the choice of a proper study design may result in a reduction of the variance estimate of the treatment effect (such as in longitudinal studies versus cross-sectional studies). In comparing competing experimental designs, an efficient design is one that can achieve the same precision as other designs but with fewer resources.

### 1.2.2 Data structure

In veterinary epidemiology there are two types of study designs: observational and experimental studies [16, chapters: 7-12]. In experimental studies, subjects are randomly allocated to different comparison groups, whereas in observational studies, subjects are observed and their data is recorded. In general, experimental studies permit drawing stronger conclusions than observational studies, but often observational studies are the only visible option [24, Chapter 2]. Within the context of experimental studies, two types of data structure "repeated measures" and "hierarchical" are selected to form the basis of the data structures discussed though-out the thesis. Generally, failure to account for the consequences of such type of data structure may result in a violation of regression model assumptions and result in a poor fit of a statistical model and a questionable statistical inference ([15, Chapter 7] and [17]).

#### 1.2.2.1 Hierarchical data structure

In veterinary epidemiology, animals (subjects) within the same herd (cluster) are more alike, compared to animals from different herds. Animals within a particular herd share the experience of being in the same environment (food, management practice,..etc.) which may lead to increased homogeneity over time [16, chapter 21]. This type of data structure is called hierarchical, multilevel [77] or clustered [16, chapter 20] data structure.

The goal of multilevel analysis is to account for all the variation in the outcome, including the contributed information from each level of clustering in the data. In multilevel data, the outcome is usually measured at the lowest level of the hierarchy. One advantage of a multilevel data structure is its flexibility to allow researchers to combine multiple levels of analysis in a single comprehensive model by specifying predictors at different levels. It is also possible to include cross-level interactions to determine the dependence of lower-level predictors on higher level predictors [26]. A consequence of the hierarchical structure implies that the observations from subjects within the same cluster are similar. i.e., the same covariance structure between the measurements on subjects within the same cluster, usually termed an exchangeable covariance structure. One way to account for the similarity between responses is by modeling the covariance structure of the outcome.

A common approach for the analysis of hierarchical data with a continues outcome variable is the linear mixed model, also known as multilevel model [77], or variance component model [76]. These models account for the hierarchical structure of the data by specifying random effects for all levels above the bottom level. Then the variability in the outcome can be split into variances at different levels, i.e., each level contributes to the variation in the outcome. Goldstein *et al.* [28] presented a measure of the percentage of variability attributable to cluster over total variability, called the intra-class correlation (ICC) or variance partition coefficient (VPC). They described also how to extend VPC to binary response models. The ICC measures the degree of similarity of measurements within a cluster. It takes values between 0 and 1. Goldstein [26] suggests using "intra-unit" instead of intra-class correlation and replace unit with an appropriate term (i.e., herd, hospital, etc.).

#### 1.2.2.2 Repeated measures data structure

Repeated measures data structure exists when repeated measurements are taken on the same subject at different ordered times or various conditions [12, Chapter 2]. Longitudinal data [15] are a common form of repeated measures where measurements are recorded on subjects over a period of time. However, throughout the thesis, repeated measures and times series are used to refer to longitudinal data setting. The statistical objective in longitudinal data design is making inference about the expected value of outcomes, in terms of treatment effects and how such effects change over time. A longitudinal study design has the advantage over a cross-sectional design in that changes over time in treatment effects can be estimated [15, Chapter 1]. The positively correlated measurements per subject in longitudinal studies may reduce the variance estimate of treatment effect in comparison to cross-sectional studies. Thus the design has a potential for substantial gains in efficiency ([15, Chapter 1], [22]).

Similarly as for the hierarchical data structure, repeated measures data structure implies that the multiple measurements on the same subject are correlated. The correlation $\rho(j, j')$ between observations (e.g., at times) $j$ and $j'$ can be expressed in a range of correlation structures, including independent ($\rho(j, j') = 0$), exchangeable ($\rho(j, j') = \gamma$), and autoregressive (AR) ($\rho(j, j') = \gamma^{j-j'}$). The autoregressive process implies that correlation between the two measurements on the same subject that are close in time is higher than the two that are further apart. The within-subject dependence (as a result of the correlated observations) violates the basic assumption for simpler statistical methods that observations are independent. Similarly as to the hierarchical data structure, the within-subject dependence is usually accounted for by modelling the covariance structure [16].

#### 1.2.2.3 Repeated measures data with additional hierarchical structure

Repeated measures data with additional hierarchical structure exists when multiple records are taken over time on the same subjects (e.g., animals or farms) which are nested within some (physical) clusters (e.g., hospitals, herds, provinces). In multi-level modelling terminology [77], this may be termed three-level repeated measures data, with observations corresponding to level one and clusters to level three. Such data structures are encountered across a wide range of applications in veterinary and human epidemiology. An example of this type of data structure is the records of presence or absence of bacteria in monthly milk samples from cows housed in multiple herds. Thus, the hierarchical structure is the clustering of cows in herds, and the repeated measures are the monthly test records based on the milk samples.

Dealing with the hierarchical structure in addition to the repeated measures will at the very least increase the complexity (conceptual and numerical) of the model/analysis considerably. Some procedures (GEE in Section 1.4.2.1) were designed for two-level structures and offer no straightforward estimation to three-level structures. Other procedures (ML in Section 1.4.1.1) may be affected in their performance by the increased model complexity and size of datasets. Comparison of procedures for repeated measures with additional hierarchical structure exist for single datasets [65] but no comprehensive review has to our knowledge been undertaken.

### 1.2.3 Experimental design

Generally, one of the basic principles in experimental design is the reduction of variation between the treated units (experimental error). Often this can be achieved through the randomization of the treated units [13] and blocking groups of similar experimental units. The characteristic of an experiment usually involves, the imposition of $a$ treatments randomly to $n$ experimental units, in which their responses are measured. The experimental units can be divided into $a$ groups based on the treatments they receive, or to treatments per block, or multiple blocks of homogenous units per treatments.

#### 1.2.3.1 Treatments between subjects: parallel group design

In randomized controlled clinical trials with two treatments ($a = 2$), the eligible subjects are randomly assigned into two groups with the objective to compare the effect of the two treatments ([16, Chapter 11] and [13, Chapter 3]). The results are then analyzed by the comparison of the groups. An implication of design is that differences between the subjects contribute to the variability of measurements.

#### 1.2.3.2 Treatments within subjects: cross-over design

In a cross-over study with two treatments ($a = 2$), each eligible subject is assigned to receive both treatments in sequence, with a time period before the adminstration of the second treatment, usually termed "wash-out" [16, Chapter 11]. Each subject is randomly assigned its first treatment. Then, the outcome is monitored during each period of treatment, and in this way each subject can serve as its own control.

In a repeated measurements design, it may be of interest to randomly expose each subject in the study to a sequence of treatments to reduce the error (within subject) variance as well as to enable an unbiased estimate of treatment effects, by having each subject serve as its own control. One major advantage of the within-subjects design is that it eliminates almost all confounding effects that may be caused by the subject differences.

Another advantage of this type of treatment adminstration is that it reduces the sample size requirements. One disadvantage of this design, is the potential of confounding of the order effect of treatment. This can be usually be avoided by randomly assigning the sequence of treatments and ensuring an adequate wash-out period to eliminate the effect of one treatment on subsequent treatment(s).

#### 1.2.3.3 Incomplete block design

The design of many experimental studies may face some logistical constraints (e.g., sample size space or time limitations) to allocate ($a > 2$) treatments to all $e$ blocks. Thus only a portion of the treatments can be allocated to each block. This design is called the "incomplete block design". Designs for incomplete blocks include balanced and unbalanced block designs. Multiple types of balanced and partially incomplete designs exist. Dean and Voss [13, Chapter 11] presented balanced incomplete block design, group divisible designs, and cyclic designs. The classical balanced incomplete block design (BIBD) exists for certain combinations of the number of treatments, blocks, and block sizes. This design requires that every pair of treatments occurs together within the same block an equal number of times [13, Chapter 11]. In the group divisible design, treatments are divided into groups, and within each group the same requirement as for a BIBD is imposed. In the cyclic design, the experimental units are grouped into different blocks of different sizes, where each block is obtained from its previous block by cycling the treatments. However, many experimental studies fall into the unbalanced incomplete block design where the above requirements can not be fulfilled. In general a block design where experimental units are nested within blocks can be thought of as hierarchical, multilevel [77], clustered [16, Chapter 20] data structure. However, in certain situations, it can be considered also as a form of repeated measures data [15] structure, where blocks refer to a sequence of measurements over time.

## 1.3 Statistical models

### 1.3.1 Generalized linear models

Generalized linear models (GLMs) constitute a framework that unifies the regression models for independent outcomes [55, Chapter 4]. It consists of three main items, a distribution function $f$ that is a member of an exponential family, a linear predictor $\eta$ and a link function $g$, so a simple regression model takes the following form:

$$\mathrm{E}(y) = \mu = g^{-1}(\eta)$$

where E$(y)$ is the mean of $y$, and Var$(y)$ =V$(\mu)$=V$(g^{-1}(\eta))$.

Consider independent binary outcomes $y_i, i = 1, \ldots, n$, and a set of $p$ explanatory variables $x_{i1}, \ldots, x_{ip}$. For some specific functions $a(.)$, $b(.)$ and $c(.)$, the likelihood function of the exponential family [12, Chapter 9] takes the following form:

$$f(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i; \phi)\right\}, \tag{1.2}$$

where $\theta$ is a canonical parameter and $\phi$ is a dispersion parameter, $b(\theta)$ is called cumulant function [55, Chapter 2] that does not depend on the data. As an example, for binomial random variable $y$, with $(n, \mu)$, where $n$ is the number of trial and $\mu$ is the probability of success, the probability mass function can be expressed in the following exponential family form:

$$\begin{aligned} f(y|n, \mu) &= \binom{n}{y}\mu^y(1-\mu)^{n-y} \\ &= \exp\left[\log\binom{n}{y} + y\log\mu + (n-y)\log(1-\mu)\right] \\ &= \exp\left[y\log\left(\frac{\mu}{1-\mu}\right) - n\log(1-\mu) + \log\binom{n}{y}\right], \end{aligned}$$

where $(y\log(\frac{\mu}{1-\mu})$, $(n\log(1-\mu)$ and $\log\binom{n}{y})$ refer to $(y\theta, b(\theta)$ and $c(y))$ respectively in the likelihood function (1.2) (see, e.g., [12, Chapter 9]).

The log likelihood of (1.2) takes the form:

$$\log f(y_i;\theta_i,\phi) = \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i;\phi), \tag{1.3}$$

so,

$$\frac{\partial}{\partial\theta}\log f(y_i;\theta_i,\phi) = \frac{y_i - b'(\theta_i)}{a(\phi)},$$

then taking the expected value and equating it to zero (e.g., [12, Chapter 9]) implies that the mean is,

$$\mu_i = \frac{\partial}{\partial\theta}\log f(y_i;\theta_i,\phi) = b'(\theta_i)$$

It follows from (1.3) that the variance of $y_i$ is,

$$\mathrm{Var}(y_i) = \frac{\partial^2}{\partial\theta^2}\log f(y_i;\theta_i,\phi) = b''(\theta_i)a(\phi).$$

The linear predictor incorporates all the information for the explanatory variables $x_{i1},\ldots,x_{ip}$ into the model, i.e. $\eta_i = \beta_0+\beta_1 x_{i1}\ldots+\beta_p x_{ip}$, where $\beta$ is vector of the regression coefficient. The link function establishes a relationship between the linear predictor and the mean of the distribution by mapping the (0,1) interval into a whole real line $(-\infty,\infty)$, i.e. $\mu_i = E(y_i) = g^{-1}(\beta_0+\beta_1 x_{i1},\ldots,+\beta_p x_{ip})$. Various link functions commonly used [55, Chapter 4] for binomial distributions include: logit, probit, cloglog and log-log; the logit and probit models are discussed in more detail in Section 1.3.1.1.

Given the relationship between $\theta$ and $\beta$ through the link function and the variance function, the $f(y; \theta, \phi)$ can be expressed as $f(y; \beta, \phi)$, so the likelihood of $\beta$ and $\phi$ takes the following form:

$$l(\beta, \theta) = \exp\left\{ \frac{\sum_{i=1}^{n} \theta_i(\beta) y_i - b(\theta_i(\beta))}{a(\phi)} + c(y_i; \phi) \right\}. \qquad (1.4)$$

Then, the regression coefficients are estimated by solving the following estimating equation which equates the score function to zero:

$$S_\beta(\beta, \phi) = \frac{\partial}{\partial \beta} \log l(\beta, \phi) = \sum_{i=1}^{n} (\frac{\partial \mu_i}{\partial \beta})' Var^{-1}(y_i)(y_i - \mu_i) = 0. \qquad (1.5)$$

This corresponds to maximizing the (log) likelihood function, i.e. ML estimation. In the absence of the assumption about a full specification of the distribution belonging to the exponential family, the equation (1.5) is solved for the regression coefficients ($\beta$) by iterative weighted least squares (i.e. quasi-likelihood estimation, see 1.4.1.3).

#### 1.3.1.1 Logistic and probit regression

The logit link function defined as $g(\mu) = \log(\mu/(1-\mu))$ [5] is widely used due to its simple interpretation in terms of the odds ratio [55, Chapter 4]. Logistic regression refers to model with logit link which in its simple form can be written

$$\text{logit}(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} \ldots + \beta_p x_{ip}, \tag{1.6}$$

where $\mu_i = \text{E}(y_i) = \Pr(y_i = 1)$. Similarly, the probit link $g(\mu) = \Phi^{-1}(\mu)$, where $\Phi$ is the cumulative distribution function of the standard normal distribution $\text{N}(0, 1)$. This leads to the following probit regression model

$$\Phi^{-1}(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} \ldots + \beta_p x_{ip}, \tag{1.7}$$

The logit and probit regression models are similar. However, their regression parameter estimates are scaled: the logit coefficients exceed the probit coefficients by the approximation factor $\pi/\sqrt{3} = 1.814$. However, even adjusting regression coefficients by this factor, a slight difference between the logistic regression and the probit coefficients may still exist due to the difference between logistic and normal curves. Haley [33] showed that the logistic distribution, whose cumulative distribution function (cdf) takes the simple form $F(x) = (1 + e^{-x})$ with density function $f(x) = F(x)(1 - F(x))$, is very close to a normal distribution. Moreover, Haley showed that the maximum difference $|F(dx) - \Phi(x)|$ between the logistic cdf, $F(dx)$, with zero mean and scale parameter $d = \sqrt{3}/\pi$ and the standard normal cdf, $\Phi(x)$, is about 0.0228. Kotz [43, Chapter 22] showed graphically that the difference is minimized to 0.009 when

$d = (\sqrt{3}/\pi)(16/15)$.

### 1.3.2 Generalized linear mixed models

The generalized linear mixed models (GLMMs) extend the generalized linear models (GLMs) [55, Chapter 4] by incorporating random effects for each subject. Thus called random effects models [15, Chapter 7] or subject-specific models [31, Chapter 5]. The idea is to link the mean of the response to the linear predictor $\eta$ conditional on the random effects [56, Chapter 1], as well as to reflect the natural heterogeneity across subjects [15, Chapter 7]. Suppose we have a collection of binary observations $y_{ij}$ on each of $n$ subjects ($i = 1, \ldots, n$) at $t$ time points ($j = 1, \ldots, t$), as well as a set $x_1, \ldots, x_p$ of explanatory variables recorded for each subject at every time point. A random effects logistic regression model, often termed a logistic random intercept model, takes the following form:

$$\text{logit}(\Pr(y_{ij} = 1|u_i)) = \mu_{ij} = \beta_0 + \beta_1 x_{1ij} + \ldots + \beta_p x_{pij} + u_i, \qquad (1.8)$$

where $\Pr(y_{ij} = 1|u_i)$ is the conditional probability and $u_1, \ldots, u_n$ are independent random variables and commonly assumed normally distributed, say $u_i \sim N(0, \sigma^2)$, where $\sigma^2$ represents the heterogeneity (variance) between subjects. A more general form of the model (1.8) is to replace the single random effect $u_i$ for subject $i$ by a series of autocorrelated random effects resulting in a repeated measures random effects model [15, Chapter 11] that takes the form:

$$\text{logit}(\Pr(Y_{ij} = 1|u_{ij})) = \beta_0 + \beta_1 x_{1ij} + \ldots + \beta_p x_{pij} + u_{ij}. \qquad (1.9)$$

Where the $u_{i1}, \ldots, u_{it}$ are series of autocorrelated random effects with $\rho(u_{ij}, u_{ij'}) = \rho^{|j-j'|}$. The most commonly assumed distribution is the Gaussian (normal), say $u_{ij} \sim N(0, \sigma^2)$ where $\sigma^2$ represents the heterogeneity (variance) between subjects. Both models (1.8) and (1.9) are for the conditional probability of an "event" given the random effects of the $i$th subject. However, model (1.9) forms a better basis for random effects modelling of repeated measures data because of its ability to incorporate autocorrelation structure between the repeated measurements [15, Chapter 11].

Several alternative approaches have been suggested to allow for non-exchangeable correlation structures, such as, a multivariate approach involving estimation of all correlations between measures on the same subject [85]. However, this approach seems unsuited to deal with long time series. Another approach has been proposed to model correlations between lowest level residuals, conditional upon the random effects in (1.8), by an autoregressive function of time [4]; this method is implemented in a macro for the MLwiN software.

#### 1.3.2.1 Intra-class correlation

Typically, generalized linear mixed models provide an approximate estimate of the dependence of the outcomes $\rho(y_{ij}, y_{ij'})$ within a subject or a cluster (ICC, see 1.2.2.1). It depends on the mean (and therefore the fixed effects), the distribution of the random effects and their correlation structure. For model (1.8) with normally distributed random effects and in the absence of time-dependent predictors in the fixed effects, any two observations on the same subject are correlated to the same degree. No exact formula for the ICC is available but several approximations have been developed with the simplest of these, based on latent variable interpretation of the binary outcome [77]. By this interpretation, a binary event ($y = 1$) is created whenever a continuous latent variable exceeds a threshold. For example, a subject may succumb when its severity of disease exceeds a threshold, or a subject may become diseased when exposure exceeds a threshold. Mathematically, a binary outcome can always be represented by a latent variable and a threshold, although their interpretation can be only hypothesized. In a logistic model, the latent variable can be shown to have a logistic distribution with a variance of $\pi^2/3$. Therefore, the following formula for the ICC is exact for the latent variable and may be used as an approximation for the observed binary outcome:

$$\mathrm{ICC} \approx \frac{\sigma^2}{\sigma^2 + \pi^2/3}. \tag{1.10}$$

Likewise, in model (1.9) the correlation between two observations $k$ time steps apart can be expressed approximately as,

$$\mathrm{ICC} \approx \frac{\rho^k \sigma^2}{\sigma^2 + \pi^2/3}, \quad \text{where } k = 1, \ldots, t-1. \tag{1.11}$$

### 1.3.3 Generalized linear marginal models

Marginal (population-averaged or PA) models [86] are expressed in terms of the marginal expectation of the outcomes without conditioning on the random effects. Then the marginal expectation (or probability of an event) is modeled as a function of the explanatory variables and regression parameters through the link function in a GLM. A marginal logistic regression model takes the following form:

$$\mathrm{logit}(\mu_{ij}) = \eta_{ij} = \beta_0 + \beta_1 x_{1ij} + \ldots + \beta_p x_{pij}, \tag{1.12}$$

where $\mu_{ij} = \mathrm{E}(y_{ij}) = \Pr(y_{ij} = 1)$, note that in marginal and random effects models the regression parameters are not equal and their effects have different interpretations. A method to scale random effects parameters to marginal parameters is available (see Section 1.5.1).

The avoidance of such scaling by separating fixed and random effects estimates was one of the key ideas behind the development of marginalized models [34]. Marginalized models employ the marginal model (1.12) for the regression coefficients and the random effects model (1.8) for the correlation structure. In the latter model, the fixed part is replaced by the equivalent of the marginal model fixed part for the conditional probability [34]. Marginalized models may be programmed using flexible statistical optimization tools (such as the `nlmixed` procedure in SAS; [32]), but to our knowledge these models are not yet available in standard statistical software, or as an add-on package.

## 1.4 Statistical estimation procedures

### 1.4.1 Random effects estimation procedures

The likelihood contribution of subject $i$ in model (1.8) involves an integral over the random effect distribution, and takes the following form:

$$l(\beta, \sigma_u^2) = \int_{-\infty}^{+\infty} \prod_{j=1}^{t} e^{(\eta_{ij}+u_i)y_{ij}} (1 + e^{\eta_{ij}+u_i})^{-1} \frac{1}{\sqrt{2\pi\sigma_u}} e^{\frac{-1}{2\sigma_u^2} u_i^2} du_i. \quad (1.13)$$

In general, there is no analytic expression available for the equation (1.13) and a numerical procedure is needed. Alternatively several approximation algorithms have been proposed aimed at producing estimates close to the global ML estimate without actually computing the likelihood function [6]. These algorithms carry a number of different names and acronyms typically involving “weighted least squares” and “quasi”- or “pseudo-likelihood”.

#### 1.4.1.1 Maximum likelihood estimation via numerical integration

The exponential part in equation (1.13) makes the Gauss-Hermite quadrature procedure [53] a logical method to evaluate it numerically. The adaptive quadrature procedure is preferable for normally distributed random effects [68]. In adaptive quadrature, the quadrature points are rescaled and shifted to the shape of the log likelihood function. ML estimation by numerical integration for model (1.8) has become available in several statistical packages in recent years, the most flexible of these being the (`gllamm`) macro for Stata for latent variable models (including the generalized linear mixed models) [67]. In addition, Stata offers (`xtmelogit`) procedure for multilevel models.

#### 1.4.1.2 Markov Chain Monte Carlo

The Bayesian statistical framework is based on the well-known Bayes theorem [23]. One major distinction from classical (frequentist) statistics is that in Bayesian statistics the parameters are stochastic variables with prior and posterior distributions. Our interest is in the full posterior distribution, which depends on the likelihood function and the prior distribution over the unknown parameters in the model of interest. For a density function $p$, parameter $\theta$, observed data $D$, and a prior distribution $p(\theta)$, the posterior distribution takes the following form:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} \tag{1.14}$$

where $p(D|\theta)$ is the likelihood function. Markov chain Monte Carlo (MCMC) offers techniques to generate samples from Markov chains which in a wide variety of models can be devised to converge to the posterior distribution of $\theta$ (for more details, see e.g., [8]). Our focus here is on using MCMC techniques as an estimation algorithm for the frequentist model (1.8), rather than exploring genuine Bayesian models with informative prior distributions. By this approach, prior distributions are generally taken as vague ("non-informative"), and the inference is based on posterior distributions using a posterior median (or mean) as a substitute for a maximum likelihood estimate and 95% probability intervals instead of confidence intervals. A common non-informative prior for the fixed effects is $\mathrm{N}(0, 10^6)$, for inverse variances or precisions gamma $(10^{-3}, 10^{-3})$ [9] or for the standard deviation a uniform distribution (0,100)) [46, 23].

MCMC techniques exist to compute "real" maximum likelihood estimates (see e.g., [29, Chapter 14]) but these are beyond the present scope. The MCMC approach avoids computation of the full likelihood function, and has been shown to perform well across a range of settings including multilevel random intercept models [9]. The essential statistical software for Bayesian analysis is WinBugs; in addition, a range of multi-level models can be fitted in MLwiN.

The flexibility of MCMC allows us to implement complicated models. Congdon [11, Chapter 7] describes one way of constructing a series of autocorrelated random variables, such as $(u_{i1}, \ldots, u_{it})$ in model (1.9) for MCMC analysis,

$$u_{ij} = \rho u_{ij-1} + \epsilon_{ij} \tag{1.15}$$

where $\epsilon_{ij}$ is an uncorrelated random variable $\sim N(0, \sigma_\epsilon^2)$, $u_{ij} \sim N(0, \sigma^2)$ and $u_{i0} \sim N(0, \sigma_0^2)$. The correlation between $(u_{ij}, u_{ij-t})$ is established through the variances, where a first order autoregressive process is assumed, i.e. $\sigma_0^2 = \sigma_\epsilon^2(1-\rho^2)$ and $\sigma^2 = \rho\sigma_0^2$ (for more details, see e.g., [11, Chapter 7]).

#### 1.4.1.3 Quasi-likelihood method

Quasi-likelihood is a term used to describe a function that has similar properties to the likelihood function (1.4), but without being strictly derived from a probability distribution. The quasi-likelihood requires a known specification of a relation between the mean and the variance of the observations, i.e., for a set of independent binary variables $y_1, \ldots, y_n$, the $\mathrm{Var}(y_i) = V(\mu_i) = \phi\mu_i(1-\mu_i)$ [56, Chapter 5]. $\phi$ is a scale (or dispersion) parameter, and is usually estimated from the data. Alternatively, one may fix the scale parameter to a value of 1 to reflect the actual relationship in the binomial distribution. McCullagh and Nelder [55, Chapter 9] refer to the following integral (if it exists) as the log quasi-likelihood for $\mu_i$ given $y_i$:

$$Q(\mu_i, y_i) = \int_{y_i}^{\mu_i} \frac{y_i - \mu}{\phi V(\mu)} d\mu. \tag{1.16}$$

Then, the regression coefficients are estimated by solving the following estimating equations which equate the $j$th element of the score function to zero:

$$S_\beta(\beta, \phi) = \sum_{i=1}^{n} \frac{\partial}{\partial \beta_j} Q(\mu_i, y_i) = \sum_{i=1}^{n} (\frac{\partial \mu_i}{\partial \beta_j}) \frac{y_i - \mu_i}{\phi V(\mu_i)} = 0, \quad j = 1, \ldots, p \tag{1.17}$$

the parameter $\phi$ can be estimated separately using

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\mu_i)} = \frac{\chi^2}{n-p},$$

where $\chi^2$ is the generalized Pearson statistic (e.g., [55, Chapter 2]). The inclusion of the scale factor $\phi$ in quasi-likelihood models, give them the ability to directly accommodate overdispersion, and since $\phi$ is constant, equation (1.17) is identical to equation (1.5).

For model (1.8), an iterative weighted least squares procedure iteratively applies mixed linear model estimation to an "adjusted" variate obtained by Taylor approximation of the outcome around its current estimated mean, until convergence, using either ML or REML, thus results in IGLS "iterative generalized least square" or RIGLS "restricted iterative generalized least square", respectively. The resulting regression coefficient estimates are called maximum quasi-likelihood estimates because they can be obtained from optimizing a quasi-likelihood function which only involves first and second order conditional moments, augmented with a penalty term on the random effects [59]. Breslow and Clayton [7] presented two estimation procedures based on quasi-likelihood function called penalized quasi-likelihood (PQL) and marginal quasi-likelihood (MQL). The MQL estimates are derived under random effects model assumptions [25]. Both procedures use Laplace approximation to approximate the likelihood function. One major difference between the two algorithms is that MQL does not incorporate the random effects $u_i$ in the linearization of the mean [59, Chapter 14]. It has been suggested to refine the approximations by including a second-order term in the Taylor expansions, usually denoted as second order PQL and MQL procedures [27, 71]. These quasi-likelihood algorithms are implemented in MLwiN by adapting an iterative generalized least squares for binary series by combination of Taylor series approximation.

#### 1.4.1.4 Pseudo-likelihood method

Pseudo-likelihood (PL) is a term used to describe a function of the data that has similar properties to the likelihood function (1.4) ([59, Chapter 9] and [84]). Wolfinger and O'Connell [84] suggest a pseudo-likelihood approach for generalized linear mixed models, based on Gaussian approximation and Taylor's theorem. It differs from the quasi-likelihood approach by using a true joint likelihood function in an iterative estimation process. It involves assuming the regression parameters are known, then applies a linear mixed model to estimate the dispersion $\phi$ and the variances parameters, and then assume the variances are known and estimates the regression parameters. The use of either ML or REML in the linear mixed model estimation process, resulted in either pseudo-likelihood (PL) or restricted pseudo-likelihood (REPL), respectively [84]. The (restricted) pseudo-likelihood approach allows for both random effects in the linear predictor and correlation structure in the observations scale errors conditional (on the mean) [84]. Intuitively, one would expect this procedure to be suitable for models such as model (1.9). Modelling by correlation structure only yields marginal estimates [59]. Adding random effects effectively yields a random effects model with serial correlation [59, Chapter 22]. The (restricted) pseudo-likelihood approach is implemented in SAS (`proc glimmix`) and R (`glmmPQL` library), in addition to the (restricted) pseudo-likelihood approach, the SAS procedure offers first order PQL and MQL estimation ([84] and [59, Chapter 15]).

### 1.4.2 Marginal estimation procedures

Likelihood-based marginal approaches do exist, namely, Dale and Bahadur models (see, e.g., [59, Chapter 7] and [15, Chapter 8]). However, these approaches became unattractive due to the extensive computational requirements. In Bahadur model [15, Chapter 98], the within-subject dependence is captured via marginal correlations. One drawback of this model is that, the correlations among binary responses are constrained by marginal means [15, Chapter 8] and the parameters increase rapidly with $t$ (the number of repeated measurements per subject).

Two alternative and more common approaches for longitudinal data are generalized estimating equations (GEE) [48, 86, 36] and alternating logistic regression (ALR) [10]. These procedures are often referred to as semi-parametric because they do not assume a specific form of the dependence between observations on the same subject, i.e. the within-subject correlation structure. Both GEE and ALR estimation yield PA estimates that are asymptotically unbiased and can be nearly efficient relative to the maximum-likelihood estimate in a fully and correctly specified model [15, Chapter 8]. The GEE procedure is available in most major statistical packages (e.g., SAS, S-Plus/R and Stata), with only slight differences in their implementation, and ALR is available in the former two packages.

#### 1.4.2.1 Generalized estimating equations

The GEE extension to generalized linear models for the analysis of longitudinal data was introduced in a series of papers [48, 86]. For binary observations $y_{ij}$ on each of $n$ subjects ($i = 1, \ldots, n$) at $t$ time points ($j = 1, \ldots, t$), as well as a set of explanatory variables $x_1, \ldots, x_p$, the "estimating equation" takes the following form:

$$S_\beta(\beta, \phi) = \sum_{i=1}^{n} \sum_{j=1}^{t} (\frac{\partial \mu_j}{\partial \beta_i}) \frac{y_{ij} - \mu_{ij}}{\phi V(\mu_{ij})} = 0, \tag{1.18}$$

the parameter $\phi$ can be estimated separately using

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \sum_{j=1}^{t} \frac{(y_{ij} - \hat{\mu}_{ij})^2}{V(\mu_{ij})},$$

where $\hat{r}_{ij} = (y_{ij} - \hat{\mu}_{ij}) / \sqrt{V(\mu_{ij})}$ is the Pearson residual. The solution of the multidimensional estimating equation (1.18) determines the parameter estimates, usually obtained in a stepwise (iterative) manner where an iterated and updated equation between regression and within-subject dependence estimates is solved in each step and the process terminates when the solution no longer changes ("convergence"). Specifically, the GEE procedure involves a user-specified "working" correlation matrix to approximate the true within-subjects correlation structure. When using a robust variance estimation method ("Huber/White" or "sandwich", [36]) the statistical properties for the PA estimates hold even for a misspecified working correlation structure; this is often referred to as a robustness property of the GEE procedure. However, a correct specification of the correlation structure enhances its efficiency [83]. Most software implementations offer a range of correlation structures, including independent, exchangeable, and autoregressive (AR) (see Section 1.2.1.2). The estimated Pearson residual $\hat{r}_{ij}$ is used to estimate the correlation [36, Chapter 3]. In an autoregressive correlation structure ($\rho(j,j') = \gamma^{j-j'}$), one way of estimating the scalar $\gamma$ is by the following equation [73]:

$$\hat{\gamma} = \frac{1}{(n(t-1)-p)\hat{\phi}} \sum_{i=1}^{n} \sum_{j=1}^{t-1} \hat{r}_{i,j} \hat{r}_{i,j+1},$$

where $p$ in the number of fixed effects parameters. The correlation matrix can then be built from the autoregressive structure implied by the AR correlation [36, Chapter 3]. GEE is limited to the classical two-level settings in repeated measures data.

#### 1.4.2.2 Alternating logistic regression

Generally, the correlations among binary data are constrained by the (marginal) probabilities [64]. Thus the GEE estimates of the association among the binary outcomes can be inefficient [10]. To overcome this problem Carey *et al.* [10] suggested using pairwise odds ratios to model the association between pairs of the outcomes and they proposed a procedure called alternating logistic regression (ALR). It refers to a procedure that iterates between a logistic regression using GEE to estimate regression coefficients and a logistic regression for modeling within-subject dependence in terms of pairwise odds ratios. For binary observations $y_{ij}$ on each of $n$ subjects $(i = 1, \dots, n)$ at $t$ time points $(j = 1, \dots, t)$, the odds ratio parameter [10] for each unique pair of outcomes within subjects $(y_{ij}, y_{ij'})$ takes the following form:

$$\psi_{j,j'} = \frac{\Pr(y_{ij} = 1, y_{ij'} = 1)\Pr(y_{ij} = 0, y_{ij'} = 0)}{\Pr(y_{ij} = 1, y_{ij'} = 0)\Pr(y_{ij} = 0, y_{ij'} = 1)}, \tag{1.19}$$

The ALR approach has the same robustness properties of the GEE procedure with respect to regression parameters, and is considered efficient in estimating the association parameter [10]. The ALR procedure has the advantage of providing standard errors for the association parameters $\psi$ between the pairs of responses, and is numerically more efficient than GEE for large clusters [10]. The ALR has the ability to accommodate up to three levels of hierarchical structure, where it allows one to distinguish between odds-ratios within clusters and within subjects; however, both the within-cluster correlation and the within-subject correlation must be modelled as exchangeable. To illustrate by a numerical example, an ALR analysis of the scc40 dataset of [16, Chapter 27] gave a common log odds ratio within subjects (cows) of 2.27 (OR = 9.68) and a common log odds-ratio for within clusters (herds) of 0.22 (OR = 1.25). The within-subject pairwise odds-ratio is relating two observations from the same cow, and a value of 9.68 suggests a positive outcome in a cow at one time point increases the odds for a positive outcome at another time point (*in the same cow*) almost 10-fold. In essence, some cows are at higher risk of a positive outcome than others. The within-cluster pairwise odds-ratio of 1.25 indicates that a positive outcome in a cow increases the odds of a positive outcome in another cow (*in the same herd*) by 25%. In essence, this corresponds to clustering of positive outcomes in farms.

## 1.5 Relationship and performance of marginal and random effects models

### 1.5.1 Relationship between marginal and random effects models

The relation between random effects and marginal estimates has been discussed and described [86, 60]; see also the summary by Diggle *et al.* [15]. The inferential goal of a marginal model is the marginal probability (averaged across the population of subjects), thus provides a population average interpretation of the estimates. On the other hand the inferential goal of the random effects model is the probability conditional on the unobserved (subject) random effects. This provides a subject-specific interpretation of the estimates. Zeger *et al.* [86] provided a conversion formula for logistic regression with normally distributed random effects:

$$\beta^{PA} \approx (c^2\sigma^2 + 1)^{-1/2}\beta^{SS}, \quad \text{where} \;\; c = 16\sqrt{3}/(15\pi) = 0.588. \tag{1.20}$$

For a probit model, the above conversion formula becomes an exact formula (see, e.g., [56, Chapter 8]):

$$\beta^{PA} = (\sigma^2 + 1)^{-1/2}\beta^{SS}. \tag{1.21}$$

Both formulas can be used to relate subject specific to population average models/estimates under the assumption that random effects are normally distributed. Without any distributional assumptions on the random effects it holds that the marginal regression parameters are attenuated or diluted (towards zero) relative to the random effects parameters, unless the variance is zero [56, Chapter 8].

### 1.5.2 Performance of random effects estimation procedures

The performance of random effects estimation procedures rely on the ability of the statistical algorithm to approximate the log likelihood function. The estimation procedures based on adaptive quadrature to maximize the log likelihood (ML) ([63, Chapters: 2-4] and [62, 68]) are preferred and produce reliable estimates of the regression parameters. However, caution should be taken in their use because “even with adaptive Gaussian quadrature and with relatively simple models, convergence to a global maximum can be difficult to obtain” [47]. Rabe-Hesketh *et al.* [68] showed that adaptive quadrature to approximate the integral for maximum likelihood performs better than PQL. The performance of MCMC as a maximum likelihood estimation procedure was evaluated by Browne and Draper [9], they found that MCMC produced the closest reproduction of true model values in comparison with PQL and MQL.

In many studies, PQL showed a tendency to give biased estimates [71]; in particular, the variance components were biased towards zero [31]. The PQL procedure was shown to perform poorly in datasets with small numbers of repeated measurements per subject [59, Chapter 14], and an improvement was noticed by increasing both the number of subjects as well as the number of measurements per subject. One study [18] reported the performance of REPL with a focus on the variance parameters. This study indicated that REPL suffers from convergence problems and produces biased estimates for the interclass correlation, especially for a small number of subjects with a small number of repeated measurements. However for large numbers of clusters, it seems to converge to steady but biased estimates especially when the variance is large.

### 1.5.3 Performance of marginal estimation procedures

The performance of GEE has been studied by many researchers over the last decade. In summary, the use of an independent working correlation in GEE provides highly efficient regression estimates [86]. Pepe and Anderson [61] reported that the use of non-independent working correlations may lead to biased regression estimates and indicated that there is an advantage in using the independent structure for models that include time-varying covariates. However, Fitzmaurice *et al.* [20] found that the independence structure may lead to a substantial loss of efficiency for models including time-varying covariates. Sutradhar and Das [79] have shown that the use of misspecified correlation structures in GEE leads to loss of efficiency for regression estimates. Different studies ( e.g., [54, 87, 83, 74]) have shown that the independent working correlation produces efficient estimates only for very restricted cases and are subject to a substantial loss in efficiency even when the design is balanced. Wang and Carey [83] concluded that the choice of working correlation in GEE has a substantial impact on the efficiency of regression estimates. They recommended the choice of the working correlation should coincides with the true correlation of the data and can be chosen based on either statistical criteria or biological background. Wang and Carey [83] recommended also carrying out a simulation study based on the covariate structure to evaluate the impact of the working correlation in practical data analysis.

Breslow and Clayton [7] showed that MQL is a marginal procedure. Nevertheless, the performance of MQL has been studied by many researchers as a random effects procedure (e.g., [71, 9]). Rodríguez and Goldman [71] reported in their simulation that MQL produce biased regression estimates and underestimated variances. Browne and Draper [9] demonstrated that MQL performed worse than PQL when the random effects variances are large. MQL was reported to perform poorly in datasets with a small number of repeated measurements per subject [59, Chapter 14]. Goldstein and Rasbash [27] and Rodríguez and Goldman [71] showed that second order MQL performs only slightly better than first order MQL.

### 1.5.4 Comparison of marginal and random effects procedures

For binary repeated measures outcomes, random effects and marginal estimation procedures handle the within subject dependence differently and provide different parameter estimates with different interpretations. In the context of a longitudinal smoking prevention trial, Hu *et al.* [39] compared the traditional stratified analysis, ordinary logistic regression, random effects logistic model and GEE. They reported that the absolute values of the random effects estimates were larger than those from GEE models. They indicated that the correlation between the repeated measures play a role in the discrepancy between the estimates from the two models. They also reported that the marginal estimates of the fitted random effects models (random effects estimates converted using 1.20) were similar to GEE estimates. In the context of longitudinal comparative studies, Kuchibhatla and Fillenbaum [44] compared three procedures, ordinary logistic regression, random intercept model and GEE. They reported that the absolute values of the random intercept estimates and their standard error were larger than those from the ordinary logistic and GEE models. The ordinary logistic regression under- and over-estimated the standard errors of time invariant covariates and time varying covariates, respectively. However, an argument regarding these findings can be made that these differences may be due to the difference between the subject-specific and population average estimates. Preisser *et al.* [65] presented a comparison of ALR, GEE and random-effects logistic regression for analysis of a single dataset on patterns of occupational illness. They reported that ALR is a useful method for estimating the regression parameters and detecting the clustering in longitudinal data.

In general, the choice of procedure, in particular the choice between marginal and random effects procedures should first and foremost be guided by the desired interpretation of effects. Diggle *et al.* [15, Chapter 7] argue that PA effects are of primary interest in clinical trials because "the average difference between control and treatment is most important, not the difference for any one individual". Lindsey and Lambert [49] warn that the population average may hide individual effects, and that "in extreme cases, a marginal analysis can show an average positive treatment effect when the effect would in fact be judged negative for each individual".

## 1.6 Missing values

By missing values in binary repeated measures data we mean data with incomplete records over time on the same subjects (e.g., animals or farms). 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. Missing data in experimental studies may occur by design where some logistical restrictions force an unbalancedness of the data, such as in the incomplete block design.

Generally, missingness in longitudinal data presents a potential source of bias. In part, the bias could be due to the changes in data structure from being balanced to being unbalanced, which in turn may raise technical difficulties, especially for those statistical methods that can only cope with balanced data [15, Chapter 13]. If the process of the observations being missing (the missingness mechanism) varies from subject to subject, the distribution of the observed data may not be the same as for the full data.

### 1.6.1 Classification of missing data

Despite the large body of literature on missing data [52, 45, 14, 19, 51, 37, 38], most authors agree that handling missing values is not a trivial task and that in many instances there is a need for sensitivity analyses [40]. Thus, additional information about the missingness mechanism is required. Missing data mechanisms have been classified into three categories [52]: missing completely at random; missing at random; not missing at random.

Within the context of binary repeated measures data, let $y_{ij}$ refer to complete binary records on each of $n$ subjects ($i = 1, \ldots, n$) at $t$ time points ($j = 1, \ldots, t$). Furthermore, let $y_{ij} = (y_{ij}^o, y_{ij}^m)$ where $y_{ij}^o$ is the observed subset of the data, and $y_{ij}^m$ is the subset of the data that would have been available had they not been missing. Note that the $y_{ij}^m$ is therefore unobserved or latent. Let $r_{ij}$ be an indicator of missing $y_{ij}$. Little and Rubin [52] consider the conditional distribution $f(r_{ij}|y_{i.}, x_{i.}, \phi)$ for $r_{ij}$ given $y_{ij}$ where $y_{i.}$ represents all the intended repeated measurements of the response of subject $i$, and $x_{i.}$ is for all repeated measurements of a particular predictor for subject $i$. The $\phi$ denotes unknown parameter(s) involved in the modeling of the missing data process.

In the above notation, a subject $i$ drops out from the study at time $d$, if $r_{id-1} = 0$ and $r_{ij} = 1$ for all $j \geq d$.

Missing completely at random (MCAR) [52, 45] refers to a missing data mechanism that does not depend on either prior observed or unobserved outcome values. Then, the conditional distribution for $r_{ij}$ takes the form: $f(r_{ij}|y^o_{i.}, y^m_{i.}, x_{i.}, \phi) = f(r_{ij}|x_{i.}, \phi)$. Little and Rubin [52] indicated that under a large sample assumption, the maximum likelihood estimator obtained from the observed data is equivalent to that obtained from the full dataset, i.e. the missing process can be ignored.

Diggle and Kenward [14] introduced a completely random drop-out (CRD) process that assumes missing completely at random. One implication of the MCAR assumption is that the distribution of the observed outcomes 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 outcomes is unaffected by the drop-out.

Missing at random (MAR) [52, 45] or random drop-out (RD) [14] refers to a missing data (drop-out) process that depends on the observed values only, (i.e., there are no unknown or unmeasured factors that influence the probability of an observation being missing). In this case the conditional distribution for $r_{ij}$ takes the form: $f(r_{ij}|y^o_{i.}, y^m_{i.}, x_{i.}, \phi) = f(r_{ij}|y^o_{i.}, x_{i.}, \phi)$. Little and Rubin [52] showed how to simplify the full likelihood function of the model data when MAR holds. They concluded that under a large sample assumption, the maximum likelihood estimator obtained from the observed data is equivalent to that obtained from the full dataset. Diggle and Kenward [14] proposed a logistic model for the drop-out process:

$$\text{logit}(\Pr(r_{ij} = 1)) = \beta_0 + \beta_1 \text{time}_j + \beta_2 y_{ij-1}, \tag{1.22}$$

where $\Pr(r_{ij} = 1)$ is the probability that subject $i$ drops out at time $j$.

Not missing at random (NMAR, sometimes also MNAR) [52, 45] or informative drop-out (ID) [14] refers to a drop-out missing data mechanism that depends on the unobserved outcome (current or future missing values). The conditional distribution for $r_{ij}$, $f(r_{ij}|y^o_{i.}, y^m_{i.}, x_{i.}, \phi)$, does not permit any reduction. Little and Rubin [52] indicated that inference based on the likelihood function ignoring the missing data mechanism is biased and concluded that a NMAR missing process can not be ignored under likelihood inference. Contrary to MAR, the NMAR process implies that the distribution of outcomes prior to a drop-out is not the same for those subjects who drop-out and those who do not.

### 1.6.2 Impact of missing values

The impact of missing values has been studied and several approaches have been proposed to handle it. These approaches range from imputation to statistical modeling. Several imputation algorithms have been proposed, including last observation carried forward (LOCF); unconditional mean imputation [52]; and conditional mean imputation [59, Chapter 27]. They all make the strong assumption about the data missing process to be completely at random, which may not be always the case. The simplicity of these approaches is one motivation behind their use.

Several approaches have been proposed to assess and account for missing values [19], including the complete case method (also termed "listwise deletion" [58, Chapter 5]). By this method, subjects with at least one missing value are dropped from the analysis. Fitzmaurice [19] and Little and Rubin [52] 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 (or "pairwise deletion" [58, Cahpter 5] and [52, 19]). Fitzmaurice [19] argued that a weighted version of GEE (WGEE) falls under this approach. Kim and Curry [42] showed that for a 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 [50] and Little and Rubin [52] explained that these methods assume the strong MCAR process. Little and Rubin [52] argued that neither the complete case method nor the available case method is generally satisfactory.

Little and Rubin [52] showed that an MAR process can be ignored when using likelihood-based inference. Hogan *et al.* [38] defined ignorability as the situation where “the missing data model can be left unspecified or ignored”. The GEE estimation procedure [48, 86] requires the stronger assumption MCAR about missing values. Robins *et al.* [70] showed that ordinary GEE does not allow a 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 Janson *et al.* [21]. In brief, the weight for each subject can be calculated by fitting a marginal logistic regression for the binary indicators of previous drop-outs. Then the predicted values from this model can be used to compute probability weights $w_{ij}$ for the subject $i$, as the inverse probabilities of not dropping out up to the current time point. By introducing a probability weight $w_{ij}$ into the estimating equation (1.18) this leads to the following:

$$S_\beta(\beta, \phi) = \sum_{i=1}^{n} \sum_{j=1}^{t} (\frac{\partial \mu_j}{\partial \beta_j}) \frac{y_{ij} - \mu_{ij}}{\phi V(\mu_{ij})} w_{ij} = 0, \tag{1.23}$$

Fitzmaurice [19] argued that WGEE falls under the available case method, because it uses only the observed data. Ali and Talukder [1] demonstrated the application of weighted GEE for MAR, and GEE for MCAR; they concluded that WGEE is valid for MAR. Touloumi *et al.* [80] reported that the degree of bias in GEE estimates increases with the severity of non-randomness and with the proportion of MAR data.

## 1.7 Purpose and overall objective

The purpose of this research project was to assess the performance of statistical procedures belonging to marginal and random effects models for the analysis of binary longitudinal data in veterinary science, specifically, to describe and quantify their performance in terms of statistical properties such as unbiasedness, confidence interval coverage and efficiency.

In summary, binary records made on the same subjects over time are likely to be correlated [57, 75] or clustered [16]. A within-subject dependence violates the basic assumption of logistic regression that observations are independent, and may, if not accounted for, lead to biases in parameter estimates and standard errors ([15, Chapter7] and [17]). Such data structures challenge the statistical methods to hold its properties, such as asymptotic unbiasedness and nominal confidence interval coverage. Marginal and random effects procedures (models) ([15, Chapter 7-9], [60]) have been proposed for the analysis of binary repeated measures data. However, none of these approaches and methods combines perfectly with an additional hierarchical structure.

We will motivate and illustrate all aspects of these models in veterinary epidemiology research. In this thesis we will discuss the performance of these statistical models through simulation studies in the context of binary repeated measures with/without additional hierarchical structure. Emphasis will be placed on assessing the existing statistical methods through simulation studies. In order to realistically reflect the choice an applied researcher faces when it comes to data analysis, only procedures implemented in broadly accessible statistical software are included. The goal of the assessment is to establish some practical guidelines for the choice of statistical procedures for the analysis of longitudinal binary repeated measures data in veterinary science.

The overall objective of this thesis is to carry out a statistical assessment and comparison of marginal and random effects procedures, in terms of statistical properties such as unbiasedness, confidence interval coverage and efficiency. In addition the study will explore the effect of design parameters such as the length of time series, the hierarchical structure, the number of replicate subjects, the level at which the treatments are applied (between versus within subjects), and the impact of missing values, in a longitudinal design. There are four specific objectives:

1: The first objective is to give a statistical assessment of marginal and random effects procedures, in terms of properties such as unbiasedness, efficiency and confidence interval coverage, in a two-level balanced longitudinal design (Chapter 2).

2: The second objective is to explore and compare marginal and random effects estimation procedures for the analysis of binary repeated measures data with additional hierarchical structure (Chapter 3).

3: The third objective is to assess the impact of missing values on the performance of different statistical estimation procedures for the analysis of binary repeated measures data with additional hierarchical structure (Chapter 4).

4: The fourth objective is to explore statistical approaches to assess and account for specific correlation structures in hierarchical data arising from incomplete experimental designs (Chapter 5).

## 1.8 References

## References

[1] Ali, M. W., Talukder, E., 2005. Analysis of longitudinal binary data with missing data due to dropouts. *Journal of Biopharmaceutical Statistics*, **15**, 993–1007.

[2] Anderson, K., Brooks, A. S., Morrison, A. L., Reid-Smith, R. J., Martin, S. W., Benn, D. M., Peregrine, A. S., 2004. Impact of *Giardia* vaccination on asymptomatic *Giardia* infections in dogs at a research facility. *Canadian Veterinary Journal* **45**, 924–930.

[3] Averill, T., Rekaya, R., Weigel, K., 2006. Random regression models for male and female fertility evaluation using longitudinal binary data. *Journal of Dairy Science* **89**, 3681–3689.

[4] Barbosa, B., Goldstein, H., 2000. Discrete multilevel response models. *Quality and Quantity* **34**, 323–330.

[5] Berkson, J., 1951. Why I Prefer Logits to Probits. *Biometrics* **7**, 327–339.

[6] Breslow, N. E., 2003. Whither PQL?. University of Washington Biostatistics. Working Paper Series **192**.

[7] Breslow, N. E., Clayton, D. G., 1993. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.

[8] Browne, W. J., 1998. Applying MCMC Methods to Multi–level Models. PhD thesis University of Bath.

[9] Browne, W. J., Draper, D., 2006. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis* **3**, 473–514.

[10] Carey, V., Zeger, S. L., Diggle, P., 1993. Modeling multivariate binary data with alternating logistic regressions. *Biometrika* **80**, 517–526.

[11] Congdon, P., 2002. *Bayesian Statistical Modelling.* Wiley, New York.

[12] Davis, C., 2002. *Statistical Methods for the Analysis of Repeated easurements.* Springer-Verlag, New York, Inc.

[13] Dean, A. M., Voss, D. T., 1999. *Design and analysis of experiments.* Springer-Verlag Inc., New York.

[14] Diggle P., Kenward M. G., 1994. Informative dropout in longitudinal data analysis. *Applied Statistics* **43**, 49–93.

[15] Diggle, P. J., Heagerty, P., Liang, K.-Y., Zeger, S. L., 2002. *Analysis of Longitudinal Data*, 2nd ed., Oxford University Press, Oxford.

[16] Dohoo, I. R., Martin, S. W., Stryhn, H., 2003. *Veterinary Epidemiologic Research.* AVC Inc., Charlottetown, Canada; web-site: `http://www.upei.ca/ver`.

[17] Dohoo, I. R., Stryhn, H., 2006. Simulation studies on the effects of clustering. XI*th* International Conference of Veterinary Epidemiology and Economics, Cairns, Australia, August 2006.

[18] Evans, B. A., Feng, Z., Peterson, A. V., 2001. A comparison of generalized linear mixed model procedures with estimating equations for variance and covariance parameter estimation in longitudinal studies and group randomized trials. *Statistics in Medicine* **20**, 3353–3373.

[19] Fitzmaurice, G. M., 2003. Methods for handling dropouts in longitudinal clinical trials. *Statistica Neerlandica* **57**, 75-99.

[20] Fitzmaurice, G., Molenberghs, G., Lipsitz, S., 1995. Regression Models for Longitudinal Binary Responses with Informative Drop-Outs. *Journal of the Royal Statistical Society, Series B* **57**, 691–704.

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

[22] Gail, M. H., Mark, S. D., Carroll, R. J., Green, S. B., Pee, D., 1998. On design consideration and randomization-based inference for community intervention trials. *Statistics in Medicine* **15**, 1069–1092.

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

[24] Glantz, S. A., 2005 *Primer of Biostatistics.* McGraw-Hill Professional, United States.

[25] Goldstein, H., 1991. Nonlinear multilevel models with an application to discrete response data. *Biometrika* **78**, 45–51.

[26] Goldstein, H., 1995. *Multilevel Statistical Models.* Institute of Eduacation. Multilevel project. London.

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

[28] Goldstein, H., Browne, W. J., Rasbash, J., 2002. Partitioning variation in multilevel models. *Understanding Statistics* **1**, 223–231.

[29] Gilks, W. R., Richardson, S., Spiegelhalter, D. J., 1996. *Markov Chain Monte Carlo in Practice.* Chapman & Hall, New York.

[30] Green, M. J., Burton, P. R., Green, L. E., Schukken, Y. H., Bradley, A. J., Peeler, E. J., Medley, G. F., 2004. The use of Markov Chain Monte Carlo for analysis of correlated binary data: patterns of somatic cells in milk and the risk of clinical mastitis in dairy cows. *Preventive Veterinary Medicine* **64**, 157–174.

[31] Gregoire, T. G., Brillinger, D. R., Diggle, P. J., Russek-Cohen E., Warren W. G. , Wolfinger, R. D., Neuhaus, J. M., Segal, M. R., 1997. *Modelling Longitudinal and Spatially Correlated Data: Methods, Applications, and Future Directions.* Chapter: An assessment of approximate maximum likelihood estimators in generalized linear mixed models. Springer.

[32] Griswold, M. E., Zeger, S. L., 2004. On marginalized multilevel models and their computation, Johns Hopkins University, Dept. of Biostatistics, Working Papers No. **99**.

[33] Haley, D. C., 1952. Estimation of the dosage Mortality Relationship When the Dose is Subject to Error. (Technical Report No. 15) CA: Stanford University, Applied Mathematics and Statistics Labs.

[34] Heagerty, P. J., Zeger, S. L., 2000. Marginalized multilevel models and likelihood inference. *Statistical Science* **15**, 1–19.

[35] Hamilton, D. C., 1992. Analysis of Fish Behaviour Data. *Canadian Journal of Statistics* **20**, 228–233.

[36] Hardin, J. W., Hilbe, J. M., 2003. *Generalized estimating equations.* Chapman & Hall/CRC, Boca Raton.

[37] Heyting, A., Tolboom, J. T., Essers, J. G., 1992. Statistical Handling of Drop-Outs in Longitudinal Clinical Trials. *Statistics in Medicine* **11**, 2043–2061.

[38] Hogan, J. W., Roy, J., Korkontzelou, C., 2004. Biostatistics tutorial: Handling dropout in longitudinal data. *Statistics in Medicine* **23**, 1455-1497.

[39] Hu, F. B., Goldberg, J., Hedeker, D., Flay, B. R., Pentz, M. A., 1998. Comparison of Population-Averaged and Subject-Specific Ap-

proaches for Analyzing Repeated Binary Outcomes. *American Journal of Epidemiology* **147**, 694–703.

[40] Kenward, M. G., Goetghebeur, J. T. Molenberghs, G., 2001. Sensitivity analysis for incomplete categorical data. *Statistical Modelling* **1**, 31–48.

[41] Kimber, W., 2007. Comparison of Phenobarbital and Potassium Bromide Monotherapies in the Treatment of Canine Epilepsy. MSc Thesis, Department of Biomedical Sciences, Atlantic Veterinary College, Charlottetown, Canada.

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

[43] Kotz, J., 1970. *Distributions in Statistics: Continuous Univuriate Distributions.* Wiley, New York.

[44] Kuchibhatla, M., Fillenbaum, G., 2003. Comparison of methods for analyzing longitudinal outcomes: cognitive status as an example. *Aging & mental Health* **7**, 462–468.

[45] Laird, N. M., 1988. Missing data in longitudinal studies. *Statistics in Medicine* **7**, 305–315.

[46] Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R., 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.

[47] Lesaffre, E., Spiessens, B., 2001. On the effect of the number of quadrature points in a logistic random-effects model: an example. *Journal of the Royal Statistical Society, Series C* **50**, 325–335.

[48] Liang, K. Y., Zeger, S. L., 1986, Longitudinal Data-Analysis Using Generalized Linear-Models. *Biometrika* **73**, 13–22.

[49] Lindsey, J. K., Lambert, P., 1998. On the appropriateness of marginal models for repeated measurements in clinical trials. *Statistics in Medicine* **17**, 447–469.

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

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

[52] Little, R. J. A., Rubin D. B., 2002. *Statistical Analysis With Missing Data.* Wiley-Interscience, Hoboken, NJ.

[53] Liu, Q., Pierce, D. A., 1994. A Note on Gauss-Hermite Quadrature. *Biometrika*, **81**, 624–629.

[54] Mancl, L., Leroux, B., 1996. Efficiency of regression estimates for clustered data. *Biometrics* **52**, 500–511.

[55] McCullagh, P., Nelder., J. A., 1989. *Generalized Linear Models.* Chapman & Hall, London.

[56] McCulloch, C. E., Searle, S. R., 2001. *Generalized Linear and Mixed Models.* Wiley, New York.

[57] McDermott, J. J., Schukken, Y. H., Shoukri, M. M., 1994. Methods for analysing data collected from clusters of animals. *Preventive Veterinary Medicine* **18**, 175–192.

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

[59] Molenberghs, G., Verbeke, G., 2005. *Models for Discrete Longitudinal Data.* Springer, New York.

[60] Neuhaus, J. M., 1992. Statistical methods for longitudinal and clustered design with binary responses. *Statistical Methods in Medical Research* **1**, 249–273.

[61] Pepe, M. S., Anderson, G. L., 1994. A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics: Simulation and Computation* **23**, 939–951.

[62] Pinheiro, J. C., Bates, D. M., 1995. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics* **4**, 12–35.

[63] Pinheiro, J. C., Bates, D. M., 2000. *Mixed effects models in S and S-Plus.* Springer-Verlag, New-York.

[64] Prentice, R. L., 1988. Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44**, 1033–1048.

[65] Preisser, J. S., Arcury, T. A., Quandt, S. A., 2003. Detecting patterns of occupational illness clustering with alternating logistic regressions applied to longitudinal data. *American Journal of Epidemiology* **158**, 495–501.

[66] Olde Riekerink, R. G. M., Barkema, H. W., Stryhn, H., 2007. The effect of season on somatic cell count and the incidence of clinical mastitis. *Journal of Dairy Science* **90**, 1704–1715.

[67] Rabe-Hesketh, S., Skrondal, A., 2008. *Multilevel and Longitudinal Modeling using Stata,* 2nd ed. Stata Press.

[68] Rabe-Hasketh, S., Skrondal, A., Pickles, A., 2002. Reliable estimation of generalised linear mixed models using adaptive quadrature. *The Stata Journal* **2**, 1–21.

[69] Rasmussen, M. D., Bjerring, M., Justesen, P., Jepsen, L., 2002. Milk quality on Danish farms with automatic milking systems. *Journal of Dairy Science* **85**, 2869–2878.

[70] 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.

[71] Rodríguez, 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.

[72] Sanchez, J., Dohoo, I. R., Nødtvedt, A., Keefe, G. P., Markham, F., Leslie, K., DesCôteaux, L., Campbell, J., 2002. A longitudinal study of gastrointestinal parasites in Canadian dairy farms. The value of an indirect *Ostertagia ostertagi* ELISA as a monitoring tool. *Veterinary parasitology* **107**, 209–226.

[73] SAS Institute Inc. 2004. *SAS/SAT 9.1 User's Guide.* Cary, NC: SAS Institute Inc.

[74] Schildcrout, J. S., Heagerty, P. J., 2005. Regression analysis of longitudinal binary data with time-dependent environmental covariates: bias and efficiency. *Biostatistics* **6**, 633–652.

[75] Schukken, Y. H., Grohn, Y. T., McDermott B., McDermott J. J., 2003. Analysis of correlated discrete observations: background, examples and solutions. *Preventive Veterinary Medicine* **59**, 223–40.

[76] Searle, S. R., Casella, G., McCulloch, C. E., 1992. *Variance Components.* Wiley, New York.

[77] Snijders, T. A. B., Bosker, R. J., 1999. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modelling.* Sage Publishers, London.

[78] Speer, C. A., Scott, M. Cathy, Bannantine, John, P., Waters, W. Ray, Mori, Yasuyuki, Whitlock, Robert H., Eda, Shigetoshi., 2006. A novel enzyme-linked immunosorbent assay for diagnosis of mycobacterium avium subsp. paratuberculosis Infections (Johne's Disease) in Cattle. *Clinical And Vaccine Immunology* **13**, 535–540.

[79] Sutradhar, B. C., Das, K., 1999. On the efficiency of regression estimators in generalized linear models for longitudinal data. *Biometrika* **86**, 459–465.

[80] 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.

[81] Veissier, I., Boissy, A., dePassillé, A. M., Rushen, J., van Reenen, C. G., Roussel, S., Andanson, S., Pradel, P., 2001. Calves' responses to repeated social regrouping and relocation. *Journal of Animal Science* **79**, 2580–2593.

[82] Virginie, R., Kiros, B., Duncan, C. T., 2005. A three-level model for binary time-series data: the effects of air pollution on school absences in the Southern California Children's Health Study. *Statistics in Medicine* **24**, 1103–1115.

[83] Wang, Y. G., Carey, V., 2003. Working correlation structure misspecification, estimation and covariate design: Implication for generalised estimating equation performance. *Biometrika* **90**, 29–41.

[84] Wolfinger, R., O'Connell, M., 1993. Generalized linear mixed models: a pseudo-likelihood approach. *Communications in Statistics: Simulation and Computation* **48**, 233–243.

[85] Yang, M., Goldstein, H., Heath, A., 2000. Multilevel models for repeated binary outcomes: attitudes and voting over the electoral cycle. *Journal of the Royal Statistical Society, Series A* **163**, 49–62.

[86] Zeger, S. L., Liang, K. Y., Albert, P. S., 1988. Models for longitudinal data - a generalized estimating equation approach. *Biometrics* **44**, 1049–1060.

[87] Ziegler, A., Kastner, C., Blettner, M., 1998. The generalized estimating equations: an annotated bibliography. *Biometrical Journal* **40**, 115–139.

# A simulation study to assess statistical methods for binary repeated measures data

## 2.1 Abstract

Binary repeated measures data are commonly encountered in both experimental and observational veterinary studies. Among the wide range of statistical methods and software applicable to such data, one major distinction is between marginal and random effects procedures. The objective of the study was to review and assess the performance of marginal and random effects estimation procedures for the analysis of binary repeated measures data. Two simulation studies were carried out, using relatively small, balanced, two-level (time within subjects) datasets. The first study was based on data generated from a marginal model with first order autocorrelation, the second on a random effects model with autocorrelated random effects within subjects. Three versions of the models were considered in which a dichotomous treatment was modelled additively, either between or within subjects, or modelled by a time interaction. Among the studied statistical procedures were: Generalized Estimating Equations (GEE), Marginal Quasi Likelihood, Likelihood based on numerical integration, Penalized Quasi Likelihood, Restricted Pseudo Likelihood and Likelihood based approximation by Bayesian Markov Chain Monte Carlo. Results for the marginal model data showed autoregressive GEE to be highly efficient when treatment was within subjects, even with strongly correlated responses. For treatment between subjects, random effects methods also performed well in some situations; however, a small number of subjects with short time series proved a challenge for both marginal and random effects methods. Results for the random effects model data showed bias in estimates from random effects methods while the marginal model produced estimates close to the marginal parameters.

## 2.2 Introduction

Repeated measures studies refer to data with multiple records over time on the same subject (e.g., animal or farm) with the objective of making inference about the expected value of outcomes, in terms of treatment effects and how such effects change over time. This type of study design, also referred to as longitudinal, has the advantage over a cross-sectional design that changes over time in treatment effects or in individuals can be estimated [10, Chapter 1]. The design also has a potential for substantial gains in efficiency.

Binary repeated measures data are encountered across a wide range of applications in veterinary science and veterinary epidemiology. The most evident examples of two-level data are records of presence or absence of disease conditions over time. Disease conditions may be detected clinically (e.g., mastitis) or by a test such as bacterial culture [34], faecal egg counts [1] or antibody determination for parasites [41]. Other examples are success of fertilization (e.g., in repeated reproduction cycles [2]), occurrence of certain behaviours in animal welfare studies [18, 48], or of treatment side effects in clinical trials (e.g., treatments for diabetes in dogs [23]). If the binary outcome is created by thresholding a quantitative outcome at a predefined cut-off value (e.g., ELISA for the diagnosis of Johne's disease; [45]) a substantial loss of information is implied but the dichotomous outcome may be of greater interest than the quantitative measurement. Another range of applications occur in the context of farm-level monitoring of product quality (e.g., milk [38]).

Binary records made on the same subject (or unit) over time are likely to be correlated [30, 43] or "clustered" [11, Chapters: 20-21]. A within-subject dependence violates the basic assumption of logistic regression that observations are independent, and may, if not accounted for, lead to biases in parameter estimates and standard errors ([10, Chapter 7] and [12]).

Several procedures (models) have been proposed for the analysis of binary repeated measures data, and they are usually classified into different models: marginal (population-averaged), random effects (subject-specific), and transition models ([33] and [10, Chapters 7-10]).

In marginal, random effects, and transitional models the treatment effects have different interpretations. Generally speaking, the choice of model should be guided by the data structure, the available information as well as the scientific questions of interest. The inferential goal of a marginal model is the marginal probability (averaged across the population of subjects), while for random effects models it is the probability conditional on the unobserved (subject) random effects. In transitional models, the inferential goal is the probability conditional on the previous response, i.e. the (transition) probability of moving from one binary state to the next state. Treatment effects refer to the impact of a treatment on these probabilities. Apart from an approximate conversion formula from random effects to marginal estimates (discussed below) no simple analytical links exist between the treatment estimates of the three models. In some situations, the question of interest largely determines the preferable model, for example if the interest is in transition probabilities and effects. If the factor of primary interest represents an inherent trait of the subjects, a subject-specific interpretation makes little sense [11, Chapter 22]. In practice, the choice between a marginal and a random effects model is often open to additional considerations such as software accessibility and statistical efficiency. Therefore, and by the fundamentally different interpretation of transition effects already noted, this study is focused on the choice between marginal and random effects models.

Despite the large body of literature on binary repeated measures data, the applied researcher may find little specific guidance on the choice of method for the data at hand (see however, [28]). Analysis of a single dataset by multiple procedures (e.g., [22]) does not necessarily provide much insight into which procedures provide the right answers, and does not cover all aspects of statistical inference. Statistical assessments of marginal and random effects procedures for clustered binary data are abundant (e.g., more recently [21]), though often without addressing all issues related to the repeated measures. One study for repeated measures focused entirely on variance and correlation parameters [13]. The assessments are usually based on statistical simulation, whereby artificial datasets are generated according to a statistical model with fixed and known parameters (true model). The parameter estimates from the analyses of simulated datasets by different statistical procedures are then compared to the known (true) parameters. This approach depends critically on the relevance of the selected true model. In the present context, the true model should reflect the longitudinal character of the data by allowing for autocorrelation, i.e. the dependence being stronger between observations on the same subject obtained close in time than distant in time. Moreover, Stryhn *et al.* [46] suggest that its data structure might be matched to the data at hand as closely as possible. Longitudinal data structures range from balanced two-level structures (e.g., randomized clinical trials with no structural dependence between subjects) to unbalanced, incomplete multi-level structures (e.g., observational records of farm animals). The focus here will be on the former, simpler structure while exploring the effect of other design parameters such as the length of the time series, the number of replicate subjects and the level at which the treatments are applied (between versus within subjects).

The objective of this study is to give a statistical assessment of marginal and random effects procedures, in terms of properties such as unbiasedness and efficiency, in a two-level balanced longitudinal design. The assessment includes a range of different design parameters as well as true model assumptions of either marginal or random effects type. In order to realistically reflect the choice an applied researcher faces when it comes to data analysis, only procedures implemented in broadly accessible statistical software are included. The goal of the assessment is to establish some practical guidelines for the choice of statistical procedure for the analysis of balanced, binary repeated measures data.

## 2.3 Statistical models and estimation procedures

Consider binary records (e.g., presence or absence of bacteria in monthly milk samples) $y_{ij}$ on each of $n$ subjects $(i = 1, \ldots, n)$ at $t$ time points $(j = 1, \ldots, t)$, as well as a set $x_1, \ldots, x_p$ of explanatory variables recorded for each subject at every time point.

### 2.3.1 Marginal or population-averaged (PA) model

A marginal logistic regression model takes the following form:

$$\text{logit}(\mu_{ij}) = \beta_0^{PA} + \beta_1^{PA} x_{1ij} + \ldots + \beta_p^{PA} x_{pij}, \tag{2.1}$$

where, $\mu_{ij} = \text{E}(y_{ij}) = \Pr(y_{ij} = 1)$. Thus, the marginal expectation (or probability of an "event") is modelled as a function of the explanatory variables through the logit link function. Furthermore, the marginal variance is related to the marginal expectation by the equation $\text{Var}(y_{ij}) = \phi\mu_{ij}(1 - \mu_{ij})$, where $\phi$ is a scale parameter, and subjects are assumed independent. Hereafter, $\beta^{PA}$ refers to a marginal, or population-averaged [52] regression parameter.

#### 2.3.1.1 Marginal model estimation procedures

The most commonly used marginal estimation procedures, generalized estimating equations (GEE) and alternating logistic regression (ALR), are often referred to as semi-parametric because they do not make assumptions about the specific form of a dependence between observations on the same subject, i.e. the within-subject correlation structure. Both GEE and ALR estimation yield estimates for $\beta^{PA}$ that are asymptotically unbiased and can be nearly efficient relative to the maximum-likelihood estimate in a fully and correctly specified model [10].

The GEE approach to analysis of longitudinal data by generalized linear models was introduced in a series of papers [27, 52].

The "estimating equation" refers to a (multidimensional) equation whose solution determines the parameter estimates, usually in a stepwise (iterative) manner where an updated equation is solved in each step and the process terminates when the solutions no longer change ("convergence"). The GEE procedure involves a user-specified "working" correlation matrix to approximate the true within-subjects correlation structure. Most software implementations offer a range of correlation structures, including independent ($\rho(j,j') = 0$), exchangeable ($\rho(j,j') = \rho$), and (first order) autoregressive ($\rho(j,j') = \rho^{j-j'}$), where $\rho(j,j')$ is the correlation between observations at times $j$ and $j'$. When using a robust variance estimation, the statistical properties for the estimates of $\beta^{PA}$ hold even for a misspecified working correlation structure; this is often referred to as a robustness property of the GEE procedure. However, a substantial loss of efficiency due to misspecification of the working correlation structure may occur as has been shown in studies involving different data structures [29, 53, 49, 42].

The ALR procedure uses the same estimating equation for $\beta^{PA}$ as GEE, but differs from GEE by modelling the association among responses in terms of pairwise odds ratios, and is numerically more efficient as the cluster size gets large [7].

The GEE procedure is available in most major statistical packages (e.g., SAS, S-Plus/R and Stata), with only slight differences in their implementation, and ALR is available in the former two packages.

### 2.3.2 Random effects or subject-specific (SS) model

The simplest random effects model, often termed a random intercept model, takes the following form:

$$\operatorname{logit}(\Pr(y_{ij} = 1|u_i)) = \beta_0^{SS} + \beta_1^{SS} x_{1ij} + \ldots + \beta_p^{SS} x_{pij} + u_i, \qquad (2.2)$$

where $u_1, \ldots, u_n$ are independent random variables with the same distribution. The most commonly assumed distribution is the Gaussian (normal), say $u_i \sim \mathrm{N}(0, \sigma^2)$ where $\sigma^2$ represents the heterogeneity (variance) between subjects. Model (2.2) is for the conditional probability of an "event" given the random effect $u_i$ of the $i$th subject, rather than the marginal probability in model (2.1). Hereafter, $\beta^{SS}$ refers to a random effects, or subject-specific [52], regression parameter.

The relation between random effects and marginal estimates has been discussed and described([52, 32]; see also the summary by Diggle *et al.* [10, Chapter 7]. Without any distributional assumptions on the random effects it holds that the marginal regression parameters are attenuated or diluted (towards zero) relative to the random effects parameters, unless the subject variance $\sigma^2$ is zero. For normally distributed random effects, the following approximation formula holds:

$$\beta^{PA} \approx (c^2\sigma^2 + 1)^{-1/2}\beta^{SS}, \quad \text{where } c = 16\sqrt{3}/(15\pi) = 0.588. \qquad (2.3)$$

#### 2.3.2.1 Random intercept model estimation procedures

Random effects models for binary outcomes do not have a closed form of the full log likelihood function. As the likelihood involves an integral over the random effect distribution, numerical integration by Gauss-Hermite quadrature is a possibility (for normally distributed random effects). The preferable form of the integration is adaptive quadrature, whereby the quadrature points are successively adapted to the shape of the log likelihood function. Statistical estimation procedures based on numerical integration via adaptive quadrature to maximize the log likelihood (ML), produce reliable estimates of the regression parameters [37]. One should be cautioned that "even with adaptive Gaussian quadrature and with relatively simple models, convergence to a global maximum can be difficult to obtain" [26].

Before numerical integration became computationally feasible in practice, several approximation algorithms aimed at producing estimates close to the global ML estimate without actually computing the likelihood function were developed (see [4], for a recent review). These algorithms carry a number of different names and acronyms typically involving "weighted least squares" and "quasi"- or "pseudo-likelihood". The algorithms iteratively employ mixed linear model estimation to an "adjusted" variate obtained by Taylor approximation of the outcome around its current estimated mean, until convergence. 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., [39, 40]). A "second order" PQL procedure eliminates some of the bias [15]. A marginal version of the algorithms (e.g., termed MQL) yields parameter estimates with a marginal interpretation [5], although computed under random effects model assumptions.

ML estimation by numerical integration for generalized linear mixed models has become available in several statistical packages in recent years, the most flexible implementation being the *gllamm* package for Stata [36]. Weighted least squares approximation algorithms are available in most statistical software packages (e.g., SAS, S-Plus/R and Stata) as well as in special-purpose multilevel software (e.g., MLwiN (including the 2nd order PQL option) and HLM).

#### 2.3.2.2 Bayesian modeling and estimation procedures

The focus here is on using Markov chain Monte Carlo (MCMC) techniques within a Bayesian framework as an estimation algorithm for the frequentist model (2.2), rather than exploring genuine Bayesian models with informative prior distributions. The MCMC approach avoids computation of the full likelihood function, and has been shown to perform well across a range of settings [6]. The essential statistical software for Bayesian analysis is WinBugs; in addition, a range of multi-level models can be fitted in MLwiN.

#### 2.3.2.3 Random effects repeated measures models and estimation procedures

A serious objection against model (2.2) for longitudinal data is that it implicitly assumes an exchangeable correlation structure whereby any two observations on the same subject are correlated to the same degree. As the variances and correlations in generalized linear (mixed) models depend on the fixed effects, this statement is only strictly true if the fixed effects include no time-dependent predictors. Intuitively, one would expect the correlation between two observations to decrease with their distance in time. Several approaches have been suggested to allow for non-exchangeable correlation structures (e.g., [3, 51]; and [31, Chapter 22]), but to our knowledge the only one in widespread use and implemented in standard statistical software is the restricted pseudo likelihood approach (REPL; SAS: `proc glimmix`, R: `glmmPQL` library). It is based on a similar weighted least squares algorithm approximation algorithm as described above, but allows for both random effects and error correlation structure in the linear mixed model estimation of the "adjusted variate" [50]. The correlation structures are for the binary repeated measures; modelling by correlation structure alone therefore yields PA estimates [31, Chapter 9]). Adding random effects effectively yields a random effects model with serial correlation [31, Chapter 22].

Another idea is to replace the single random effect $u_i$ for subject $i$ by a series $(u_{i1}, \ldots, u_{it})$ of $\mathrm{N}(0, \sigma^2)$ distributed, autocorrelated random effects (as in the marginal model, $\rho(j, j') = \rho^{j-j'}$). The extension of model (3.2) then takes the form,

$$\operatorname{logit}(\Pr(y_{ij} = 1 | u_{i1}, \ldots, u_{it})) = \beta_0^{SS} + \beta_1^{SS} x_{1ij} + \ldots + \beta_p^{SS} x_{pij} + u_{ij}. \quad (2.4)$$

If $\rho = 1$, model (2.4) reduces to the random intercept model (2.2). In our view, model (2.4) forms a better basis for random effects modelling of repeated measures data because of its ability to incorporate autocorrelation [10]. In principle, model (2.4) can be set up and estimated in a Bayesian framework using MCMC methods ([9, Chapter 5], [10, Chapter 11]), e.g., in WinBUGS software. In our experience, however, it is a non-trivial task to achieve acceptable trajectories of the resulting Markov chains.

A further refinement of this idea, and an amalgamation of marginal and random effects procedures, marginalized models employ the marginal model (2.1) for the regression coefficients and the random effects model (2.4) for the correlation structure. In the latter model, the fixed part is replaced by the equivalent of the marginal model fixed part for the conditional probability [20]. Marginalized models may be programmed using flexible statistical optimization tools (such as the `nlmixed` procedure in SAS; [17]), but to our knowledge these models are not yet available in standard statistical software, or as an add-on package.

## 2.4 Simulation studies

### 2.4.1 Models for simulated data

Two simulation studies were carried out using relatively small, balanced, two-level (time within subjects) datasets. The first study was based on the marginal logistic regression model (2.1) with a first order autocorrelation between the binary outcomes. The second study used the random effects model (2.4) with autocorrelated subject random effects. The fixed part of the models included a dichotomous treatment and a linear effect of time. The treatment was "applied" either to subjects (between-subjects (BS) design) or to two periods within each subject (WS) in a balanced cross-over type design. Three versions of the fixed part structure were studied: additive time and treatment effects in BS and WS designs, as well as an interaction model for the BS design.

All studies and designs were furthermore assessed in different settings intended to reflect a range of experimental data encountered in practice. The number of subjects was set to either small or large ($n = 20$ and 100, respectively). Datasets with substantially less than 20 subjects per treatment would usually not be analyzed by random effects methods, and therefore fall beyond the scope of this study. The length of longitudinal series on each subject was short, medium or long ($t = 4$, 8 and 16, respectively). The marginal autocorrelation between adjacent observations was high, moderate, or low ($\rho = 0.7$, 0.5 and 0.2, respectively). In the random effects model, the between-subjects standard deviation was set at $\sigma = 1$, and the same correlation values as above were used for the autocorrelated subject random effects. We also included the special case $\rho = 1$ corresponding to a random intercept model. 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 $\pi^2/3$ [44, Chapter 14] yields an intra-class correlation of $\sigma^2/(\sigma^2 + \pi^2/3) = 0.23$ and a first-order correlation of $\rho\sigma^2/(\sigma^2 + \pi^2/3)$, and the values 0.16, 0.12 and 0.05 for $\rho = 0.7, 0.5, 0.2$, respectively.

Simulated data from the marginal model were generated by the `bindata` package in R software [25]. The algorithm generates binary random variables with a given correlation structure by converting multivariate random variables into binary variables. Correlations among binary data are constrained by the (marginal) probabilities [35], which, in a logistic regression, depend on the predictors. Hence, the fixed effects parameters were chosen to avoid extreme probabilities that would make the desired correlation structures infeasible. The fixed parameters were set at: $\beta_0 = -0.5$, $\beta(\text{treatment}) = 0.35$, $\beta(\text{time}) = 0.10$, $\beta(\text{interaction}) = -0.15$. The autocorrelated random effects of each subject were obtained by multiplying a vector of $t$ independent variables by the upper triangular factor of the Cholesky decomposition of the correlation matrix as described by Congdon [9]. Generation of the binary outcomes then followed the usual scheme for random effects logistic regression models [46].

### 2.4.2 Software and settings for estimation procedures

The GEE estimation procedures used the implementation in R version 2.1.0 software (`gee` version 4.13.10) with different working correlation structures: independence, autoregressive ($\hat{\beta}^{PA}_{\text{AR}}$), exchangeable, and autoregressive with known (true) correlation ($\hat{\beta}^{PA}_{\text{F}}$). The ALR estimation procedure ($\hat{\beta}^{PA}_{\text{ALR}}$) was carried out in R version 1.9 software (`alr` 4.2 package) and used an exchangeable correlation structure. The random effects procedures used the first order MQL ($\hat{\beta}^{PA}_{\text{MQL}}$) and second order PQL ($\hat{\beta}^{SS}_{\text{PQL}}$) procedures implemented in the MLwiN software (version 2.02), the REPL procedure ($\hat{\beta}^{ss}_{\text{REPL}}$) in proc glimmix of SAS (version 9.1), as well as the adaptive quadrature algorithms for ML estimation ($\hat{\beta}^{ss}_{\text{ML}}$) implemented in Stata version 9 software (xtlogit and gllamm commands), and the non-adaptive quadrature glmmML (version 0.26) package for R software. The REPL procedure was set up with subject random effects and a first order autoregressive correlation structure (including also an additional overdispersion parameter); in addition, a marginal REPL procedure without subject random effects was included.

The Bayesian estimation procedures ($\hat{\beta}^{ss}_{\text{MCMC}}$) were implemented in WinBUGS version 1.4 called from the R software using the R2WinBUGS package [47]. Vague ("non-informative") prior distributions (i.e. $\text{N}(0, 10^6)$) were used for all fixed effects parameters. The recently recommended uniform distribution for inverse variances [24, 14] proved sensitive to trap messages, even after truncation of the distribution, so we reverted to the classical gamma distribution $(10^{-3}, 10^{-3})$ for the inverse between-subjects variance [6]. The Markov chains were run with 300 burn-in samples, and the subsequent estimates (posterior distribution medians) were based on 1000 samples. These burn-in and estimation sample sizes were arrived at after inspecting MCMC diagnostics for selected datasets; Browne *et al.* [6] used a somewhat larger burn-in period of 500 samples.

In order to reduce the computing time, the datasets generated by a model including an interaction term were analyzed only by a restricted set of estimation procedures: GEE with autoregressive correlation, ALR, ML by numerical integration as implemented in Stata, and MCMC.

### 2.4.3 Analysis of and performance of simulated data

For the analysis of the simulated marginal model data, the estimates and standard errors of random effects estimation procedures (except MQL) were converted to marginal parameter estimates by the formula (2.3), using the estimated between-subject variance. The bias-adjusted relative efficiency of an estimate $\hat{\beta}$ was computed by the formula,

$$\text{relative efficiency} = \frac{\text{Var}(\hat{\beta}_{true}) + (\text{E}(\hat{\beta}_{true}) - \beta_0)^2}{\text{Var}(\hat{\beta}) + (\text{E}(\hat{\beta}) - \beta_0)^2}, \qquad (2.5)$$

where $\beta_0$ refers to the true model parameter, $\hat{\beta}_{true}$ refers to an autoregressive GEE analysis with a correlation structure fixed at the true value, denoted by $\hat{\beta}^{PA}_{\text{F}}$ [49, 8]. Thus, the relative efficiency measures the variance around the true value, caused by either random variation or bias, relative to a “correct” estimation procedure. The means and variances in formula (2.5) were computed from the distribution of the corresponding estimate across the simulated datasets. Note that by the lack of a reference method for the autoregressive random effects model, no analogous relative efficiency could be computed.

The presence of statistically significant bias in the estimates was assessed by a $z$-test based on the true value and the standard deviation among simulations. The statistical significance of bias in the standard errors was assessed by comparing the mean standard error to a 95% confidence interval for the standard deviation based on the simulations. This simple procedure was considered acceptable because the statistical variation in the estimated standard deviation was generally much larger than that of the mean standard error. The confidence intervals were computed by the large-sample normal approximation based on the standard error; for the GEE procedures, the robust standard error was used. The coverage of 95% confidence intervals was computed as the proportion of simulated datasets for which the confidence interval (in the Bayesian analysis: the credibility interval) contained the true parameter.

If non-convergence or non-sensible estimates occurred for a certain method and dataset, the analysis was attempted by the same or a similar method in a different software implementation. For example, in some small data settings the autoregressive GEE procedures failed to produce a useful (an extreme value of) robust standard error. By the close agreement between model-based and robust standard errors in other settings, the model-based standard error was used in such instances. Also, the ALR estimate was in some cases obtained from SAS instead of R, and different optimization techniques were tried in for the REPL procedure in SAS if the default quasi-Newton method failed. If the problems persisted across different implementations, the corresponding dataset was omitted.

## 2.5 Results of performance analysis

Generally, results are shown only for the treatment parameter. Means of estimates and of the associated standard errors, as well as standard deviations of estimates among the simulations are shown in tables (Tables 2.1–2.4). A summary of performance measures on the bias, confidence interval coverage and relative efficiency are shown graphically (Figures 2.1–2.6). In the interest of clarity and space, all results of the interaction models were excluded from the presentation. Furthermore, no results are shown for GEE estimation with independence and exchangeable correlation structures because throughout they were very close to those of ALR estimation. We focus here is on ALR estimation because to our knowledge its performance and robustness properties have not been reported. Also, the marginal REPL analysis has been omitted from the results, because its close agreement with GEE procedures is well-documented [31, Chapter 9]. For ML estimation by numerical integration, only the results from the `gllamm` implementation are shown because of its greater flexibility than the `xtlogit` command in Stata and the `glmmML` package in R. Additional tables (A.1–A.8) of results are reprinted in Appendix A.

### 2.5.1 Marginal model data

The estimates of the two GEE procedures agreed closely (Tables 2.1 and 2.2). The robust and model-based standard errors (not shown) of the autoregressive GEE procedures were generally close, and in agreement with the standard deviation across the simulations. However, in small datasets ($n = 20$) the agreement was best for the model-based standard errors whereas the robust standard errors were up to 10% lower. The efficiency for $\hat{\beta}^{PA}_{\mathrm{AR}}$ relative to $\hat{\beta}^{PA}_{\mathrm{F}}$ was close to 100% in all settings (Figures 2.2 and 2.4). The estimates of ALR and MQL procedures agreed closely, and were close to the GEE estimates except for one setting (WS; $(n, t, \rho) = (20, 16, 0.7)$).

The REPL estimates were close to the GEE estimates in most settings, with scattered deviations in some of the smallest datasets ($n = 20$ and $t \leq 8$). The estimates of the three other random effects procedures were close in many settings (in particular the largest datasets) and then differed substantially from the marginal estimates; however, overall there was less agreement among these random effects procedures than the marginal procedures. Compared to the other random effects procedures, REPL produced markedly lower estimates of the between-subjects variances, which in turn affected the scaling of random effects to a marginal estimates and overall lead to a performance of REPL akin to a marginal procedure.

The variability of estimates, as expressed by the standard deviations, decreased with increasing size of the dataset (increasing $n$ or $t$), and also decreased with decreasing correlation in BS designs. The standard deviations were higher in BS than WS datasets with the same settings.

#### 2.5.1.1 Treatment between subjects (BS)

The estimates of marginal estimation procedures were generally unbiased, except for a general upwards bias in the smallest datasets ($(n, t) = (20, 4)$) with moderate to large correlation (Table 2.1). The coverage of confidence intervals was close to nominal for $n = 100$ but underestimated (lowest value 91%) in several small dataset settings (Figure 2.1). The relative efficiency of the ALR and MQL procedures ranged between 0.9 and 1.0 (Figure 2.2).

The REPL estimates were unbiased even in the smallest datasets with large correlation; this lead to a relative efficiency above 1. The CI coverage was close to nominal (range 94–97%), and the relative efficiency was never below 1. Analysis allowing for extra-binomial dispersion showed a minor underdispersion with values ranging down to 0.8 (results not shown)

The ML and MCMC random effects estimation procedures showed more instances of an upwards bias than the marginal procedures, and the bias tended to increase with increasing $\rho$. The PQL estimates were on the average close to the true value, and only some biased settings were noted (Table 2.1). The CI coverage for the PQL procedure was close to nominal (range 93–96%), and the relative efficiency ranged between 0.9 and 1.0, except for the smallest dataset where PQL performed better than the reference procedure. However, the comparison in this setting was obscured by the fact that due to convergence problems for the second order PQL procedure, a first order procedure was often used. Both ML and MCMC procedures had relative efficiencies down to 0.78 for highly correlated data, and more variable CI coverages; in particular, MCMC showed less than nominal coverage (lowest 91%) in most settings. It is notable that for data with the strongest autocorrelation, the between-subjects variance estimates were more extreme with ML and MCMC procedures than with PQL; the latter values ranged up to $\hat{\sigma} > 5$ in the smallest datasets $((n, t) = (20, 4))$.

#### 2.5.1.2 Treatment within subjects (WS)

The marginal estimation procedures only showed an upward bias for the longest series with small number of subjects ($(n, t) = (20, 16)$; Table 2.2). The CI coverage for the GEE and ALR procedures was close to nominal or moderately below (ranging down to 91.5%; Figure 2.3). The MQL procedure suffered from substantial undercoverage (down to 74%) unless the series was short ($n = 4$) or the correlation was low ($\rho = 0.2$). This was a result of severely underestimated standard errors (Table 2.2). The relative efficiency of ALR and MQL procedures dropped down to close to 0.5 for the longer series with high correlation (Figure 2.4). For the autoregressive GEE procedures, the standard deviations among estimates seemed to peak at an intermediate true correlation $\rho$, except for the shortest series where they decreased with increasing $\rho$.

The REPL procedure performed similarly to the GEE procedures, except for a single uncharacteristic downward bias for $(n, t, \rho) = (20, 4, 0.7)$. The other random effects procedures gave fairly similar average estimates and standard deviations, and were all subject to upwards bias in most settings. The relative efficiency was as low as for the ALR procedure, ranging down to 0.5 for the longer series with high correlation. The confidence intervals showed strong undercoverage (down to 66%) except for the shortest series ($t = 4$), again owing to at times grossly underestimated standard errors of the procedures. It may be noted that PQL tended to give larger estimates in comparison with ML and MCMC, as $\rho$ increased.

### 2.5.2 Random effects model data

The estimates of marginal estimation procedures (GEE, ALR and MQL) can be compared either to the true subject-specific parameter value (0.35) or the true marginal parameter value (0.302), obtained from the conversion formula (2.3) with the known between-subjects standard deviation $\sigma = 1$. The indicated significance for bias in Tables 2.3–2.4 and the coverage of confidence intervals in Figures 2.5–2.6 refer to the marginal parameter. This comparison is, however, theoretical and hypothetical from a practical point of view because $\sigma$ is not known and no estimate is provided for $\sigma$ from marginal estimation procedures (except MQL). Further discussion of the implications of choosing a marginal procedure is deferred to Section 2.6.4.

The estimates and standard deviations from marginal estimation procedures generally agreed closely (Tables 2.3–2.4), except for MQL in the $((n, t, \rho) = (20, 4, 0.2))$ setting. The random effects estimation procedures, except REPL, also agreed fairly well, and some differences may be due to convergence problems experienced for the ML, PQL and MQL procedures, especially for $\rho < 1$. Similarly with the marginal data, the REPL estimates were generally closer to the estimates of marginal than random effect estimation procedures.

For the random intercept model ($\rho = 1$), the random effects procedures showed no appreciable bias in the treatment effect except for the smallest dataset in the WS design (Tables 2.3–2.4). Across all settings and procedures, the average between-subject standard deviations were close to the true value (data not shown). The CI coverage was close to nominal (range 93–96%, Figures 2.5–2.6). The autoregressive random effects model ($\rho < 1$) generally showed a downward bias in treatment effects (except for the smallest dataset in both designs), and the mean estimates were in most cases closer to the marginal than the random effects parameter. The between-subject standard deviations were strongly underestimated, with values decreasing with both $t$ and $\rho$, and ranging from of 0.80 ($(t, \rho) = (4, 0.7)$) to 0.12 ($(t, \rho) = (16, 0.2)$). The confidence intervals showed strongest undercoverage, down to 87%, in those large datasets ($n = 100$) where the bias in the standard deviations was most pronounced.

Irrespective of $\rho$, the marginal estimation procedures (including GEE with exchangeable and independence correlation structures) gave estimates centered around the marginal parameter although a few settings showed a minor bias in either direction (Tables 2.3–2.4). The CI coverages were generally above 90% in the larger datasets, owing to an underestimated standard error, but close to nominal in datasets with less information.

For $\rho = 1$, the estimated between-subjects standard deviation produced by MQL showed a clear downward bias (range 0.73–0.82).

Convergence problems were encountered with several procedures, most severely so in small datasets with low correlation. The quasi-likelihood procedures (MQL and PQL) were strongest affected, and in some cases the analysis could not be completed in 30-40% of the datasets (e.g $((n, t, \rho) = (100, 4, 0.2), (20, 16, 02), (20, 4, \leq 0.5)))$.

### 2.5.3 Interaction model in between subjects design

The parameters of interest were the difference between treatments at the average time point (treatment main effect when the time predictor is centered) and the change over time in treatment effect (treatment by time interaction). Convergence problems due to the complexity of the interaction model were encountered with most of the procedures; in particular in the smallest dataset $((n, t) = (20, 4))$, so it was excluded from the analysis. Although results are not shown, the findings are summarized briefly for both marginal and random effects data.

#### 2.5.3.1 Marginal model data

The qualitative statement can be made that estimates for the main and interaction effects showed behaviours similar to the BS and WS designs, respectively. Generally, marginal estimation procedures (GEE and ALR), and random effects procedures (ML and MCMC), gave estimates centered around the true value, except for $((n, t, \rho) = (20, 8, \leq 0.7)$. Random effects procedures showed more instances of deviations from the true value than the marginal procedures.

For the interaction parameter, the GEE and ALR estimation procedures showed no substantial bias and a CI coverage close to nominal even if ranging down to 92%. The mean standard error and standard deviation among simulations agreed closely for marginal estimation procedures, whereas for random effects procedures, the latter was always larger. The MQL and random effects procedures suffered from CI undercoverage (down to 73 and 65%, respectively) in the presence of high correlation. All procedures except the autoregressive GEE showed some loss in relative efficiency, ranging down to 0.7 for random effects procedures in the presence of high correlation.

### 2.5.3.2 Random effects model data

The estimates of marginal estimation procedures (GEE and ALR) can be compared either to the true subject-specific parameter value of interaction and treatment main effect ($-0.15$ and $0.35$) or the true marginal parameter value ($-$ 0.129 and 0.302) respectively, obtained from the conversion formula (2.3) with the known between-subjects standard deviation $\sigma = 1$. Findings refer to the marginal parameters.

The estimates, standard deviations and the mean standard errors from marginal estimation procedures (GEE and ALR) generally agreed closely. For both treatment main effect and interaction effect, and irrespective of $\rho$, the marginal estimation procedures gave estimates centered around the marginal parameter, although a few settings deviated in either direction (e.g., $((n, t, \rho) = (100, 16, \leq 0.5)$ and $(n, t, \rho) = (20, 8, \leq 0.7)))$. The CI coverages were generally above 91%.

The estimates, standard deviations and the mean standard errors from marginal estimation procedures (GEE and ALR) generally agreed closely. For both treatment main effect and interaction effect, and irrespective of $\rho$, the marginal estimation procedures gave estimates centered around the marginal parameter, although a few settings deviated in either direction. The CI coverages were generally above 91%.

For the random intercept model ($\rho = 1$), the random effects procedures showed estimates close to the true values in most settings. The CI coverages were generally close to nominal and down to 92% in some settings specially for MCMC procedure. Both random effects procedures (ML and MCMC) gave estimates for the between-subject standard deviations close to the true value (data not shown). As with the random effects model data, the autoregressive random effects model data ($\rho < 1$) generally showed similar patterns for the mean estimates and between-subject standard deviations.

### 2.5.4 Summary of performance

The performance of the estimation procedures in terms of bias, coverage of confidence intervals and efficiency (for marginal model data only) were summarized to yield Tables 2.5 and 2.6. For each procedure and effect type (BS or WS), the tendency across all data settings were assessed as either 0 (no bias, nominal coverage, 100% efficiency), $-$ (underestimation, undercoverage, <95% efficiency), or as + for converse findings. Where multiple patterns existed across data settings, additional symbols were given in parenthesis for findings present in more than two settings, with the more common patterns listed first.

Across all marginal data settings, the marginal estimation procedures predominantly performed well, the one exception being MQL for the WS design. REPL performed as a marginal procedure, and at times better than the reference GEE method. The other random effects procedures performed acceptably in the BS design, despite some loss in relative efficiency, but failed in many settings of the WS design on all performance parameters assessed.

For the random effects data settings, the marginal estimation procedures performed reasonably well, when compared to the true PA parameter, despite some tendency towards negative bias. As a random effects procedure, REPL was compared to the true SS parameter, which was generally underestimated irrespective of the true value of $\rho$. The other random effects procedures performed well in the compound symmetry setting ($\rho = 1$) but showed similar underestimation and undercoverage as REPL for $\rho < 1$.

## 2.6 Discussion

Although stated generally, the conclusions in the following are evidently confined to the range of procedures and settings covered by the study.

### 2.6.1 Marginal estimation procedures

The GEE estimation procedure with autoregressive working correlation matrix remained highly efficient across all marginal data settings, and the model-based and robust standard errors agreed closely. The procedure performed on par with other marginal procedures for the random effects data (with exchangeable, and non-autoregressive correlation structures). Although these results covered only a small range of non-autoregressive correlation structures, they are in our view supportive of the use of an autoregressive working correlation structure when using GEE procedures for longitudinal binary data [42].

Other correlation structures for GEE (exchangeable and independence) showed substantial loss in efficiency relative to the autoregressive structure in the marginal model data, in particular for the WS design with a moderate to large autocorrelation. The independence structure has been described as sufficient for datasets up to moderate size [19, Chapter 3], but the loss in efficiency was observed even in the small ($n = 20$) datasets. The estimates for ALR were very close to those of GEE with exchangeable correlation structure. If there is, within the marginal estimation framework, interest in a parameter quantifying the within-subjects dependence, the log odds-ratio of the ALR procedure is in our view preferable to values of the GEE working correlation matrix. Evidently the odds-ratio is a better and more commonly used measure of association between binary outcomes than the correlation [10]. Also, the ALR procedures provides a standard error of the estimated association so that a confidence interval can be constructed. However, the ALR procedure does not in its current implementations allow for repeated measures correlations such as autocorrelation.

The demonstrated downward bias in MQL estimates of the between-subjects variance is well-known (e.g., [39]). As a marginal procedure [5], MQL performed on par with other procedures involving the exchangeable correlation assumption (ALR, GEE), except for a strong CI undercoverage (similar to the random effects procedures) in the WS and interaction datasets. As the undercoverage is largely a result of underestimated standard errors, it is suggested to add robust ("sandwich") variance estimation to the MQL procedure; the usefulness of this suggestion remains however to be assessed in practice.

In the smallest BS datasets ($n = 20$), all marginal procedures experienced problems with some CI undercoverage and upwards bias in estimates (only $t \leq 8$). These findings are agreement with previous findings, e.g., summarized in a recommendation to apply GEE only "if the number of clusters is at least 30 for a cluster size of about 4 for low to moderate correlation" [53].

### 2.6.2 Random effects estimation procedures

All the random effects procedures (except REPL discussed below) performed well and had fairly similar estimation errors in the data generated from random intercept models ($\rho = 1$). Recently, Heo and Leon [21] concluded that the full likelihood approach "appears to be preferable for the analysis of clustered binary observations with underlying random effects models". On the other hand, Browne and Draper [6] reported the closest reproduction of true model values with MCMC procedures, a finding that could not be reproduced in the current study settings, in which, with one exception, the procedures showed no bias.

For marginal model data, the PQL procedure seemed less affected by the model misspecification than the ML and MCMC procedures, in particular for the BS design, and the MCMC procedure had lowest efficiency and CI coverage in several settings. These tendencies may be linked to the higher estimated between-subjects variances for ML and MCMC procedures. As the data were generated from a marginal model no true value exists for assessment of these estimates; however, one may speculate that the PQL estimates were biased towards zero as previously mentioned (e.g., [40]). Another possible explanation is a selection bias resulting from convergence problems for all random effects procedures, most severely for the PQL procedure.

For the autoregressive random effects model data ($\rho < 1$), all regression estimates of random effects procedures were similar but downward biased and close to the marginal estimates. This can be seen largely as a scaling effect caused by the underestimation of the random effect variances. The avoidance of such scaling problems by separating fixed and random effects estimates was one of the key ideas behind the development of marginalized models [20]. Finally, the agreement between the two Stata implementations of ML estimation based on adaptive quadrature (`xtlogit` and `gllamm`) suggested the use of the former, and faster, procedure for simple two-level models.

The inclusion of both random effects and a correlation structure in the REPL procedure makes it difficult to characterize the resulting approach in terms of the PA/SS dichotomy. This is due to the modelling of parts of the variance/correlation structure on different scales [31, Chapter 22]. For both marginal and random effects datasets, the estimated variance by REPL was substantially lower than by other random effects procedures. This could in part be due to the well-known attenuation of variance parameters by PQL in certain settings [13], but could also be due to the "competing" explanation of the variance/correlation structure on the two scales. In effect, REPL performed mostly as a marginal estimation procedure, and showed no promise for estimation of the variance and autoregressive parameter in the autoregressive random effects data. The performance was actually similar to that of a REPL approach without any random effects (results not shown) which was already noted to be similar to GEE, except for an increased sensitivity to convergence problems in datasets with a small number of subjects and long series, in accordance with the findings reported by Molenberghs and Verbeke [31, Chapter 14].

### 2.6.3 Issues related to statistical design

The precision of parameter estimates and performance of estimation procedures were generally better in within-subjects (WS) than between-subjects (BS) designs, in agreement with the general notion that the former, whenever logistically and biologically feasible, are the more powerful. The smallest ($n = 20$) marginal BS designs presented problems for all estimation procedures whereas performance in the corresponding WS designs was clearly better for the marginal estimation procedures. Also, the WS designs produced more pronounced differences in performance than the BS designs. Different impacts of autocorrelation on the precision of parameter estimates were noted in the BS and WS designs. In the former, precision decreased with increasing $\rho$ whereas in the latter precision seemed to be non-monotonic as a function of $\rho$ (see also the discussion by Diggle *et al.* [10]). At closer scrutiny, in some settings with low $\rho$ (both marginal and random effects data) the difference in precision between BS and WS estimates was fairly small. Thus, the implied gain in precision by a WS design may not always be the determining factor when planning an experimental study (when a large number of subjects are available).

In the presence of an interaction, qualitatively different behaviours for main effect and interaction parameters were observed. These behaviours agreed with the intuitive perception that the main effect corresponds to a fixed point in time and is estimated between subjects in a cross-sectional manner, whereas the interaction is a time-varying effect estimated within each subject just like the treatment in the WS design.

### 2.6.4 Marginal versus random effects estimation procedures

We conclude with a discussion of the implications of the findings for the choice of procedure, in particular the choice between marginal and random effects procedures. As was already stated in the introduction, the researcher should first and foremost be guided by the desired interpretation of effects. Diggle *et al.* [10, Chapter 7] argue that PA effects are of primary interest in clinical trials because "the average difference between control and treatment is most important, not the difference for any one individual". Lindsey and Lambert [28] warn that the population average may hide individual effects, and that "in extreme cases, a marginal analysis can show an average positive treatment effect when the effect would in fact be judged negative for each individual". We think that a good study design that pays attention to the randomization and the allocation process of subjects to different comparison groups should help to control any confounders and avoid such extreme situations.

If a PA interpretation is desired, the semi-parametric marginal estimation procedures have to their credit the robustness implicit in making no specific assumptions about random effects and correlation structure. As the random effects procedures under study here, excluding REPL, all make the conceptually unreasonable assumption that residual correlations are constant over time, the question for application of such random effects procedures is the sensitivity of the results to that assumption. For marginal model data, either of the WS or interaction design, the random effects procedures displayed severe deficiencies, in terms of both efficiency and CI coverage, which increased with the size of the dataset and the true autocorrelation. For the BS design, the random effects procedures showed a minor loss of efficiency, but for the small datasets also a CI coverage closer to nominal than marginal procedures. In the smallest dataset ($(n, t) = (20, 4)$), both marginal and random effects procedures lead to a marked upwards bias. For the random effects model and estimation of its corresponding marginal parameter, the marginal procedures performed similarly to the procedures for misspecified correlation structure, considering the low correlation between binary outcomes. Furthermore, in the presence of autocorrelation between random effects the random effects procedures were closer to the marginal parameter instead of the true, subject-specific value. In summary, the use of random effects procedures to estimate a marginal parameter is not recommended generally but may be acceptable in certain settings. In particular, in some datasets that were too small for the asymptotic properties of GEE procedures to guarantee approximately unbiased estimates and close to nominal CI coverage, the random effects procedures estimates had slightly better properties. However, we advise against random effects procedures if the effect of interest is time-varying and there is a strong decay in correlations, even for a series as short as 4 time points.

If an SS interpretation is desired, a marginal estimation procedure is of little use unless (unrealistically) the between-subjects variance is known. Nor are marginal procedures attractive in situations where a between-subjects variance is of genuine interest. The MQL procedure does provide a variance estimate, but because of its downward bias a back-conversion of PA to SS estimates using the conversion formula (2.3) does not yield an unbiased estimate of the SS parameter. One advantage of using random effects procedures is the ability to model and predict effects at the individual level. The properties of random effects procedures under exchangeable correlations has been extensively studied; however, the focus was on situations with decaying correlation over time. Goldstein *et al.* [16] described how underdispersion may arise as a result of unmodelled autocorrelation; the results for the marginal model did not show any substantial underdispersion. In the presence of autocorrelation, the random effects procedures failed to reproduce the subject-specific value, and for this situation we cannot point to any procedures among those covered in the study to obtain subject-specific estimates with acceptable performance. Thus, we advise to wait the development of marginalized models to see whether they could become the first choice in such situations. Until then the researcher's best option might be to try to reduce the unexplained autocorrelation in the models by incorporating time-varying fixed effects (in particular, the time points themselves) and possibly random slopes of such predictors. In the WS design, a random slope of the treatment effect can also be suggested to effectively split the series on each subject into the two treatment series.

## 2.7 References

## References

[1] Anderson, K., Brooks, A. S., Morrison, A. L., Reid-Smith, R. J., Martin, S. W., Benn, D. M., Peregrine, A. S., 2004. Impact of *Giardia* vaccination on asymptomatic *Giardia* infections in dogs at a research facility. *Canadian Veterinary Journal* **45**, 924–930.

[2] Averill, T., Rekaya, R., Weigel, K., 2006. Random regression models for male and female fertility evaluation using longitudinal binary data. *Journal of Dairy Science* **89**, 3681–3689.

[3] Barbosa, B., Goldstein, H., 2000. Discrete multilevel response models. *Quality and Quantity* **34**, 323–330.

[4] Breslow, N. E., 2003. Whither PQL?. University of Washington Biostatistics Working Paper Series **192**.

[5] Breslow, N. E., Clayton, D. G., 1993. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.

[6] Browne, W. J., Draper, D., 2006. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis* **3**, 473–514.

[7] Carey, V., Zeger, S. L., Diggle, P., 1993. Modeling multivariate binary data with alternating logistic regressions. *Biometrika* **80**, 517–526.

[8] Chaganty, N. R., 2004. Efficiency of generalized estimating equations for binary responses. *Journal of the Royal Statistical Society, Series B* **66**, 851–860.

[9] Congdon, P., 2003. *Applied Bayesian Modelling.* Wiley, New York.

[10] Diggle, P. J., Heagerty, P., Liang, K.-Y., Zeger, S. L., 2002. *Analysis of Longitudinal Data*, 2nd ed., Oxford University Press, Oxford.

[11] Dohoo, I. R., Martin, S. W., Stryhn, H., 2003. *Veterinary Epidemiologic Research.* AVC Inc., Charlottetown, Canada; web-site: `http://www.upei.ca/ver`.

[12] Dohoo, I. R., Stryhn, H., 2006. Simulation studies on the effects of clustering. XI*th* International Conference of Veterinary Epidemiology and Economics, Cairns, Australia, August 2006.

[13] Evans, B. A., Feng, Z., Peterson, A. V., 2001. A comparison of generalized linear mixed model procedures with estimating equations for variance and covariance parameter estimation in longitudinal studies and group randomized trials. *Statistics in Medicine* **20**, 3353–3373.

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

[15] Goldstein, H., Rasbach, J., 1996. Improved approximations for multilevel models with binary responses, *Journal of the Royal Statistical Society, Series A* **159**, 505–513.

[16] Goldstein, H., Browne, W. J., Rasbach, J., 2002. Multilevel modelling of medical data. *Statistics in Medicine* **21**, 3292–3315.

[17] Griswold, M. E., Zeger, S. L., 2004. On marginalized multilevel models and their computation, Johns Hopkins University, Dept. of Biostatistics, Working Papers No. **99**.

[18] Hamilton, D. C., 1992. Analysis of Fish Behaviour Data. *Canadian Journal of Statistics* **20**, 228–233.

[19] Hardin, J. W., Hilbe, J. M., 2003. *Generalized Estimating Equations.* Chapman & Hall/CRC, Boca Raton.

[20] Heagerty, P. J., Zeger, S. L., 2000. Marginalized multilevel models and likelihood inference. *Statistical Science* **15**, 1–19.

[21] Heo, M., Leon, A. C., 2005. Comparison of statistical methods for the analysis of clustered binary observations. *Statistics in Medicine* **24**, 911–923.

[22] Hu, F. B., Goldberg, J., Hedeker, D., Flay B. R., Pentz, M. A., 1998. Comparison of Population-Averaged and Subject-Specific Approaches for Analyzing Repeated Binary Outcomes. *American Journal of Epidemiology* **147**, 694–703.

[23] Kimber, W., 2007. Comparison of Phenobarbital and Potassium Bromide Monotherapies in the Treatment of Canine Epilepsy. MSc Thesis, Department of Biomedical Sciences, Atlantic Veterinary College, Charlottetown, Canada.

[24] Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R., 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.

[25] Leisch, F., Weingessel, A., Hornik, K., 1998. On the generation of correlated artificial binary data. Vienna University of Economics and Business Administration, Working Papers Series **13**.

[26] Lesaffre, E., Spiessens, B., 2001. On the effect of the number of quadrature points in a logistic random-effects model: an example. *Journal of the Royal Statistical Society, Series C* **50**, 325–335.

[27] Liang, K. Y., Zeger, S. L., 1986, Longitudinal Data-Analysis Using Generalized Linear-Models. *Biometrika* **73**, 13–22.

[28] Lindsey, J. K., Lambert, P., 1998. On the appropriateness of marginal models for repeated measurements in clinical trials. *Statistics in Medicine* **17**, 447–469.

[29] Mancl, L., Leroux, B., 1996. Efficiency of regression estimates for clustered data. *Biometrics* **52**, 500–511.

[30] McDermott, J. J., Schukken, Y. H., Shoukri, M. M., 1994. Methods for analysing data collected from clusters of animals. *Preventive Veterinary Medicine* **18**, 175–192.

[31] Molenberghs, G., Verbeke, G., 2005. *Models for Discrete Longitudinal Data.* Springer, New York.

[32] Neuhaus, J. M., Kalbfleisch, J. D., Hauck, W. W., 1991. A comparison of cluster-specific and population averaged approaches for analyzing correlated binary data. *International Statistical Review* **59**, 25–36.

[33] Neuhaus, J. M., 1992. Statistical methods for longitudinal and clustered design with binary responses. *Statistical Methods in Medical Research* **1**, 249–273.

[34] Olde Riekerink, R. G. M., Barkema, H. W., Stryhn, H., 2007. The effect of season on somatic cell count and the incidence of clinical mastitis. *Journal of Dairy Science* **90**, 1704–1715.

[35] Prentice, R. L., 1988. Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44**, 1033–1048.

[36] Rabe-Hesketh, S., Skrondal, A., 2008. *Multilevel and Longitudinal Modeling using Stata*, 2nd ed. Stata Press.

[37] Rabe-Hasketh, S., Skrondal, A., Pickles, A., 2002. Reliable estimation of generalised linear mixed models using adaptive quadrature. *The Stata Journal* **2**, 1–21.

[38] Rasmussen, M. D., Bjerring, M., Justesen, P., Jepsen, L., 2002. Milk quality on Danish farms with automatic milking systems. *Journal of Dairy Science* **85**, 2869–2878.

[39] Rodríguez, 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.

[40] Rodríguez, 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.

[41] Sanchez, J., Dohoo, I. R., Nødtvedt, A., Keefe, G. P., Markham, F., Leslie, K., DesCôteaux, L., Campbell, J., 2002. A longitudinal study of gastrointestinal parasites in Canadian dairy farms. The

value of an indirect *Ostertagia ostertagi* ELISA as a monitoring tool. *Veterinary parasitology* **107**, 209–226.

[42] Schildcrout, J. S., Heagerty, P. J., 2005. Regression analysis of longitudinal binary data with time-dependent environmental covariates: bias and efficiency. *Biostatistics* **6**, 633–652.

[43] Schukken, Y. H., Grohn, Y. T., McDermott B., McDermott J. J., 2003. Analysis of correlated discrete observations: background, examples and solutions. *Preventive Veterinary Medicine* **59**, 223–40.

[44] Snijders, T. A. B., Bosker, R. J., 1999. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modelling.* Sage Publishers, London.

[45] Speer, C. A., Scott, M. Cathy, Bannantine, John P., Waters, W. Ray, Mori, Yasuyuki, Whitlock, Robert H., Eda, Shigetoshi., 2006. A novel enzyme-linked immunosorbent assay for diagnosis of mycobacterium avium subsp. paratuberculosis Infections (Johne's Disease) in Cattle. *Clinical And Vaccine Immunology* **13**, 535–540.

[46] Stryhn, H., Dohoo, I. R., Tillard, E., 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, August 2000.

[47] Sturtz, S., Ligges, U., Gelman, A., 2005. R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software* **12**, 1–16.

[48] Veissier, I., Boissy, A., dePassillé, A. M., Rushen, J., van Reenen, C. G., Roussel, S., Andanson, S., Pradel, P., 2001. Calves' responses to repeated social regrouping and relocation. *Journal of Animal Science* **79**, 2580–2593.

[49] Wang, Y. G., Carey, V., 2003. Working correlation stucture misspecification, estimation and covariate design: Implication for generalised estimatting equation performance. *Biometrika* **90**, 29–41.

[50] Wolfinger, R., O'Connell, M., 1993. Generalized linear mixed models: a pseudo likelihood approach. *Communications in Statistics: Simulation and Computation* **48**, 233–243.

[51] Yang, M., Goldstein, H., Heath, A., 2000. Multilevel models for repeated binary outcomes: attitudes and voting over the electoral cycle. *Journal of the Royal Statistical Society, Series A* **163**, 49–62.

[52] Zeger, S. L., Liang, K. Y., Albert, P. S., 1988. Models for longitudinal data - a generalized estimating equation approach. *Biometrics* **44**, 1049–1060.

[53] Ziegler, A., Kastner, C., Blettner, M. 1998. The generalized estimating equations: an annotated bibliography. *Biometrical Journal* **40**, 115–139.

Table 2.1: Mean estimate of between-subjects (**BS**) treatment effect (true value = 0.35), followed in parenthesis by standard deviation among simulations and mean standard error, based on analyses of 1000 simulated marginal (**PA**) datasets per setting ($n$ = number of subjects, $t$ = number of time points, $\rho$ = autocorrelation). Analysis by procedure B of type $A$ is designated by $\hat{\beta}^{A}_{\mathrm{B}}$, where $A = PA$ (population-averaged) or $SS$ (subject-specific), and B = F (generalized estimating equations (GEE) with fixed autoregressive correlation), AR (GEE with autoregressive correlation), ALR (alternating logistic regression), MQL (marginal quasi-likelihood), REPL (restricted pseudo-likelihood), PQL (2nd order penalized quasi-likelihood), ML (maximum likelihood), MCMC (Bayesian Markov chain Monte Carlo).

| | | | Statistical Methods$^{\top}$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | $t$ | $\rho$ | $\hat{\beta}^{PA}_{\mathrm{F}}$ | $\hat{\beta}^{PA}_{\mathrm{AR}}$ | $\hat{\beta}^{PA}_{\mathrm{ALR}}$ | $\hat{\beta}^{PA}_{\mathrm{MQL}}$ | $\hat{\beta}^{SS}_{\mathrm{REPL}}$ | $\hat{\beta}^{SS}_{\mathrm{PQL}}$ | $\hat{\beta}^{SS}_{\mathrm{ML}}$ | $\hat{\beta}^{SS}_{\mathrm{MCMC}}$ |
| 100 | 16 | .7 | .359 (.22,.22) | .359 (.22,.22) | .360 (.23,.23) | .360 (.23,.23) | .357 (.22,.23*) | .375$^{\ddagger}$(.24,.24) | .381$^{\ddagger}$(.24,.24) | .383$^{\ddagger}$(.24,.24) |
| | | .5 | .359 (.17,.17) | .360 (.17,.17) | .360 (.17,.18) | .360 (.17,.18) | .360 (.17,.18*) | .369$^{\ddagger}$(.18,.18) | .370$^{\ddagger}$(.18,.18) | .371$^{\ddagger}$(.18,.18) |
| | | .2 | .352 (.13,.13) | .352 (.13,.13) | .352 (.13,.13) | .352 (.13,.13) | .353 (.13,.13) | .355 (.13,.13) | .355 (.13,.13) | .355 (.13,.13) |
| | 8 | .7 | .344 (.27,.27) | .345 (.27,.27) | .342 (.28,.28) | .342 (.28,.28) | .340 (.27,.28*) | .348 (.28,.30*) | .371$^{\dagger}$(.30,.30) | .376$^{\ddagger}$(.30,.30) |
| | | .5 | .348 (.22,.22) | .348 (.22,.22) | .348 (.22,.23) | .348 (.23,.23) | .348 (.22,.23*) | .361 (.23,.23) | .366$^{\dagger}$(.24,.24) | .368$^{\dagger}$(.24,.24) |
| | | .2 | .349 (.17,.17) | .349 (.17,.17) | .349 (.17,.17) | .349 (.17,.17) | .353 (.17,.18*) | .355 (.17,.17) | .355 (.17,.17) | .351 (.17,.17) |
| | 4 | .7 | .342 (.34,.33) | .342 (.34,.33) | .341 (.35,.33) | .341 (.35,.34) | .340 (.33,.32) | .319$^{\ddagger}$(.33,.36*) | .366 (.37,.36) | .362 (.37,.35*) |
| | | .5 | .341 (.30,.29) | .341 (.30,.29) | .343 (.30,.29) | .343 (.30,.30) | .347 (.30,.29) | .351 (.31,.30) | .366 (.32,.31) | .344 (.31,.31) |
| | | .2 | .339 (.24,.23) | .340 (.24,.29) | .340 (.24,.23) | .340 (.24,.24) | .340 (.24,.24) | .349 (.25,.23*) | .350 (.25,.24) | .341 (.25,.23*) |
| 20 | 16 | .7 | .351 (.54,.49*) | .352 (.54,.49*) | .351 (.55,.51*) | .351 (.55,.53) | .346 (.52,.53) | .363 (.57,.55) | .371 (.58,.53*) | .358 (.58,.57) |
| | | .5 | .347 (.40,.37*) | .347 (.40,.37*) | .348 (.41,.38*) | .348 (.41,.40) | .346 (.40,.42) | .357 (.42,.41) | .357 (.42,.39*) | .364 (.43,.42) |
| | | .2 | .347 (.30,.27*) | .347 (.30,.27*) | .347 (.30,.29) | .347 (.30,.29) | .348 (.29,.31*) | .350 (.30,.29) | .349 (.30,.29) | .372$^{\dagger}$(.30,.30) |
| | 8 | .7 | .377 (.63,.61) | .379 (.64,.61) | .379 (.66,.63*) | .381 (.66,.66) | .348 (.60,.61) | .374$^{\ddagger}$(.65,.69*) | .405$^{\dagger}$(.70,.68) | .397 (.70,.71) |
| | | .5 | .371 (.51,.49) | .373 (.51,.49) | .373 (.53,.50*) | .373 (.53,.53) | .359 (.50,.53*) | .382 (.54,.54) | .386$^{\dagger}$(.55,.52*) | .386$^{\dagger}$(.55,.57) |
| | | .2 | .363 (.39,.37*) | .364 (.39,.37*) | .364 (.40,.37*) | .364 (.39,.39) | .365 (.39,.41) | .370 (.40,.39) | .369 (.40,.38*) | .393$^{\ddagger}$(.41,.40) |
| | 4 | .7 | .410$^{\dagger}$(.82,.75*) | .413$^{\dagger}$(.82,.75*) | .406$^{\dagger}$(.83,.76* | .406$^{\dagger}$(.83,.76*) | .339 (.71,.70) | .365 (.73,.71) | .414$^{\dagger}$(.84,.82) | .428$^{\ddagger}$(.85,.81*) |
| | | .5 | .397$^{\dagger}$(.70,.65*) | .400$^{\dagger}$(.70,.65*) | .395 (.71,.66*) | .394 (.71,.69) | .353 (.63,.63) | .362 (.65,.63) | .414$^{\ddagger}$(.74,.70*) | .409$^{\dagger}$(.74,.77) |
| | | .2 | .382 (.57,.52*) | .381 (.57,.52*) | .379 (.57,.52*) | .379 (.57,.55) | .365 (.56,.55) | .387$^{\dagger}$(.58,.55*) | .384 (.58,.55*) | .409$^{\ddagger}$(.61,.61) |

$^{\dagger}$ significant bias in estimate at $P < 0.05$; $^{\ddagger}$ significant bias in estimate at $P < 0.01$; $^{*}$ significant bias in standard error at $P < 0.05$

$^{\top}$ Note that SS estimates were converted to PA value (see text).

Table 2.2: Mean estimate of within-subjects (**WS**) treatment effect (true value = 0.35), followed in parenthesis by standard deviation among simulations and mean standard error, based on analyses of 1000 simulated marginal (**PA**) datasets per setting ($n$ = number of subjects, $t$ = number of time points, $\rho$ = autocorrelation). See Table 2.1 for coding of statistical methods.

| | | | Statistical Methods $^{\top}$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | $t$ | $\rho$ | $\hat{\beta}^{PA}_{\mathrm{F}}$ | $\hat{\beta}^{PA}_{\mathrm{AR}}$ | $\hat{\beta}^{PA}_{\mathrm{ALR}}$ | $\hat{\beta}^{PA}_{\mathrm{MQL}}$ | $\hat{\beta}^{SS}_{\mathrm{REPL}}$ | $\hat{\beta}^{SS}_{\mathrm{PQL}}$ | $\hat{\beta}^{SS}_{\mathrm{ML}}$ | $\hat{\beta}^{SS}_{\mathrm{MCMC}}$ |
| 100 | 16 | .7 | .352 (.14,.14) | .352 (.14,.14) | .356 (.18,.18) | .356 (.18,.11$^*$) | .357 (.14,.14) | .371$^{\ddagger}$(.19,.10$^*$) | .369$^{\ddagger}$(.19,.10$^*$) | .368$^{\ddagger}$(.19,.10$^*$) |
| | | .5 | .359$^{\dagger}$(.14,.14) | .358 (.14,.14) | .359 (.16,.16) | .359 (.16,.11$^*$) | .361$^{\dagger}$(.14,.14) | .369$^{\ddagger}$(.16,.10$^*$) | .368$^{\ddagger}$(.16,.10$^*$) | .369$^{\ddagger}$(.16,.10$^*$) |
| | | .2 | .355 (.12,.12) | .355 (.12,.12) | .355 (.13,.12) | .355 (.13,.11$^*$) | .352 (.12,.12) | .357 (.13,.11$^*$) | .357 (.13,.11$^*$) | .360$^{\dagger}$(.13,.11$^*$) |
| | 8 | .7 | .350 (.15,.14) | .350 (.15,.14) | .349 (.18,.18) | .349 (.19,.14$^*$) | .357 (.15,.14$^*$) | .386$^{\ddagger}$(.20,.12$^*$) | .370$^{\ddagger}$(.20,.12$^*$) | .369$^{\ddagger}$(.20,.12$^*$) |
| | | .5 | .352 (.17,.16) | .352 (.17,.16) | .353 (.19,.18) | .353 (.19,.14$^*$) | .356 (.17,.16$^*$) | .371$^{\ddagger}$(.20,.13$^*$) | .370$^{\ddagger}$(.20,.13$^*$) | .371$^{\ddagger}$(.20,.13$^*$) |
| | | .2 | .355 (.16,.16) | .355 (.16,.16) | .355 (.16,.16) | .355 (.16,.14$^*$) | .357 (.16,.16) | .362$^{\dagger}$(.17,.14$^*$) | .361$^{\dagger}$(.17,.14$^*$) | .361$^{\dagger}$(.17,.14$^*$) |
| | 4 | .7 | .355 (.15,.15) | .354 (.15,.15) | .353 (.17,.17) | .353 (.17,.20$^*$) | .352 (.15,.14$^*$) | .410$^{\ddagger}$(.19,.16$^*$) | .378$^{\ddagger}$(.18,.15$^*$) | .386$^{\ddagger}$(.18,.15$^*$) |
| | | .5 | .359 (.19,.19) | .359 (.19,.19) | .355 (.20,.20) | .355 (.20,.20) | .361 (.19,.18) | .381$^{\ddagger}$(.21,.17$^*$) | .377$^{\ddagger}$(.21,.18$^*$) | .368$^{\ddagger}$(.20,.18$^*$) |
| | | .2 | .356 (.20,.21) | .356 (.20..21) | .355 (.20,.21) | .355 (.20,.20) | .356 (.20,.20) | .365$^{\dagger}$(.21,.19$^*$) | .365$^{\dagger}$(.21,.20$^*$) | .354 (.22,.20$^*$) |
| 20 | 16 | .7 | .366 (.31,.30) | .366 (.31,.30) | .386$^{\ddagger}$(.42,.39$^*$) | .387$^{\ddagger}$(.43,.24$^*$) | .363 (.32,.30$^*$) | .404$^{\ddagger}$(.44,.22$^*$) | .401$^{\ddagger}$(.44,.23$^*$) | .396$^{\ddagger}$(.44,.22$^*$) |
| | | .5 | .382$^{\ddagger}$(.33,.31$^*$) | .382$^{\ddagger}$(.33,.31$^*$) | .383$^{\ddagger}$(.36,.35) | .383$^{\ddagger}$(.36,.24$^*$) | .385$^{\ddagger}$(.33,.32) | .392$^{\ddagger}$(.37,.23$^*$) | .392$^{\ddagger}$(.37,.23$^*$) | .394$^{\ddagger}$(.37,.24$^*$) |
| | | .2 | .373$^{\ddagger}$(.28,.27) | .373$^{\ddagger}$(.28,.27) | .372$^{\dagger}$(.28,.27) | .372$^{\dagger}$(.28,.24$^*$) | .372$^{\dagger}$(.28,.28) | .375$^{\ddagger}$(.28,.24$^*$) | .374$^{\ddagger}$(.28,.24$^*$) | .389$^{\ddagger}$(.29,.24$^*$) |
| | 8 | .7 | .356 (.33,.31$^*$) | .354 (.33,.31$^*$) | .358 (.41,.39$^*$) | .359 (.42,.33$^*$) | .338 (.33,.30$^*$) | .392$^{\ddagger}$(.46,.28$^*$) | .377 (.44,.28$^*$) | .372 (.44,.27$^*$) |
| | | .5 | .349 (.38,.36$^*$) | .347 (.38,.36$^*$) | .356 (.42,.40) | .355 (.42,.33$^*$) | .344 (.38,.36$^*$) | .373 (.44,.30$^*$) | .371 (.44,.31$^*$) | .372 (.44,.31$^*$) |
| | | .2 | .351 (.36,.34$^*$) | .351 (.36,.34$^*$) | .350 (.37,.35$^*$) | .350 (.37,.32$^*$) | .352 (.36,.35) | .357 (.37,.31$^*$) | .356 (.37,.32$^*$) | .373$^{\dagger}$(.38,.33$^*$) |
| | 4 | .7 | .361 (.35,.33$^*$) | .357 (.35,.33$^*$) | .358 (.38,.36$^*$) | .357 (.38,.46$^*$) | .308$^{\ddagger}$(.33,.32) | .413$^{\ddagger}$(.43,.38$^*$) | .384$^{\dagger}$(.41,.36$^*$) | .406$^{\ddagger}$(.43,.37$^*$) |
| | | .5 | .361 (.43,.41$^*$) | .356 (.43,.41$^*$) | .362 (.45,.43) | .363 (.45,.46) | .338 (.42,.39$^*$) | .388$^{\dagger}$(.48,.40$^*$) | .382$^{\dagger}$(.47,.41$^*$) | .389$^{\dagger}$(.48,.40$^*$) |
| | | .2 | .353 (.47,.46) | .353 (.46,.46) | .356 (.47,.46) | .356 (.47,.46) | .344 (.46,.45) | .368 (.48,.43$^*$) | .365 (.48,.45$^*$) | .388$^{\dagger}$(.51,.47$^*$) |

$^{\dagger}$ significant bias in estimate at $P < 0.05$; $^{\ddagger}$ significant bias in estimate at $P < 0.01$; $^*$ significant bias in standard error at $P < 0.05$
$^{\top}$ Note that SS estimates were converted to PA value (see text).

Table 2.3: Mean estimate of between-subjects (**BS**) treatment effect (true value = 0.35, marginal true value = 0.302), followed in parenthesis by standard deviation among simulations and mean standard error, based on analyses of 1000 simulated random effects (**SS**) datasets per setting ($n$ = number of subjects, $t$ = number of time points, $\rho$ = autocorrelation). See Table 2.1 for coding of statistical methods.

|  |  |  | Statistical Methods |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
| $n$ | $t$ | $\rho$ | $\hat{\beta}^{PA}_{\mathrm{AR}}$ | $\hat{\beta}^{PA}_{\mathrm{ALR}}$ | $\hat{\beta}^{PA}_{\mathrm{MQL}}$ | $\hat{\beta}^{SS}_{\mathrm{REPL}}$ | $\hat{\beta}^{SS}_{\mathrm{PQL}}$ | $\hat{\beta}^{SS}_{\mathrm{ML}}$ | $\hat{\beta}^{SS}_{\mathrm{MCMC}}$ |
| 100 | 16 | 1 | .287$^{\dagger}$(.20,.19) | .286$^{\dagger}$(.20,.19) | .286$^{\dagger}$(.20,.19) | .325$^{\ddagger}$(.23,.22) | .341 (.24,.23) | .343 (.24,.23) | .345 (.24,.23) |
|  |  | .7 | .287$^{\ddagger}$(.14,.13$^{*}$) | .287$^{\ddagger}$(.14,.13$^{*}$) | .287$^{\ddagger}$(.14,.13$^{*}$) | .291$^{\ddagger}$(.14,.13) | .299$^{\ddagger}$(.14,.14) | .299$^{\ddagger}$(.14,.14) | .297$^{\ddagger}$(.15,.14) |
|  |  | .5 | .291$^{\ddagger}$(.12,.12) | .291$^{\ddagger}$(.12,.12) | .291$^{\dagger}$(.12,.12) | .293$^{\ddagger}$(.12,.12) | .297$^{\ddagger}$(.12,.12) | .296$^{\ddagger}$(.12,.12) | .297$^{\ddagger}$(.12,.12) |
|  |  | .2 | .291$^{\ddagger}$(.11,.11) | .291$^{\ddagger}$(.11,.11) | .290$^{\ddagger}$(.11,.11) | .293$^{\ddagger}$(.11,.11) | .293$^{\ddagger}$(.11,.11) | .293$^{\ddagger}$(.11,.11) | .297$^{\ddagger}$(.11,.11) |
|  | 8 | 1 | .300 (.22,.21) | .300 (.22,.21) | .300 (.22,.21) | .333$^{\dagger}$(.24,.24) | .357 (.26,.25) | .361 (.26,.25) | .364 (.27,.26) |
|  |  | .7 | .292 (.18,.17) | .292 (.18,.17) | .292 (.18,.17) | .297$^{\ddagger}$(.18,.18) | .313$^{\ddagger}$(.19,.18) | .313$^{\ddagger}$(.19,.19) | .310$^{\ddagger}$(.19,.18) |
|  |  | .5 | .293 (.16,.16) | .293 (.16,.16) | .293 (.16,.16) | .295$^{\ddagger}$(.16,.16) | .304$^{\ddagger}$(.17,.16) | .304$^{\ddagger}$(.17,.16) | .302$^{\ddagger}$(.17,.16) |
|  |  | .2 | .296 (.15,.15) | .296 (.15,.15) | .296 (.15,.15) | .296$^{\ddagger}$(.15,.16) | .306$^{\ddagger}$(.15,.15) | .300$^{\ddagger}$(.15,.15) | .299$^{\ddagger}$(.16,.15) |
|  | 4 | 1 | .306 (.26,.25) | .306 (.26,.25) | .306 (.26,.25) | .336 (.28,.27) | .361 (.30,.28$^{*}$) | .369 (.31,.30) | .368 (.31,.31) |
|  |  | .7 | .303 (.24,.23) | .302 (.24,.23) | .301 (.24,.23) | .310$^{\ddagger}$(.25,.24) | .332$^{\dagger}$(.26,.25) | .335 (.26,.26) | .331$^{\dagger}$(.26,.25) |
|  |  | .5 | .297 (.23,.22) | .296 (.23,.22) | .295 (.23,.22) | .301$^{\ddagger}$(.24,.23) | .314$^{\ddagger}$(.24,.23) | .315$^{\ddagger}$(.24,.23) | .310$^{\ddagger}$(.24,.23) |
|  |  | .2 | .300 (.22,.21) | .299 (.22,.21) | .294 (.22,.22) | .297$^{\ddagger}$(.22,.22) | .306$^{\ddagger}$(.23,.22) | .307$^{\ddagger}$(.22,.22) | .303$^{\ddagger}$(.23,.22) |
| 20 | 16 | 1 | .302 (.43,.42) | .303 (.43,.42) | .303 (.43,.44) | .343 (.48,.50) | .365 (.51,.53) | .359 (.51,.49) | .361 (.53,.55) |
|  |  | .7 | .292 (.30,.28$^{*}$) | .292 (.30,.28$^{*}$) | .298 (.30,.30) | .303$^{\ddagger}$(.31,.32) | .311$^{\ddagger}$(.32,.32) | .301$^{\ddagger}$(.31,.29$^{*}$) | .320$^{\ddagger}$(.31,.30) |
|  |  | .5 | .284$^{\dagger}$(.27,.25$^{*}$) | .284$^{\dagger}$(.27,.25$^{*}$) | .289 (.27,.28) | .290$^{\ddagger}$(.28,.29) | .298$^{\ddagger}$(.28,.29) | .289$^{\ddagger}$(.27,.27) | .309$^{\ddagger}$(.28,.27) |
|  |  | .2 | .278$^{\ddagger}$(.25,.23$^{*}$) | .278$^{\ddagger}$(.25,.23$^{*}$) | .273$^{\ddagger}$(.25,.27$^{*}$) | .266$^{\ddagger}$(.25,.27$^{*}$) | .279$^{\ddagger}$(.26,.27) | .289$^{\ddagger}$(.27,.27) | .303$^{\ddagger}$(.25,.26) |
|  | 8 | 1 | .282 (.48,.46) | .282 (.48,.46) | .282 (.48,.48) | .311$^{\dagger}$(.54,.54) | .339 (.58,.58) | .333 (.57,.55) | .344 (.59,.61) |
|  |  | .7 | .290 (.41,.37$^{*}$) | .291 (.41,.37$^{*}$) | .293 (.41,.41) | .290$^{\ddagger}$(.42,.43) | .318$^{\dagger}$(.45,.44) | .308$^{\ddagger}$(.44,.40$^{*}$) | .328 (.45,.42$^{*}$) |
|  |  | .5 | .292 (.36,.34$^{*}$) | .292 (.36,.34$^{*}$) | .305 (.36,.39$^{*}$) | .308$^{\ddagger}$(.37,.41$^{*}$) | .326 (.39,.41$^{*}$) | .304$^{\ddagger}$(.38,.37) | .327 (.38,.38) |
|  |  | .2 | .291 (.33,.32) | .291 (.33,.32) | .292 (.36,.37) | .298$^{\ddagger}$(.34,.37$^{*}$) | .318$^{\dagger}$(.35,.38$^{*}$) | .296$^{\ddagger}$(.34,.35) | .325$^{\dagger}$(.35,.36) |
|  | 4 | 1 | .290 (.59,.55$^{*}$) | .290 (.59,.55$^{*}$) | .290 (.59,.58) | .331 (.67,.66) | .349 (.72,.70) | .344 (.71,.68$^{*}$) | .365 (.76,.78) |
|  |  | .7 | .316 (.52,.50) | .314 (.51,.50) | .314 (.51,.53) | .331 (.67,.66) | .353 (.60,.62) | .348 (.57,.57) | .369 (.60,.63$^{*}$) |
|  |  | .5 | .321 (.51,.48$^{*}$) | .320 (.50,.48$^{*}$) | .320 (.50,.52) | .353 (.55,.57) | .354 (.58,.59) | .346 (.54,.54) | .377 (.57,.59) |
|  |  | .2 | .320 (.47,.45) | .319 (.47,.45) | .339 (.46,.52$^{*}$) | .366 (.50,.54$^{*}$) | .373 (.52,.56$^{*}$) | .332 (.49,.51) | .367 (.52,.55$^{*}$) |

$^{\dagger}$ significant bias in estimate at $P < 0.05$; $^{\ddagger}$ significant bias in estimate at $P < 0.01$; $^{*}$ significant bias in standard error at $P < 0.05$

Table 2.4: Mean estimate of within-subjects (**WS**) treatment effect (true value = 0.35, marginal true value = 0.302), followed in parenthesis by standard deviation among simulations and mean standard error, based on analyses of 1000 simulated random effects (**SS**) datasets per setting ($n$ = number of subjects, $t$ = number of time points, $\rho$ = autocorrelation). See Table 2.1 for coding of statistical methods.

| | | | Statistical Methods | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $n$ | $t$ | $\rho$ | $\hat\beta^{PA}_{\text{AR}}$ | $\hat\beta^{PA}_{\text{ALR}}$ | $\hat\beta^{PA}_{\text{MQL}}$ | $\hat\beta^{SS}_{\text{REPL}}$ | $\hat\beta^{SS}_{\text{PQL}}$ | $\hat\beta^{SS}_{\text{ML}}$ | $\hat\beta^{SS}_{\text{MCMC}}$ |
| 100 | 16 | 1 | .294$^\dagger$(.10,.10) | .294$^\dagger$(.10,.10) | .294$^\dagger$(.10,.11$^*$) | .336$^\ddagger$(.11,.11) | .351 (.12,.11) | .351 (.12,.12) | .353 (.12,.12) |
| | | .7 | .292$^\ddagger$(.12,.12) | .293$^\dagger$(.12,.12) | .293$^\dagger$(.12,.11$^*$) | .297$^\ddagger$(.12,.11$^*$) | .305$^\ddagger$(.13,.11$^*$) | .305$^\ddagger$(.13,.11$^*$) | .304$^\ddagger$(.13,.11$^*$) |
| | | .5 | .295$^\dagger$(.11,.11) | .295$^\dagger$(.11,.11) | .295$^\dagger$(.11,.11) | .296$^\ddagger$(.11,.11) | .301$^\ddagger$(.12,.11) | .300$^\ddagger$(.12,.11) | .303$^\ddagger$(.12,.11) |
| | | .2 | .294$^\dagger$(.11,.11) | .294$^\dagger$(.11,.11) | .294$^\dagger$(.11,.11) | .292$^\ddagger$(.11,.11) | .296$^\ddagger$(.11,.11) | .295$^\ddagger$(.11,.11) | .303$^\ddagger$(.11,.11) |
| | 8 | 1 | .286$^\ddagger$(.14,.13$^*$) | .285$^\ddagger$(.14,.13) | .285$^\ddagger$(.14,.14) | .320$^\ddagger$(.15,.15) | .342 (.16,.15$^*$) | .344 (.16,.16) | .345 (.16,.16) |
| | | .7 | .284$^\ddagger$(.15,.15) | .284$^\ddagger$(.15,.15) | .284$^\ddagger$(.15,.14$^*$) | .290$^\ddagger$(.15,.15) | .304$^\ddagger$(.16,.15$^*$) | .305$^\ddagger$(.16,.15$^*$) | .302$^\ddagger$(.16,.15$^*$) |
| | | .5 | .284$^\ddagger$(.15,.15) | .284$^\ddagger$(.15,.15) | .286$^\ddagger$(.15,.14$^*$) | .287$^\ddagger$(.15,.15) | .296$^\ddagger$(.15,.14$^*$) | .294$^\ddagger$(.15,.15) | .294$^\ddagger$(.15,.14$^*$) |
| | | .2 | .290$^\dagger$(.15,.15) | .290$^\dagger$(.15,.15) | .290$^\dagger$(.15,.14$^*$) | .295$^\ddagger$(.15,.14$^*$) | .296$^\ddagger$(.15,.14$^*$) | .294$^\ddagger$(.15,.14$^*$) | .297$^\ddagger$(.15,.14$^*$) |
| | 4 | 1 | .291 (.18,.18) | .291 (.18,.18) | .291 (.18,.20$^*$) | .313$^\ddagger$(.20,.20) | .347 (.22,.21) | .354 (.22,.22) | .353 (.22,.22) |
| | | .7 | .287$^\dagger$(.19,.20) | .287$^\dagger$(.19,.20) | .287$^\dagger$(.19,.20) | .292$^\ddagger$(.20,.20) | .316$^\ddagger$(.21,.21) | .319$^\ddagger$(.22,.21) | .316$^\ddagger$(.22,.21) |
| | | .5 | .280$^\ddagger$(.20,.20) | .280$^\ddagger$(.20,.20) | .278$^\ddagger$(.20,.20) | .280$^\ddagger$(.20,.20) | .297$^\ddagger$(.21,.20) | .300$^\ddagger$(.21,.21) | .296$^\ddagger$(.22,.21) |
| | | .2 | .283$^\ddagger$(.20,.20) | .284$^\ddagger$(.20,.20) | .280$^\ddagger$(.20,.20) | .282$^\ddagger$(.21,.20) | .291$^\ddagger$(.21,.20$^*$) | .293$^\ddagger$(.21,.26$^*$) | .290$^\ddagger$(.21,.20) |
| 20 | 16 | 1 | .283$^\ddagger$(.23,.21$^*$) | .282$^\ddagger$(.23,.21$^*$) | .282$^\ddagger$(.23,.24) | .324$^\ddagger$(.26,.25) | .340 (.28,.26$^*$) | .336 (.27,.26$^*$) | .343 (.28,.26$^*$) |
| | | .7 | .301 (.27,.26) | .301 (.27,.26) | .306 (.27,.24$^*$) | .310$^\ddagger$(.27,.25$^*$) | .321$^\ddagger$(.28,.24$^*$) | .312$^\ddagger$(.28,.24$^*$) | .323$^\ddagger$(.28,.24$^*$) |
| | | .5 | .297 (.26,.24$^*$) | .297 (.26,.25) | .301 (.26,.24) | .300$^\ddagger$(.27,.25$^*$) | .311$^\ddagger$(.27,.24$^*$) | .303$^\ddagger$(.27,.24$^*$) | .318$^\ddagger$(.27,.24$^*$) |
| | | .2 | .294 (.25,.23$^*$) | .295 (.25,.23$^*$) | .295 (.25,.24$^*$) | .298$^\ddagger$(.25,.24$^*$) | .301$^\ddagger$(.26,.24$^*$) | .297$^\ddagger$(.26,.24$^*$) | .317$^\ddagger$(.26,.24$^*$) |
| | 8 | 1 | .291 (.31,.28$^*$) | .288 (.31,.28$^*$) | .288 (.31,.32) | .323$^\dagger$(.35,.33$^*$) | .348 (.37,.35$^*$) | .344 (.37,.36) | .351 (.38,.36$^*$) |
| | | .7 | .288 (.35,.32$^*$) | .289 (.35,.32$^*$) | .294 (.35,.32$^*$) | .302$^\ddagger$(.35,.33$^*$) | .322$^\dagger$(.38,.33$^*$) | .310$^\ddagger$(.37,.34$^*$) | .322$^\dagger$(.38,.34$^*$) |
| | | .5 | .303 (.34,.32$^*$) | .304 (.34,.32$^*$) | .310 (.33,.32) | .302$^\ddagger$(.35,.33$^*$) | .329 (.36,.33$^*$) | .317$^\ddagger$(.36,.33$^*$) | .333 (.36,.34$^*$) |
| | | .2 | .307 (.33,.32$^*$) | .308 (.33,.32) | .306 (.33,.32) | .309$^\ddagger$(.33,.32) | .320$^\dagger$(.34,.33) | .314$^\ddagger$(.34,.33) | .337 (.35,.33$^*$) |
| | 4 | 1 | .317 (.42,.40) | .318 (.42,.40) | .318 (.42,.46$^*$) | .353 (.51,.45$^*$) | .391$^\dagger$(.53,.50$^*$) | .390$^\dagger$(.53,.52) | .409$^\ddagger$(.55,.53$^*$) |
| | | .7 | .326 (.47,.44$^*$) | .329 (.47,.44$^*$) | .326 (.47,.46) | .362 (.53,.45$^*$) | .380 (.55,.48$^*$) | .372 (.54,.49$^*$) | .390$^\dagger$(.56,.51$^*$) |
| | | .5 | .318 (.49,.45$^*$) | .319 (.49,.45$^*$) | .308 (.48,.46) | .334 (.53,.45$^*$) | .350 (.55,.48$^*$) | .347 (.53,.48$^*$) | .370 (.56,.50$^*$) |
| | | .2 | .318 (.48,.45) | .319 (.48,.45) | .340$^\dagger$(.47,.46) | .344 (.51,.45$^*$) | .373 (.52,.47$^*$) | .337 (.50,.48$^*$) | .369 (.53,.49$^*$ ) |

$^\dagger$ significant bias in estimate at $P < 0.05$; $^\ddagger$ significant bias in estimate at $P < 0.01$; $^*$ significant bias in standard error at $P < 0.05$

Table 2.5: Summary of performance of estimation procedures across marginal (**PA**) data settings in terms of: bias of estimate and SE, coverage of confidence intervals and relative efficiency, with multiple patterns (involving at least two settings) represented by additional symbols in parenthesis, in decreasing order of occurrence. Coding for estimation procedures: F (generalized estimating equations (GEE) with fixed autoregressive correlation), AR (GEE with autoregressive correlation), ALR (alternating logistic regression), MQL (marginal quasi-likelihood), REPL (restricted pseudo-likelihood), PQL (2nd order penalized quasi-likelihood), ML (maximum likelihood), MCMC (Bayesian Markov chain Monte Carlo).

| Design | Between-subjects (BS) | | | | Within-subjects (WS) | | | |
|---|---|---|---|---|---|---|---|---|
| Assessment | bias $\hat{\beta}$ | bias SE | CI cov. | rel.eff. | bias $\hat{\beta}$ | bias SE | CI cov. | rel.eff. |
| Procedure | | | | | | | | |
| F | 0 (+) | 0 (−) | 0 (−) | n/a | 0 (+) | 0 (−) | 0 (−) | n/a |
| AR | 0 (+) | 0 (−) | 0 (−) | 0 | 0 (+) | 0 (−) | 0 (−) | 0 |
| ALR | 0 (+) | 0 (−) | 0 (−) | 0 (−) | 0 (+) | 0 (−) | 0 (−) | − (0) |
| MQL | 0 (+) | 0 | 0 (−) | 0 (−) | 0 (+) | − (0+) | − (0+) | − (0) |
| REPL | 0 | 0 (+) | 0 | 0 (+) | 0 (+) | 0/− | 0 (−) | 0 (+) |
| PQL | 0 (+−) | 0 (+−) | 0 (−) | − (0+) | + (0) | − | − (0) | − |
| ML | 0 (+) | 0 (−) | 0 (−) | − (0) | + (0) | − | − (0) | − |
| MCMC | + (0) | 0 (−) | − (0) | − (0) | + (0) | − | − (0) | − |

0 : no significant bias, nominal CI coverage, 100% efficiency<br>
−: downwards significant bias, significant CI undercoverage, <95% efficiency<br>
+: upwards significant bias, significant CI overcoverage, >105% efficiency

Table 2.6: Summary of performance of estimation procedures across random effects (**SS**) data settings separated by true value of autocorrelation ($\rho$) in terms of: bias of estimate and SE, as well as coverage of confidence intervals, with multiple patterns (involving at least two settings) represented by additional symbols in parenthesis, in decreasing order of occurrence. For coding of estimation procedures, see Table 2.4.

| Design | | Between-subjects (BS) | | | Within-subjects (WS) | | |
|---|---|---|---|---|---|---|---|
| Assessment | | bias $\hat{\beta}$ | bias SE | CI cov. | bias $\hat{\beta}$ | bias SE | CI cov. |
| Data | Procedure | | | | | | |
| $\rho \leq 1$ | AR | 0 (−) | 0 (−) | 0 (−) | 0/− | 0 (−) | − (0) |
| | ALR | 0 (−) | 0 (−) | − (0) | 0/− | 0 (−) | 0/− |
| | MQL | 0 (−) | 0 (+) | 0 (−) | 0 (−) | 0 (−+) | 0 (−+) |
| $\rho = 1$ | REPL | 0/− | 0 | 0 | − | 0 (−) | − (0) |
| | PQL | 0 | 0 | 0 | 0 | − (0) | 0 |
| | ML | 0 | 0 | 0 | 0 | 0 | 0 |
| | MCMC | 0 | 0 | 0 (−) | 0 | 0/− | 0 |
| $\rho < 1$ | REPL | − (0) | 0 (+) | 0 (−) | − (0) | 0/− | − (0) |
| | PQL | − (0) | 0 (+) | − (0+) | − (0) | − (0) | − (0) |
| | ML | − (0) | 0 (−) | 0 (−) | − (0) | − (0) | − (0) |
| | MCMC | − (0) | 0 (+) | − (0) | − (0) | − (0) | − (0) |

0 : no significant bias, nominal CI coverage
−: downwards significant bias, significant CI undercoverage
+: upwards significant bias, significant CI overcoverage

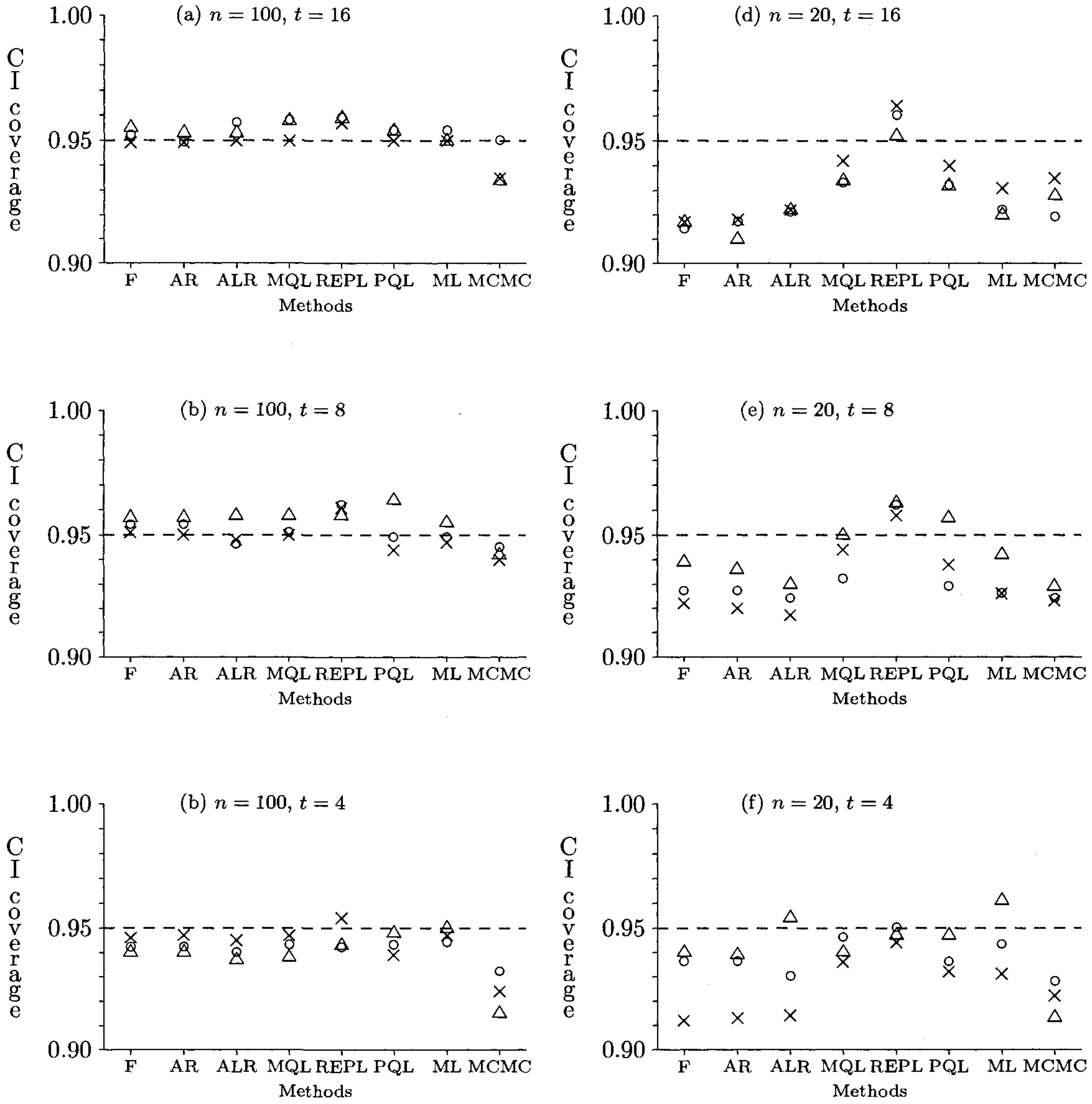

Figure 2.1: Confidence interval coverage for between-subjects (**BS**) treatment estimates of different estimation procedures, based on 1000 simulated marginal (**PA**) datasets per setting ($n$ = number of subjects, $t$ = number of time points, autocorrelation $\rho = (.7, .5, .2) \sim (\triangle, \circ, \times)$). Coding for estimation procedures: F (generalized estimating equations (GEE) with fixed autoregressive correlation), AR (GEE with autoregressive correlation), ALR (alternating logistic regression), MQL (marginal quasi-likelihood), REPL (restricted pseudo-likelihood), PQL (2nd order penalized quasi-likelihood), ML (maximum likelihood), MCMC (Bayesian Markov chain Monte Carlo).

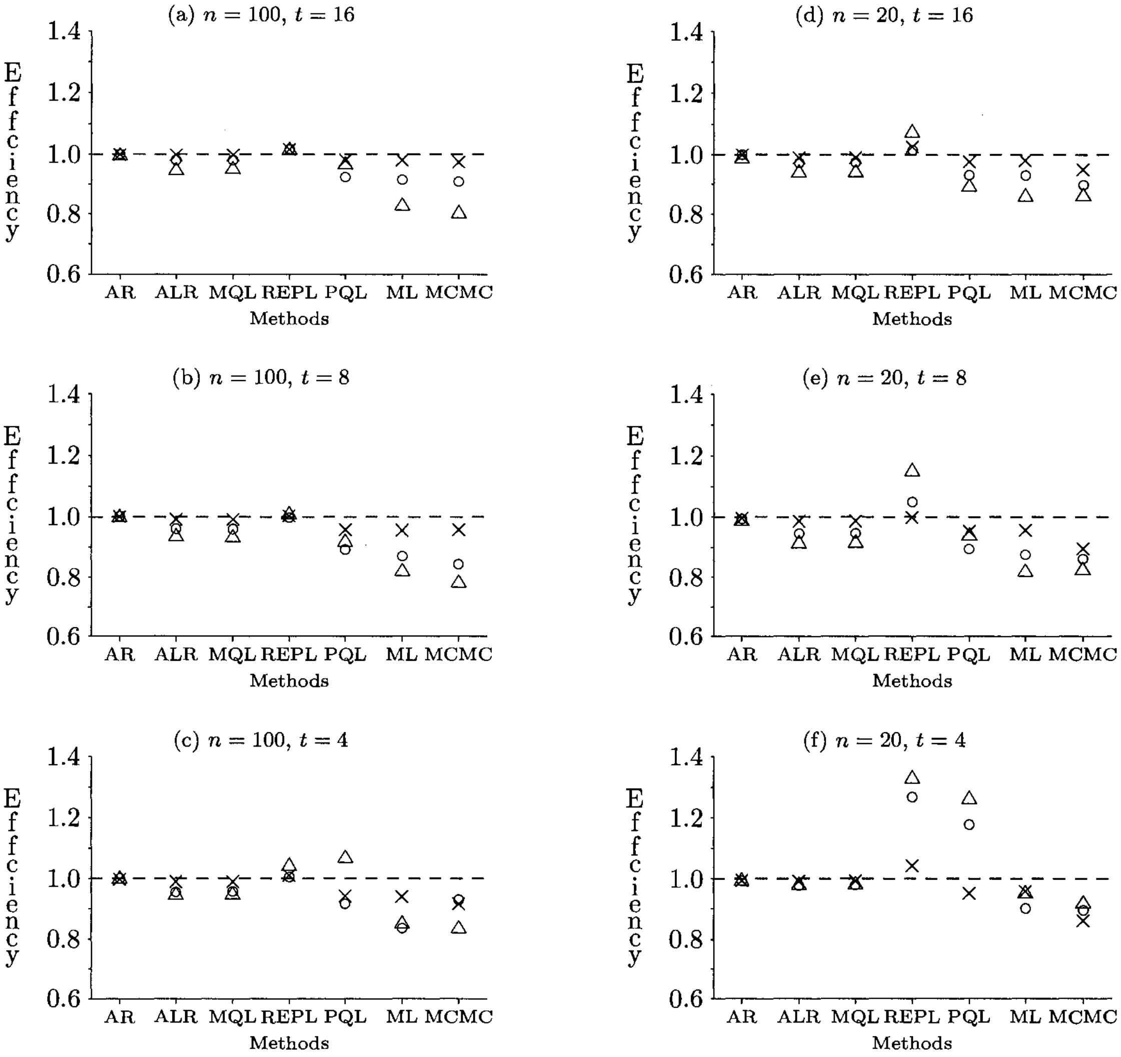

Figure 2.2: Relative efficiency (see text for definition) for between-subjects (**BS**) treatment estimates of different estimation procedures (see caption of Figure 2.1), based on 1000 simulated marginal (**PA**) datasets per setting ($n$ = number of subjects, $t$ = number of time points, autocorrelation $\rho = (.7, .5, .2) \sim (\triangle, \circ, \times)$).

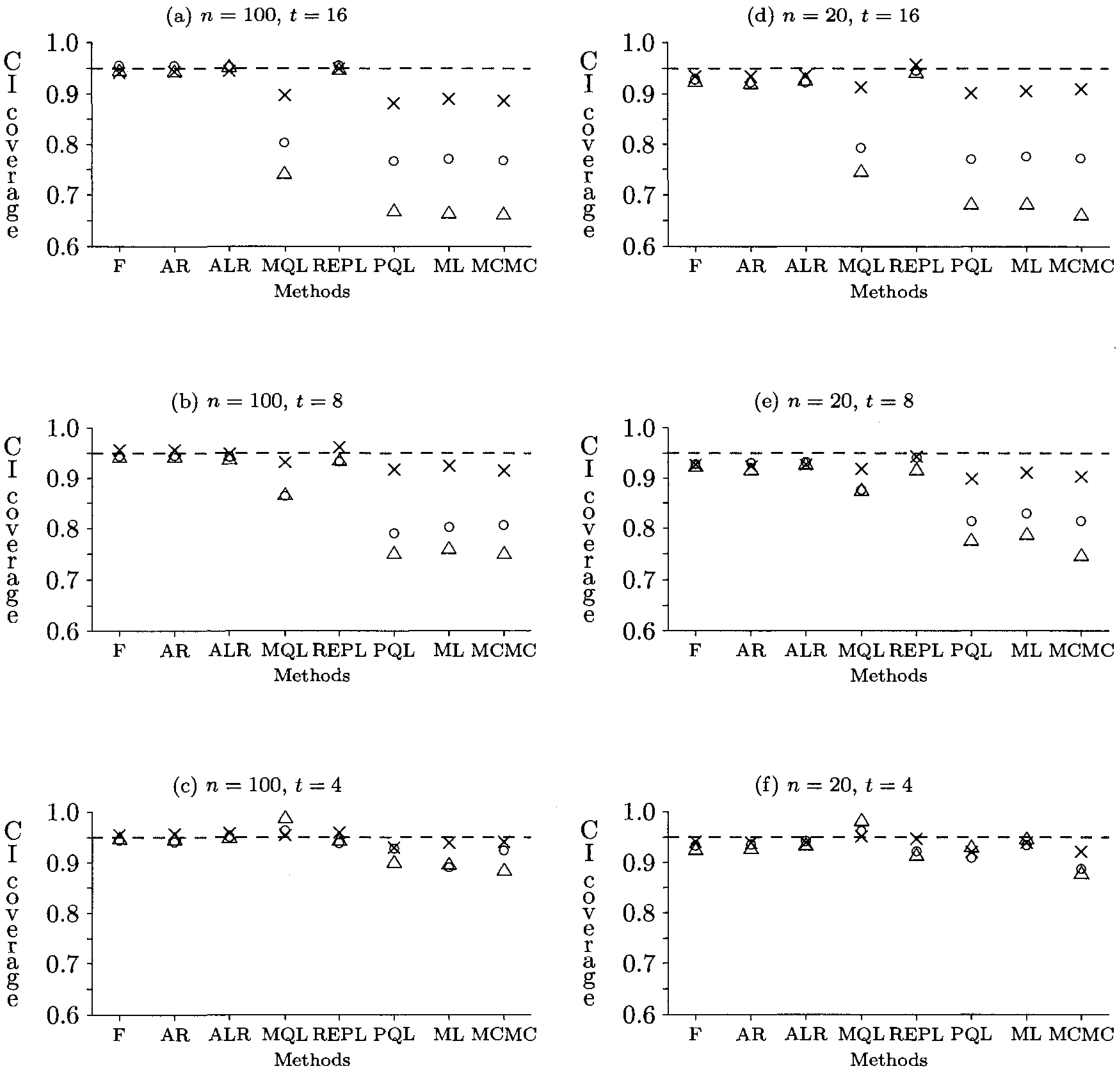

Figure 2.3: Confidence interval coverage for within-subjects (**WS**) treatment estimates of different estimation procedures (see caption of Figure 2.1), based on 1000 simulated marginal (**PA**) datasets per setting ($n$ = number of subjects, $t$ = number of time points, autocorrelation $\rho = (.7, .5, .2) \sim (\triangle, \circ, \times)$).

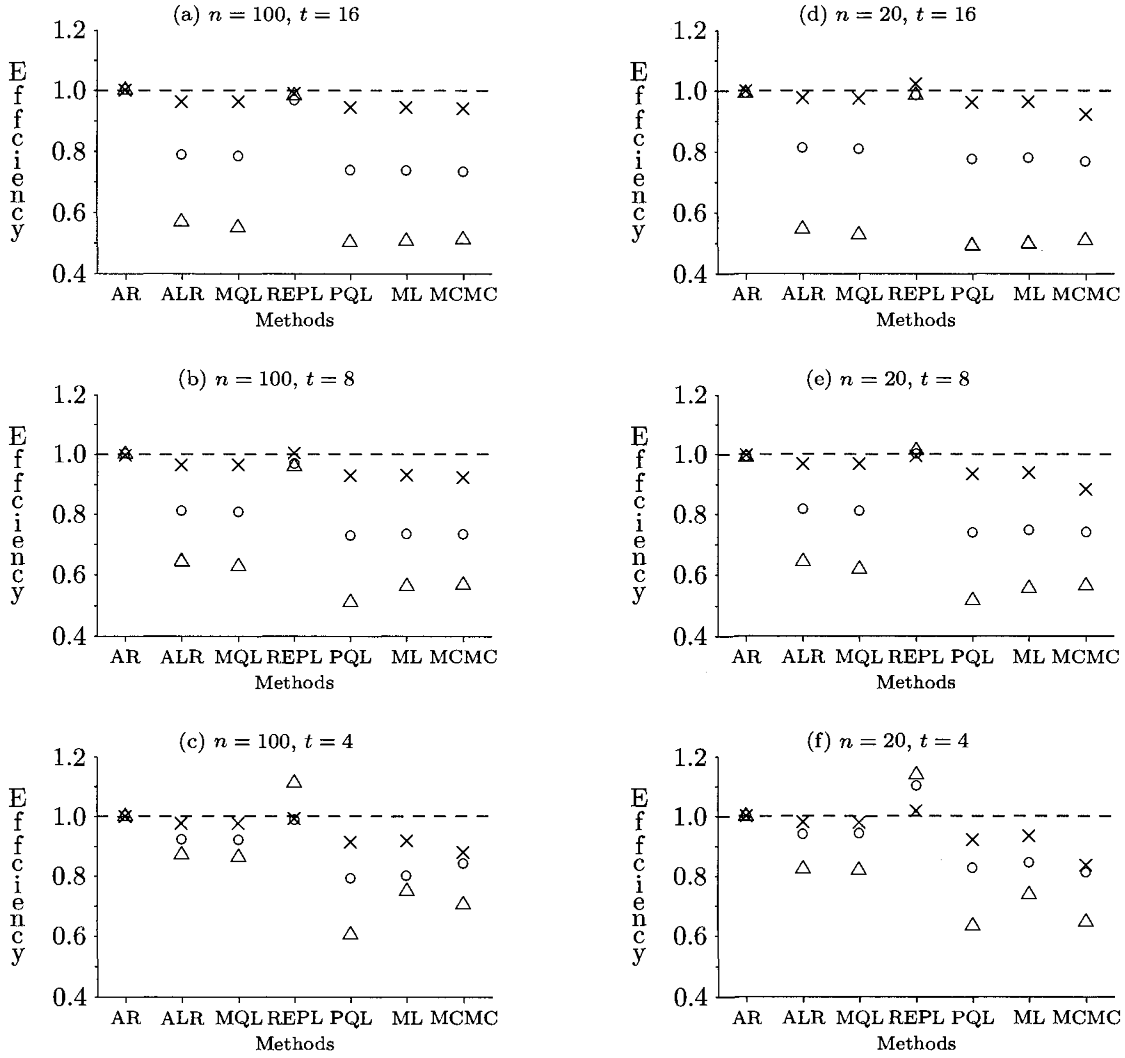

Figure 2.4: Relative efficiency (see text for definition) for within-subjects (**WS**) treatment estimates of different estimation procedures (see caption of Figure 2.1), based on 1000 simulated marginal (**PA**) datasets per setting ($n$ = number of subjects, $t$ = number of time points, autocorrelation $\rho = (.7, .5, .2) \sim (\triangle, \circ, \times)$).

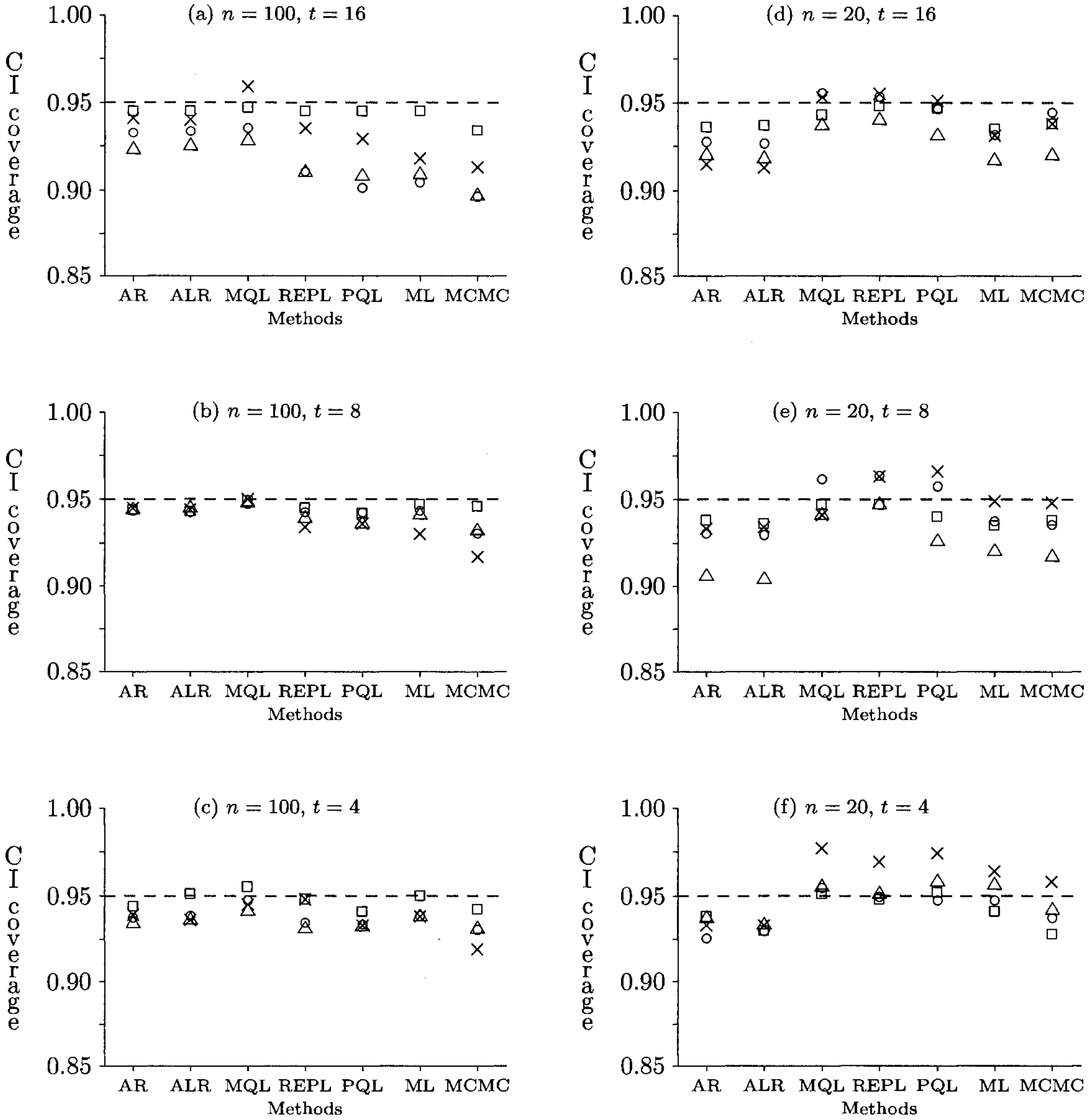

Figure 2.5: Confidence interval coverage for between-subjects (**BS**) treatment estimates of different estimation procedures (see caption of Figure 2.1), based on 1000 simulated random effects (**SS**) datasets per setting ($n$ = number of subjects, $t$ = number of time points, autocorrelation $\rho = (1, .7, .5, .2) \sim (\square, \triangle, \circ, \times)$).

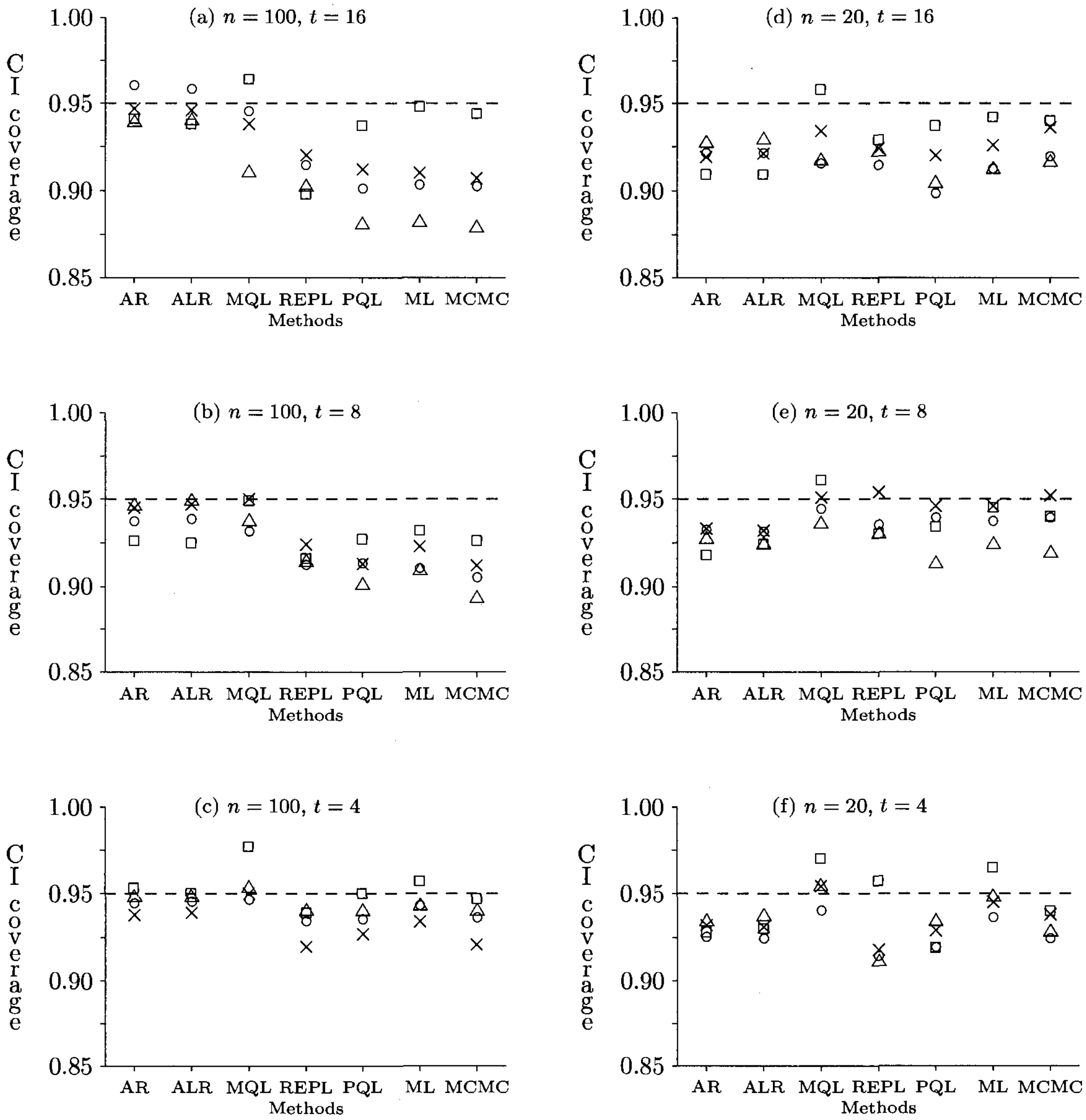

Figure 2.6: Confidence interval coverage for within-subjects (**WS**) treatment estimates of different estimation procedures (see caption of Figure 2.1), based on 1000 simulated random effects (**SS**) datasets per setting ($n$ = number of subjects, $t$ = number of time points, autocorrelation $\rho = (1, .7, .5, .2) \sim (\square, \triangle, \circ, \times)$).

# A comparison of statistical methods for the analysis of binary repeated measures data with additional hierarchical structure

## 3.1 Abstract

The objective of the study was to compare statistical methods for the analysis of binary repeated measures data with an additional hierarchical level. Such data are commonly encountered in human and veterinary epidemiological research, and one motivating setting for the present study was records of presence or absence of bacteria in milk samples obtained by approximately monthly sampling throughout the lactations of cows in dairy herds. As the basis of a simulation study, random effects true models with autocorrelated ($\rho = 1$, 0.9 or 0.5) subject random effects were used. In general, the settings of the simulation were chosen to reflect

a real somatic cell count dataset, except that the within-subject time series were balanced, complete and of fixed length (4 or 8 time points). Four fixed effects parameters were studied: binary predictors at the subject (e.g., cow) and cluster (e.g., herd) levels, respectively, a linear time effect, and the intercept. Marginal and random effects statistical procedures were considered, and their performance was compared specifically for the four fixed parameters as well as variance and correlation parameters. Among the estimation procedures considered were: ordinary logistic regression (OLR), alternating logistic regression (ALR), generalized estimating equations (GEE), marginal quasi-likelihood (MQL), penalized quasi-likelihood (PQL), pseudo likelihood (REPL), maximum likelihood (ML) estimation and Bayesian Markov chain Monte Carlo (MCMC).

The findings of this study indicate that in data generated by random intercept models ($\rho = 1$), the ML and MCMC procedures performed well and had fairly similar estimation errors. The PQL regression estimates were attenuated while the variance estimates were less accurate than ML and MCMC, but the direction of the bias depended on whether binomial or extra-binomial dispersion was assumed. In datasets with autocorrelation ($\rho < 1$), random effects estimates procedures gave downwards biased estimates, while marginal estimates were little affected by the presence of autocorrelation. The results also indicate that in addition to ALR, a

GEE procedure that accounts for clustering at the highest hierarchical level is sufficient. The REPL procedure performed poorly and produced unsatisfactory estimates regardless of autocorrelation values.

## 3.2 Introduction

Binary repeated measures data with additional hierarchical structure are data with multiple records over time on the same subjects (e.g., animals or farms), which in addition are nested within some (physical) clusters (e.g., hospitals, herds, provinces). In multi-level modelling terminology [31], this may be termed three-level repeated measures data, with observations corresponding to level one and clusters to level three. Such data structures are encountered across a wide range of applications in veterinary and human epidemiology. Our motivating example was records of presence or absence of bacteria in monthly milk samples from cows housed in multiple herds. Thus, the hierarchical structure is the clustering of cows in herds, and the repeated measures are the monthly test records based on the milk samples. Data with this structure are common in studies of dairy cow udder health (e.g., [15, 26]). Some examples from human preventive medicine include the effects of air pollution on school absences in the southern California Children's Health study [35], and the sickness episodes for workers over time [25].

Binary records made on the same subjects, nested within clusters, over time are likely to be correlated [22, 30] or clustered [8, Chapters: 20-21]. A within-subject dependence violates the basic assumption of logistic regression that observations are independent, and may, if not accounted for, lead to biases in parameter estimates and standard errors ([7, Chapter 7] and [9, Chapter 20]). Such data structures challenge the statistical methods to hold its properties, such as asymptotic unbiasedness and nominal confidence interval coverage.

Numerous 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 ([7, Chapters 7-10] and [24]). A large body of literature on statistical methods of binary repeated measures data have discussed the choice between these model types and specific procedures, see for example Diggle *et al.* [7, Chapters: 7-11] or a recent simulation study by Masaoud and Stryhn [20]. However, the added hierarchical structure poses problems for procedures of both types, and to our knowledge a comparison of statistical methods for the analysis of binary repeated measures data with such additional hierarchical structure has not yet been reported. Moreover, the impact of the added hierarchical structure would intuitively be expected to differ not only between estimation approaches but also between types of parameters in a model. The fixed part of a model could contain predictors at all three levels: the cluster level, the subject level, and observation (within subject) level. The random part of a model would involve variances and covariances.

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 were included: ordinary logistic regression (OLR), alternating logistic regression (ALR), generalized estimating equations (GEE), marginal quasi-likelihood (MQL), penalized quasi-likelihood (PQL), pseudo-likelihood (REPL, as implemented in `proc glimmix` in SAS), maximum likelihood via numerical integration (ML) and Bayesian Markov chain Monte Carlo (MCMC).

Analysis of a single dataset by multiple procedures (e.g., [25]) does not necessarily provide much insight into which procedures provide the right answers, and does not cover all aspects of statistical inference. The analytical approach taken for the present study was simulation. Statistical assessments of marginal and random effects procedures for two levels of either binary repeated measures [20] or clustered data [17] are abundant, but these studies do not address the issues related to the additional hierarchical structure.

The objective of the study is to compare marginal and random effects estimation procedures, in terms of statistical properties such as unbiasedness and confidence interval coverage, in a three-level balanced longitudinal design. The comparison includes a range of design parameters at different hierarchical levels. The goal of the comparison is to gain insight into how different estimation approaches deal with the complexity of the design, and to eventually establish some practical guidelines for the choice of statistical procedure for the analysis of balanced, binary repeated measures data with additional hierarchical structure .

## 3.3 Statistical models and estimation procedures

Consider binary records $y_{ijk}$ 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.

### 3.3.1 Random effects models

The general form of a random effects repeated measures model [7, Chapter 11] takes the following form:

$$\text{logit}(\Pr(y_{ijk} = 1 | v_k, u_{ijk})) = \beta_0 + \beta_1 x_{1ijk} + \ldots + \beta_p x_{pijk} + u_{ijk} + v_k, \quad (3.1)$$

where $(v_1, \ldots, v_m)$ are 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 $\sigma_2^2$ represents the heterogeneity (variance) between subjects and $v_k \sim N(0, \sigma_3^2)$ where $\sigma_3^2$ represents the heterogeneity (variance) between clusters. Model (3.1) is for the conditional probability of an "event" given the random effects $v_k$ and $u_{ij}$ of the $k$th cluster and of the $i$th subject at $j$th time point, respectively.

A random intercept model arises as a special case of model (3.1) when $\rho = 1$, i.e., the series $(u_{i1}, \ldots, u_{it})$ of autocorrelated random effects is replaced by the single random effect $u_i$, assumed $\sim N(0, \sigma_2^2)$, for subject $i$ in cluster $k$,

$$\text{logit}(\Pr(y_{ijk} = 1 | v_k, u_i)) = \beta_0 + \beta_1 x_{1ijk} + \ldots + \beta_p x_{pijk} + u_i + v_k, \quad (3.2)$$

with the same assumptions for $(v_k)$ as above, and the same interpretation of $\sigma_2^2$ and $\sigma_3^2$. In our view, model (3.1) forms a better basis for random effects modelling of repeated measures data than the simpler model (3.2) because of its ability to incorporate autocorrelation [7, Chapter 11].

### 3.3.2 Random effects estimation procedures

In general, there is no closed form of the full log likelihood function for models (3.1) and (3.2) and numerical procedures are needed to fit the model. Alternatively several approximation algorithms have been proposed, aimed at producing estimates close to the global ML estimate without actually computing the likelihood function [3]. These algorithms carry a number of different names and acronyms typically involving "weighted least squares" and "quasi"- or "pseudo-likelihood".

Estimation in model (3.2) by numerical approximation most commonly employs the Gauss-Hermite quadrature procedure. Adaptive quadrature [27] 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 (3.2), 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 [7] 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 [4, 20].

Breslow and Clayton [2] presented two estimation procedures based on quasi-likelihood function called penalized quasi-likelihood (PQL) and marginal quasi-likelihood (MQL). The MQL estimates are derived under random effects model assumptions [12]. 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 IGLS 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 [23, Chapter 9] whereas the PQL does. It has been also suggested to refine the approximations by the including a second-order term in the Taylor expansions, usually denoted as second order PQL and MQL procedures [13, 28]. 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., [28, 29]).

In addition, Wolfinger and O'Connell [36] suggested a similar procedure to PQL, called pseudo-likelihood (PL) procedure. It differs from the quasi-likelihood approach by using a true joint likelihood function in its estimation process. Using either ML or REML in the estimation process results in PL or restricted pseudo-likelihood (REPL), respectively. The REPL procedure allows for both random effects in the linear predictor and correlation structure in the observation scale errors conditional (on the mean) [36]. Intuitively, one would expect this procedure to be suitable for models such as model (3.1). Modelling by correlation structure only yields marginal estimates [23, Chapter 22]; adding random effects effectively yields a random effects model with serial correlation [23, Chapter 22].

### 3.3.3 Marginal estimation procedures

The most commonly used procedure to obtain marginal estimates is GEE, generalized estimating equations, which from the onset was devised to deal with repeated measures obtained from multiple subjects [19]. The terms population-averaged and subject-specific inference originate from this context [38]. However, the idea that subjects might be part of a hierarchical structure themselves was not part of the scenario studied. Despite a plethora of extensions of the originally proposed generalized estimating equations [39], to our knowledge no set of estimating equations has been proposed to deal specifically with additional hierarchical structure. Several options can be explored within the classical GEE framework for dealing with one level of clustering of the subjects in addition to the within-subject correlation structure. The simplest idea is perhaps to model clusters by fixed effects (denoted here GEEf) while retaining the usual modelling of within-subject correlation structure. Modelling hierarchical structure by fixed effects has multiple drawbacks, the most important being that it does not allow for inclusion of cluster-level predictors [8, Chapter 20]. An even more crude approach (denoted GEEs), to ignore the additional clustering, would not be expected to yield acceptable cluster-level inference. To achieve correct inference at the cluster level, the GEE handling of correlation structure must be shifted from the subject to the cluster level. This might at first seem to give up on achieving a valid and efficient within-subject inference, but the robustness of GEE procedures to misspecification of working correlation structure should ensure consistency of estimates. The standard choices of GEE working correlation structure do not allow to distinguish between within-cluster and within-subject correlations. In our view the most promising choices of cluster-level correlation structures would be independence (GEEci) and exchangeable (GEEce). Independence correlation structures, effectively corresponding to ordinary logistic regression (OLR) with robust ("sandwich") variance estimates, has been reported to work well for data comprising at least 30 subjects [39].

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 [5]. 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. 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 [20] and to maximum-likelihood estimates in a fully and correctly specified model [7, Chapter 8].

### 3.3.4 Marginal vs. subject-specific estimation

The relation between random effects and marginal estimates has been discussed and described (see e.g., [38, 24] and [7, Chapter 7]) see also Chapter 1. Zeger *et al.* [38] provided the conversion formula for logistic regression with normally distributed random effects:

$$\beta^{PA} \approx (c^2\sigma^2 + 1)^{-1/2}\beta^{SS}, \quad \text{where } c = 16\sqrt{3}/(15\pi) = 0.588. \tag{3.3}$$

For a probit model, the above conversion formula becomes an exact formula (e.g., [21, Chapter 8]):

$$\beta^{PA} = (\sigma^2 + 1)^{-1/2}\beta^{SS}. \tag{3.4}$$

Both formulas can be used to relate subject specific to population average models/estimates under the assumption that random effects are normally distributed. Without any distributional assumptions on the random effects it holds that the marginal regression parameters are attenuated or diluted (towards zero) relative to the random effects parameters, unless the variance is zero [21, Chapter 8].

## 3.4 Simulation study

The settings for the simulation study were motivated by the scc40 dataset of Dohoo *et al.* [8, Chapter 27] for repeated measures of somatic cell counts in milk samples from cows housed in multiple herds. In this observational dataset, up to 11 approximately monthly measures were taken on each cow, but missing values of different types occurred. The impact of missing values will be addressed in a forthcoming study. In order to create settings more akin to experimental studies, we consider here balanced and complete series of either $t = 4$ or $t = 8$ measurements per subject. Thirty clusters were included, with 20 subjects per cluster. In the scc40 context, factors 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:

$$\begin{aligned}
\beta_0 &= -1 && \text{(intercept centered at first time point)},\\
\beta_1 &= 0.15 && \text{(slope for time} = 0, \ldots, t-1),\\
\beta_2 &= -1 && \text{(coefficient for subject level covariate)},\\
\beta_3 &= 1 && \text{(coefficient for cluster level covariate)}.
\end{aligned}$$

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 ssc40 dataset obtained by dichotomizing the somatic cell counts at 200 000 cells/ml. High somatic cell counts are considered an indicator of subclinical mastitis. By the latent variable approximation to the variance partition coefficient [14], 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 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 $\pi^2/3$ [31, 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 [6]). Generation of the binary outcomes then followed the usual scheme for random effects logistic regression models [32].

### 3.4.1 Software and settings for estimation procedures

The GEE estimation procedure used the implementation of `proc genmod` in SAS software version 9.1, with two working correlation structures: autoregressive, and exchangeable. The ALR estimation procedure used the implementation in SAS with two pairwise odds ratios: within subcluster (subject) odds ratio and within cluster odds ratio. The random effects procedures used the first order MQL, MQLx and second order PQL, PQLx procedures, with REML option and implemented in the MLwiN software (version 2.02), the REPL procedure of SAS (`proc glimmix`), as well as the adaptive quadrature algorithms for ML estimation implemented in Stata version 10 software (`xtmelogit` command with 7 quadrature points at both the subject and cluster level). The REPL procedure was set up with cluster and subject random effects and a first order autoregressive correlation structure, and the REML option [36]. MQLx, PQLx and REPL estimation procedures included an additional overdispersion parameter. The Bayesian estimation procedures were implemented in WinBUGS version 1.4 called from the R software using the R2WinBUGS package [33]. Vague ("non-informative") prior distributions (i.e. $\mathrm{N}(0, 10^6)$) were used for all fixed effects parameters. The recently recommended uniform distribution for inverse variances, or precisions ($\tau \sim$ uniform(0,100)) was used [18, 11]. The Markov chains were run with 500 burn-in samples [4], 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.

### 3.4.2 Analysis of results for simulated data

The estimates of marginal estimation procedures can be compared either to true subject-specific parameter values obtained from the conversion formula (3.3) with $\sigma^2 = \sigma_2^2 + \sigma_3^2$. Inserting the known true variance parameters, yields $\beta^{PA} \approx 0.712 \times \beta^{SS}$. Although the marginal parameters are "theoretical" in the sense that they can only be constructed from the (unknown) variance parameters, they were used to assess the marginal estimation procedures against their expected values. Unless all variances are small, there is little prospect in using marginal estimation procedures to reconstruct the true parameters of random effects models (see [20] for discussion of the choice between random effects and marginal estimation procedures). As the fixed effects parameters are on different scales, the results were presented in terms of the relative bias defined as the difference between the average estimate among simulations ($\hat{\beta}$) and the true value (marginal or subject-specific)($\beta$) divided by the true value,

$$\text{relative bias } = \frac{\hat{\beta} - \beta}{\beta} \times 100\% \tag{3.5}$$

The presence of statistically significant bias in the estimates (of both fixed effects and variance parameters) was assessed by a $z$-test based on the true value and the standard deviation among simulations. The statistical significance of bias in the standard errors was assessed by comparing the mean standard error to a 95% confidence interval for the standard deviation based on the simulations. This simple procedure was considered acceptable because the statistical variation in the estimated standard deviation was generally much larger than that of the mean standard error. Confidence intervals (CIs) were computed by the large-sample normal approximation based on the standard error; for the GEE procedures, the robust standard error was used. The coverage of 95% CIs was computed as the proportion of simulated datasets for which the confidence interval (in the Bayesian analysis: the credibility interval) contained the true parameter.

## 3.5 Results

Presentation of results is separated by the type of estimation procedure: based on either marginal or random effects models. Relative biases of estimates and standard errors are shown in Tables 3.1–3.2, coverages of confidence intervals are shown in Figures 3.1–3.2. Table 3.3 gives relative biases of estimates and standard errors for datasets and analyses based on the probit link function.

### 3.5.1 Random effects estimation procedures

As all random effects estimation procedures except REPL are based on a random intercept model, the performance of the procedures in datasets corresponding to this true model ($\rho = 1$) is reviewed first, and subsequently we turn to the results for autoregressive datasets ($\rho < 1$).

#### 3.5.1.1 Variance parameters

The two likelihood-based procedures (ML and MCMC) produced fairly accurate variance estimates and standard errors (Table 3.1). ML estimates of the level 3 variances were slightly attenuated (biased towards zero) and the MCMC standard errors for the same parameter were somewhat inflated (biased away from zero). PQL variance estimates were less accurate, but the direction of the bias depended on whether binomial or extra-binomial dispersion was assumed. In agreement with previous findings in Chapter 2, PQL showed attenuated estimates, whereas PQLx showed both downward and upward biases ($\sigma_3^2$ for $t = 8$, and $\sigma_2^2$ for $t = 4$, respectively). The extra-binomial parameter estimates were centered around 0.80 with strongly inflated standard errors. Consequently, PQLx variance estimates were generally higher than PQL estimates. All variance estimates of the REPL procedure were strongly attenuated.

Due to the scaling by the variance parameters inherent in random effects estimation procedures, any bias in estimated variances is likely to affect the fixed effects as well (and in the same direction), in particular for moderate to large variance components. For example, all fixed effects estimates for REPL indeed showed substantial bias (range 6–16%) towards zero. Also, the attenuation of both fixed effects and variance parameters for PQL estimation was more pronounced for shorter time series, corresponding to less replication at the subject level [20].

#### 3.5.1.2 Level 3 parameters

A similar qualitative behaviour was expected for the intercept and the predictor at the highest (cluster) level, and generally the results confirmed this. Likelihood-based procedures gave unbiased estimates but in some cases slightly underestimated the standard error. PQL procedures showed the same bias in the standard error, and some instances of minor negative (PQL) or positive (PQLx) bias in the estimates. As noted above, REPL estimates were clearly biased towards zero. CI coverage (Figure 3.1) was close to or slightly below nominal for all procedures except REPL.

#### 3.5.1.3 Level 1 and 2 parameters

The subject-level parameter was estimated without any biases for ML, MCMC and PQLx; a small negative bias (3–5%) was present in PQL estimates. The regression coefficient for time (level 1) was estimated without bias by ML, MCMC and PQL (except for $t = 4$), whereas the PQLx estimates were moderately inflated, had too small standard errors and showed undercoverage of CIs.

#### 3.5.1.4 Autocorrelated data ($\rho < 1$)

For all random effects procedures, estimates of both fixed effects and variance parameters were attenuated in autoregressive datasets (Table 3.1). In the vast majority of settings, the biases were statistically significant. The relative bias was strongest for the variance parameters, increased markedly from $\rho = 0.9$ to $\rho = 0.5$, and was also somewhat larger for the long time series ($t = 8$). Standard errors were also clearly underestimated (up to 18%) in several cases, including the variance parameters and in the long series also the time effect. The bias in standard errors was less severe for the short series ($t = 4$). Contrary to the other procedures, MCMC estimation produced inflated (up to 19%) estimates for $\sigma_3^2$. Generally, the setting least affected was $(\rho, t) = (0.9, 4)$ where biases for fixed parameter estimates of ML, MCMC and PQL were below

10%, and CI coverage was above 90%. In other settings, the biases for these procedures ranged up to 24% for fixed effects and 88% for variance parameters, and CI coverage could go below 50% (for $(\rho, t) = (0.5, 8)$).

Adding the extra-binomial dispersion parameter to the PQL procedure did not alleviate the attenuation of PQL estimates substantially. For low autocorrelation ($\rho = 0.5$), both PQL procedures and the likelihood-based procedures showed downward biases and undercoverages of similar magnitudes. However, the extra-dispersion parameter tended to increase in value (show less bias) for decreasing $\rho$. Although the only random effects procedure examined which incorporated an autoregressive parameter, the REPL procedure was equally affected by the autocorrelation as the other procedures. Moreover, for both values of $\rho$ the REPL estimates had stronger bias and lower CI coverage than those of the other procedures.

### 3.5.2 Marginal estimation procedures

Generally, estimates from OLR, ALR, MQL, and the GEE procedures except GEEf agreed closely (Table 3.2), and all showed a small negative relative bias in the range 3–6%. For estimation of a marginal parameter, all the methods are asympotically consistent, and it is plausible that this bias is due to approximation error in the calculation of the marginal true value; see Section 3.5.2.3 below. As the asymptotic consistency does not rely on assumptions about the true correlation structure, the presence of autocorrelation ($\rho < 1$) in the data was expected to have less of an impact for marginal than random effects procedures.

#### 3.5.2.1 Level 3 parameters

The fixed cluster effects included in the linear predictor for GEEf precluded estimation of effects at the cluster level. For the other procedures, the relative bias remained relatively constant in the 3–6% range irrespective of the autocorrelation ($\rho$); thus, the differences in performance were essentially in the standard errors. As expected, OLR grossly underestimated the standard errors at the cluster level. Strongly underestimated standard errors were also seen for the GEEs procedure, a consequence of the lack of cluster level effects in the (variance) estimating equations. The other procedures generally showed a minor (up to 7.5%) downward relative bias in the standard errors, and CI coverage at or moderately below the nominal level (Figure 3.2), irrespective of the value of $\rho$.

#### 3.5.2.2 Level 1 and 2 parameters

Estimates from GEEf showed a small upwards bias (less than 5%), in the opposite direction of the downwards bias displayed by all other procedures. Standard errors were generally, except for some settings for the time parameter, unbiased for ALR and all GEE procedures excluding GEEs, which had substantial upwards bias for the level 2 coefficient. Both the MQL and MQLx procedures showed biases in either direction for the time parameter. Stronger biases were noted for OLR, also in both directions. CI coverages varied around the nominal level for all procedures except OLR.

#### 3.5.2.3 Overall marginal bias

We examined the small negative bias experienced by almost all marginal estimation procedures by rerunning the simulation study using a probit model both to generate the data and fit the models involved in the estimation procedures. For a probit model, the conversion formula (3.3) becomes an exact formula (3.4) [21, Chapter 8]. Table 3.3 gives the relative bias for a subset of the previously considered estimation procedures. The OLR, GEEce, GEEci and ALR procedures gave virtually unbiased estimates, although some bias in the standard errors remained (similar to the previously discussed results in Table 3.2). As the exact conversion formula thus produced unbiased estimates, we consider the approximation formula as a plausible source of the general downward bias seen in Table 3.2. The results for the probit link also confirmed the suspected positive bias in GEEf estimates. As before, the autocorrelation had only little impact on the performance of the estimation procedures.

## 3.6 Discussion

### 3.6.1 Random effects estimation procedures

In data generated by random intercept models ($\rho = 1$), the ML and MCMC procedures performed well and had fairly similar estimation errors. Generally, the estimation bias in PQL is known [21, Chapter 10] and could be partly due to the well-known attenuation of variance parameters by PQL in certain settings [10]. However, the results (especially when $\phi = 1$) were in accordance with the those of Molenberghs and Verbeke [23, Chapter 14], who reported that the performance of PQL could be improved by increasing the number of subjects. Venables and Ripley [34] concluded that allowing for $\phi$ in certain applications, yielded regression estimates that were closer to the maximum likelihood estimates, a finding that could not be reproduced in the current study settings. Our results showed an evidence of inflated standard errors for $\phi$, which does not support the suggestion by Yang *et al.* [37] of using $\phi$ as a diagnostic tool, and by Barbosa and Goldstein [1] to "allow for model misspecification". Recently, Heo and Leon [17] concluded that the full likelihood approach "appears to be preferable for the analysis of clustered binary observations with underlying random effects models".

In datasets with autocorrelation ($\rho < 1$), the downward biased estimates of random effects can be seen largely as a scaling effect caused by the underestimation of the random effect variances. Inference for all fixed effects parameters was strongly affected, and the biases of estimates were of similar magnitude. The associated pattern of increases in the inflated standard errors for $\phi$ with decreases in the correlation value, could be due to the increase of the variability within each subject as the correlation decreases. This finding may raise some concern about the approximation procedures to the log likelihood. However, more research is needed to confirm it, especially in simpler settings such as correlated binary data.

The REPL estimation procedure performed poorly and produced unsatisfactory estimates regardless of autocorrelation ($\rho = 0.5, 0.9, 1$). These findings are in support of those by Evans *et al.* [10] for variance component and by [20] for regression estimates. This could be due to the inclusion of both random effects and a correlation structure in the REPL procedure, and thus modelling parts of the variance/correlation structure on different scales [23, Chapter 22]. The ability of the REPL procedure to incorporate autocorrelation did not render the estimates less susceptible to attenuation bias in datsets with autocorrelation than the other random effects procedures.

### 3.6.2 Marginal estimation procedures

The results suggest that the small general bias observed for marginal estimation procedures could be due to the conversion formula (3.3). However, to our knowledge, the assessment of the accuracy of this formula in practice has not been reported yet. This conversion formula relies on the variance of normally distribute random effects to scale subject specific to population average models/estimates. The avoidance of such scaling problems by separating fixed and random effects estimates was one of the key ideas behind the development of marginalized models [16]. The application of marginalized models to repeated measures data with additional hierarchical structure has to our knowledge not been reported.

Marginal estimates were little affected by the presence of autocorrelation; similar performances of the different procedures were seen in all settings ($\rho = 0.5, 0.9, 1$). ALR, GEEci, and GEEce performed fairly well with only a few instances of minor statistically significant bias. MQL (apart form the fluctuation in the the standard error for time coefficient) performed on a par with ALR, and in agreement with previous finding in [20]. The results indicate that accounting for clustering at the highest hierarchical level is sufficient (GEEce, GEEci). In fact, standard errors for these procedures and ALR showed a larger bias at the cluster level than at levels below, probably an effect of of the low number of clusters [39]. This is an interesting result that we think needs more research to validate it in different settings and designs of binary data. Accounting for the additional hierarchical structure by fixed effects (GEEf) resulted in biased estimates. However, ignoring the hierarchical structure in the data in (OLR and GEEs) resulted in inflated and biased standard error of the cluster level fixed effects, this is in agreement with those reported by Diggle *et al* [7, Chapter 7] and Dohoo *et al*[9]. Allowing for $\phi$ in MQL had almost no impact on the regression estimates of fixed effects. An explanation could be that MQL do not incorporate the random effects $u_i$ in the linearization of the mean [23, Chapter 14].

### 3.6.3 Recommendations

We conclude with a discussion of the implications of the current findings for the choice of procedure. For the choice between marginal and random effects approaches, this study adds only little to existing knowledge [20]. The bias seen in marginal estimates should be of no concern for the use of marginal procedures if the interpretation in this study is correct that it is caused by the conversion formula. For autoregressive data, the random effects procedures performed poorly (as was found also in Chapter 2), therefore marginal procedures may seem more attractive, unless the time series is very short (less than 4 time points).

Generally, the likelihood-based random effects procedures (ML, MCMC) performed better than methods based on quasi-or pseudo-likelihood. The inclusion of an overdispersion parameter in the latter methods did not clearly improve their performance. As is well-documented, biases of these methods in the absence of an overdispersion parameter are towards the null [28, 29], whereas in the presence of overdispersion biases tended to be less predictable (in either direction). Moreover, we are not convinced about the usefulness of $\phi$ as a diagnostic tool. The REPL procedure performed poorly in our settings, substantiating the finding reported in by Masaoud and Stryhn [20] that REPL performs mostly as a marginal estimation procedure with no promise for estimation of the variance or an autoregressive parameter. Further research may be needed to assess its accuracy and validity for binary repeated measures data.

Among the marginal procedures, ALR and GEE with either independence or exchangeable correlation at the cluster-level performed similarly and generally well across the range of settings covered. All other attempts to incorporate the additional hierarchical level into the GEE framework produced estimates with serious deficiencies for some of the fixed effects parameters. In situations where the affected parameters are of no interest (or absent), such other schemes may be acceptable but on the other hand show no advantages over the above-mentioned generally acceptable schemes.

## 3.7 References

## References

[1] Barbosa, B., Goldstein, H., 2000. Discrete multilevel response models. *Quality and Quantity* **34**, 323–330.

[2] Breslow, N. E., Clayton, D. G., 1993. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.

[3] Breslow, N. E., 2003. Whither PQL?. University of Washington Biostatistics Working Paper Series **192**.

[4] Browne, W. J., Draper, D., 2006. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis* **3**, 473–514.

[5] Carey, V., Zeger, S. L., Diggle, P., 1993. Modeling multivariate binary data with alternating logistic regressions. *Biometrika* **80**, 517–526.

[6] Congdon, P., 2003. *Applied Bayesian Modelling.* Wiley, New York.

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

[8] Dohoo, I. R., Martin, S. W., Stryhn, H., 2003. *Veterinary Epidemiologic Research.* AVC Inc., Charlottetown, Canada; web-site: `http://www.upei.ca/ver`.

[9] Dohoo, I. R., Stryhn, H., 2006. Simulation studies on the effects of clustering. XI*th* International Conference of Veterinary Epidemiology and Economics, Cairns, Australia, August 2006.

[10] Evans, B. A., Feng, Z., Peterson, A. V., 2001. A comparison of generalized linear mixed model procedures with estimating equations for variance and covariance parameter estimation in longitudinal studies and group randomized trials. *Statistics in Medicine* **20**, 3353–3373.

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

[12] Goldstein, H., 1991. Nonlinear multilevel models with an application to discrete response data. *Biometrika* **78**, 45–51.

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

[14] Goldstein, H., Browne, W. J., Rasbash, J., 2002. Partitioning variation in multilevel models. *Understanding Statistics*, **1**, 223–231.

[15] Green, M. J., Burton, P. R., Green, L. E., Schukken, Y. H., Bradley, A. J., Peeler, E. J., Medley, G. F., 2004. The use of Markov Chain Monte Carlo for analysis of correlated binary data: Patterns of somatic cells in milk and the risk of clinical mastitis in dairy cows. *Preventive Veterinary Medicine* **64**, 157–174.

[16] Heagerty, P. J., Zeger, S. L., 2000. Marginalized multilevel models and likelihood inference. *Statistical Science* **15**, 1–19.

[17] Heo, M., Leon, A. C., 2005. Comparison of statistical methods for the analysis of clustered binary observations. *Statistics in Medicine* **24**, 911–923.

[18] Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R., 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.

[19] Liang, K. Y., Zeger, S. L., 1986, Longitudinal Data-Analysis Using Generalized Linear-Models. *Biometrika* **73**, 13–22.

[20] Masaoud, E., Stryhn, H., 2008. A simulation study to assess statistical methods for binary repeated measures data. Submitted manuscript.

[21] McCulloch, C. E., Searle, S. R., 2001. *Generalized Linear and Mixed Models.* Wiley, New York.

[22] McDermott, J. J., Schukken, Y. H., Shoukri, M. M., 1994. Methods for analysing data collected from clusters of animals. *Preventive Veterinary Medicine* **18**, 175–192.

[23] Molenberghs, G., Verbeke, G., 2005. *Models for Discrete Longitudinal Data.* Springer, New York.

[24] Neuhaus, J. M., 1992. Statistical methods for longitudinal and clustered design with binary responses. *Statistical Methods in Medical Research* **1**, 249–273.

[25] Preisser, J. S., Arcury, T. A., Quandt, S. A., 2003. Detecting patterns of occupational illness clustering with alternating logistic regressions applied to longitudinal data. *American Journal of Epidemiology.* **158**, 495–501.

[26] Olde Riekerink, R. G. M., Barkema, H. W., Stryhn, H., 2007. The effect of season on somatic cell count and the incidence of clinical mastitis. *Journal of Dairy Science* **90**, 1704–1715.

[27] Rabe-Hasketh, S., Skrondal, A., Pickles, A., 2002. Reliable estimation of generalised linear mixed models using adaptive quadrature. *The Stata Journal* **2**, 1–21.

[28] Rodríguez, 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.

[29] Rodríguez, 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.

[30] Schukken, Y. H., Grohn,Y. T., McDermott B., McDermott J. J., 2003. Analysis of correlated discrete observations: background, examples and solutions. *Preventive Veterinary Medicine* **59**, 223–40.

[31] Snijders, T. A. B., Bosker, R. J., 1999. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modelling.* Sage Publishers, London.

[32] Stryhn, H., Dohoo, I. R., Tillard, E., 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, August 2000.

[33] Sturtz, S., Ligges, U., Gelman, A., 2005. R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software* **12**, 1–16.

[34] Venables, W. N., Ripley, B. D., 2002. *Modern Applied Statistics with S (4th ed.)*. New York, Springer.

[35] Virginie, R., Kiros, B., Duncan, C. T., 2005. A three-level model for binary time-series data: the effects of air pollution on school absences in the Southern California Children's Health Study. *Statistics in Medicine* **24**, 1103–1115.

[36] Wolfinger, R., O'Connell, M., 1993. Generalized linear mixed models: a pseudo-likelihood approach. *Communications in Statistics: Simulation and Computation* **48**, 233–243.

[37] Yang, M., Goldstein, H., Heath, A., 2000. Multilevel models for repeated binary outcomes: attitudes and voting over the electoral cycle. *Journal of the Royal Statistical Society, Series A* **163**, 49–62.

[38] Zeger, S. L., Liang, K. Y., Albert, P. S., 1988. Models for longitudinal data - a generalized estimating equation approach. *Biometrics* **44**, 1049–1060.

[39] Ziegler, A., Kastner, C., Blettner, M. 1998. The generalized estimating equations: an annotated bibliography. *Biometrical Journal* **40**, 115–139.

Table 3.1: Relative bias of estimates and associated standard errors of fixed effects and variance parameters obtained by five random effects estimation procedures, based on analysis of 1000 simulated datasets per setting ($t$ = number of time points, $\rho$ = autocorrelation, parameters:$\beta_0$ (intercept), $\beta_1$ (time coefficient), $\beta_2$ (subject level factor), $\beta_3$ (cluster level factor),$\sigma_2^2$ (variance at subject level), $\sigma_3^2$ (variance at cluster level), $\phi$ (extra binomial dispersion)). Coding for estimation procedures: PQL (2nd order penalized quasi-likelihood), PQLx (2nd order penalized quasi-likelihood with extra binomial dispersion), REPL (restricted pseudo-likelihood), ML (maximum likelihood), MCMC (Bayesian Markov chain Monte Carlo).

| | | | Estimation procedures | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | PQL | | PQLx | | REPL | | ML | | MCMC | |
| t | $\rho$ | Par. | Est. | SE | Est. | SE | Est. | SE | Est. | SE | Est. | SE |
| 8 | 1 | $\beta_0$ | $0.0$ | $-5.3^*$ | $2.7^\ddagger$ | $-5.9^*$ | $-8.5^\ddagger$ | $-2.6$ | $0.9$ | $-4.9$ | $0.6$ | $-5.5^*$ |
| | | $\beta_1$ | $0.9^\dagger$ | $-2.4^*$ | $2.7^\ddagger$ | $-11.8^*$ | $-6.4^\ddagger$ | $-8.6^*$ | $0.7^\dagger$ | $-0.2$ | $0.9^\dagger$ | $-0.6$ |
| | | $\beta_2$ | $-3.0^\ddagger$ | $1.1$ | $-0.6$ | $1.3$ | $-11.4^\ddagger$ | $0.2$ | $0.0$ | $1.0$ | $0.2$ | $0.7$ |
| | | $\beta_3$ | $-1.9$ | $-5.8^*$ | $0.6$ | $-5.8^*$ | $-11.0^\ddagger$ | $-2.3$ | $1.0$ | $-5.7^*$ | $1.0$ | $-4.8^*$ |
| | | $\sigma_2^2$ | $-8.6^\ddagger$ | $-8.9^*$ | $2.7^\ddagger$ | $-14.7^*$ | $-18.8^\ddagger$ | $-6.4^*$ | $-0.1$ | $0.3$ | $1.0^\ddagger$ | $0.0$ |
| | | $\sigma_3^2$ | $-11.5^\ddagger$ | $-2.1$ | $-7.0^\ddagger$ | $-2.2$ | $-19.6^\ddagger$ | $2.1$ | $-6.2^\ddagger$ | $-0.8$ | $3.4^\ddagger$ | $12.7^*$ |
| | | $\phi$ | | | $-16.2^\ddagger$ | $44.9^*$ | $-21.4^\ddagger$ | $20.9^*$ | | | | |
| | .9 | $\beta_0$ | $-9.4^\ddagger$ | $-5.8^*$ | $-7.6^\ddagger$ | $-6.3^*$ | $-18.6^\ddagger$ | $-2.6$ | $-9.5^\ddagger$ | $-5.2^*$ | $-9.2^\ddagger$ | $-4.7^*$ |
| | | $\beta_1$ | $-8.6^\ddagger$ | $-14.2^*$ | $-6.7^\ddagger$ | $-20.2^*$ | $-17.3^\ddagger$ | $-11.4^*$ | $-9.1^\ddagger$ | $-11.7^*$ | $-8.9^\ddagger$ | $-12.1^*$ |
| | | $\beta_2$ | $-10.4^\ddagger$ | $-2.2$ | $-8.7^\ddagger$ | $-1.8$ | $-19.7^\ddagger$ | $-1.1$ | $-9.4^\ddagger$ | $-0.5$ | $-9.3^\ddagger$ | $-0.3$ |
| | | $\beta_3$ | $-9.7^\ddagger$ | $-6.0^*$ | $-7.9^\ddagger$ | $-6.0^*$ | $-19.3^\ddagger$ | $-2.8$ | $-8.8^\ddagger$ | $-5.6^*$ | $-8.4^\ddagger$ | $-4.0^*$ |
| | | $\sigma_2^2$ | $-43.1^\ddagger$ | $-10.9^*$ | $-36.2^\ddagger$ | $-15.4^*$ | $-54.6^\ddagger$ | $-2.8$ | $-39.6^\ddagger$ | $2.5$ | $-39.0^\ddagger$ | $2.9$ |
| | | $\sigma_3^2$ | $-26.1^\ddagger$ | $-1.2$ | $-23.1^\ddagger$ | $-1.1$ | $-35.7^\ddagger$ | $4.6$ | $-24.7^\ddagger$ | $1.4$ | $-16.9^\ddagger$ | $16.2^*$ |
| | | $\phi$ | | | $-12.1^\ddagger$ | $108.8^*$ | $-15.8^\ddagger$ | $60.3$ | | | | |
| | .5 | $\beta_0$ | $-22.6^\ddagger$ | $-8.6^*$ | $-21.9^\ddagger$ | $-9.0^*$ | $-28.3^\ddagger$ | $-4.6^*$ | $-22.9^\ddagger$ | $-7.8^*$ | $-23.1^\ddagger$ | $-6.2^*$ |
| | | $\beta_1$ | $-22.6^\ddagger$ | $-15.0^*$ | $-22.0^\ddagger$ | $-17.8^*$ | $-27.3^\ddagger$ | $-9.0^*$ | $-22.9^\ddagger$ | $-12.9^*$ | $-23.0^\ddagger$ | $-13.5^*$ |
| | | $\beta_2$ | $-23.6^\ddagger$ | $-3.0$ | $-23.0^\ddagger$ | $-3.1$ | $-28.4^\ddagger$ | $-0.5$ | $-23.9^\ddagger$ | $-0.4$ | $-23.8^\ddagger$ | $-0.6$ |
| | | $\beta_3$ | $-22.9^\ddagger$ | $-7.8^*$ | $-22.2^\ddagger$ | $-7.9^*$ | $-27.8^\ddagger$ | $-4.0$ | $-23.1^\ddagger$ | $-7.4^*$ | $-23.1^\ddagger$ | $-5.4^*$ |
| | | $\sigma_2^2$ | $-88.3^\ddagger$ | $-11.6^*$ | $-86.4^\ddagger$ | $-17.1^*$ | $-93.9^\ddagger$ | $-7.6^*$ | $-87.4^\ddagger$ | $-1.7$ | $-87.6^\ddagger$ | $-5.6^*$ |
| | | $\sigma_3^2$ | $-45.5^\ddagger$ | $0.3$ | $-44.6^\ddagger$ | $0.1$ | $-47.6^\ddagger$ | $5.0^*$ | $-45.9^\ddagger$ | $3.2$ | $-40.4^\ddagger$ | $19.3^*$ |
| | | $\phi$ | | | $-5.4^\ddagger$ | $187.3^*$ | $-4.4^\ddagger$ | $61.9^*$ | | | | |
| 4 | 1 | $\beta_0$ | $-4.5^\ddagger$ | $-5.1^*$ | $2.1^\dagger$ | $-6.0^*$ | $-14.8^\ddagger$ | $-4.2^*$ | $-1.0$ | $-3.6$ | $-0.8$ | $-2.8$ |
| | | $\beta_1$ | $-2.5^\ddagger$ | $0.5$ | $5.8^\ddagger$ | $-14.9^*$ | $-10.9^\ddagger$ | $-8.2^*$ | $-0.5$ | $2.7$ | $-0.2$ | $2.3$ |
| | | $\beta_2$ | $-4.7^\ddagger$ | $-0.3$ | $1.1^\dagger$ | $1.6$ | $-15.5^\ddagger$ | $-1.1$ | $-0.2$ | $1.1$ | $0.2$ | $1.2$ |
| | | $\beta_3$ | $-5.4^\ddagger$ | $-5.6^*$ | $0.4$ | $-5.2^*$ | $-16.3^\ddagger$ | $-3.1$ | $-1.2$ | $-4.6^*$ | $-0.9$ | $-2.5$ |
| | | $\sigma_2^2$ | $-14.2^\ddagger$ | $-13.8^*$ | $17.9^\ddagger$ | $-23.3^*$ | $-22.2^\ddagger$ | $-12.2^*$ | $1.2^\ddagger$ | $2.7$ | $3.1^\ddagger$ | $2.7$ |
| | | $\sigma_3^2$ | $-13.8^\ddagger$ | $-1.6$ | $-2.5^\dagger$ | $-1.1$ | $-25.9^\ddagger$ | $1.9$ | $-5.9^\ddagger$ | $0.9$ | $3.9^\ddagger$ | $14.1^*$ |
| | | $\phi$ | | | $-23.0^\ddagger$ | $19.4^*$ | $-27.7^\ddagger$ | $11.8^*$ | | | | |
| | .9 | $\beta_0$ | $-8.8^\ddagger$ | $-5.1^*$ | $-3.2^\ddagger$ | $-6.3^*$ | $-20.0^\ddagger$ | $-3.2$ | $-6.8^\ddagger$ | $-3.5$ | $-6.6^\ddagger$ | $-1.4$ |
| | | $\beta_1$ | $-6.7^\ddagger$ | $-6.9^*$ | $-0.1$ | $-18.9^*$ | $-16.4^\ddagger$ | $-10.1^*$ | $-5.6^\ddagger$ | $-4.2$ | $-5.1^\ddagger$ | $-4.8^*$ |
| | | $\beta_2$ | $-8.7^\ddagger$ | $-1.7$ | $-3.5^\ddagger$ | $-0.4$ | $-20.0^\ddagger$ | $-0.6$ | $-6.0^\ddagger$ | $1.3$ | $-5.7^\ddagger$ | $1.3$ |
| | | $\beta_3$ | $-9.3^\ddagger$ | $-6.2^*$ | $-4.1^\ddagger$ | $-6.2^*$ | $-21.0^\ddagger$ | $-2.9$ | $-6.8^\ddagger$ | $-5.3^*$ | $-6.4^\ddagger$ | $-2.9$ |
| | | $\sigma_2^2$ | $-35.5^\ddagger$ | $-18.0^*$ | $-10.7^\ddagger$ | $-28.4^*$ | $-45.3^\ddagger$ | $-11.7^*$ | $-25.4^\ddagger$ | $1.7$ | $-24.1^\ddagger$ | $0.5$ |
| | | $\sigma_3^2$ | $-21.9^\ddagger$ | $-4.7^*$ | $-12.6^\ddagger$ | $-4.8^*$ | $-34.6^\ddagger$ | $0.0$ | $-17.6^\ddagger$ | $-1.1$ | $-9.2^\ddagger$ | $13.1^*$ |
| | | $\phi$ | | | $-19.5^\ddagger$ | $37.3^*$ | $-22.7^\ddagger$ | $20.9^*$ | | | | |
| | .5 | $\beta_0$ | $-21.7^\ddagger$ | $-5.9^*$ | $-18.8^\ddagger$ | $-7.2^*$ | $-27.8^\ddagger$ | $-1.8$ | $-21.4^\ddagger$ | $-4.3^*$ | $-21.5^\ddagger$ | $-1.4$ |
| | | $\beta_1$ | $-20.5^\ddagger$ | $-7.5^*$ | $-17.5^\ddagger$ | $-14.4^*$ | $-26.4^\ddagger$ | $-4.6^*$ | $-20.3^\ddagger$ | $-5.0^*$ | $-20.1^\ddagger$ | $-5.2$ |
| | | $\beta_2$ | $-20.3^\ddagger$ | $-1.8$ | $-17.4^\ddagger$ | $-2.6$ | $-26.9^\ddagger$ | $1.6$ | $-19.8^\ddagger$ | $3.7$ | $-19.7^\ddagger$ | $3.0$ |
| | | $\beta_3$ | $-21.6^\ddagger$ | $-5.6^*$ | $-18.7^\ddagger$ | $-6.0^*$ | $-28.1^\ddagger$ | $-0.9$ | $-21.2^\ddagger$ | $-4.6^*$ | $-21.3^\ddagger$ | $-0.6$ |
| | | $\sigma_2^2$ | $-80.8^\ddagger$ | $-15.7^*$ | $-71.6^\ddagger$ | $-31.0^*$ | $-85.7^\ddagger$ | $-2.7$ | $-77.1^\ddagger$ | $0.7$ | $-77.4^\ddagger$ | $-6.1^*$ |
| | | $\sigma_3^2$ | $-41.3^\ddagger$ | $-5.5^*$ | $-36.9^\ddagger$ | $-6.7^*$ | $-45.8^\ddagger$ | $3.6$ | $-40.6^\ddagger$ | $-0.6$ | $-34.5^\ddagger$ | $13.8^*$ |
| | | $\phi$ | | | $-10.4^\ddagger$ | $87.4^*$ | $-9.7^\ddagger$ | $25.9^*$ | | | | |

$^\dagger$ significant bias in estimate at $P < 0.05$; $^\ddagger$ significant bias in estimate at $P < 0.01$;
$^*$ significant bias in standard error at $P < 0.05$

Table 3.2: Relative bias of fixed effects parameter estimates (against marginal true values) and their standard errors obtained by eight marginal estimation procedures, based on analysis of 1000 simulated datasets per setting ($t$ = number of time points, $\rho$ = autocorrelation, parameters:( $\beta_0$ (intercept), $\beta_1$ (time coefficient), $\beta_2$ (subject level factor), $\beta_3$ (cluster level factor) )). Coding for estimation procedures: OLR (ordinary logistic regression), GEEci (generalized estimating equations (GEE) with independence correlation at cluster level), GEEce (GEE with exchangeable correlation at cluster level), GEEf (GEE with fixed effects for cluster level and autoregressive correlation at subject level), GEEs (GEE with autoregressive correlation at subject level), ALR (alternating logistic regression), MQL (marginal quasi-likelihood), MQLx (marginal quasi-likelihood with extra binomial dispersion).

| | | | Marginal estimation procedures | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | OLR | | GEEci | GEEce | | GEEf | | GEEs | | ALR | | MQL | | MQLx | |
| t | $\rho$ | Par. | Est. | SE | SE | Est. | SE | Est. | SE | Est. | SE | Est. | SE | Est. | SE | Est. | SE |
| 8 | 1 | $\beta_0$ | $-3.5^{\ddagger}$ | $-58.5^{*}$ | $-6.4^{*}$ | $-4.0^{\ddagger}$ | $-6.5^{*}$ | n/a | n/a | $-3.7^{\ddagger}$ | $-38.8^{*}$ | $-3.6^{\ddagger}$ | $-6.3^{*}$ | $-3.5^{\ddagger}$ | $-5.5^{*}$ | $-3.5^{\ddagger}$ | $-4.3^{*}$ |
| | | $\beta_1$ | $-3.9^{\ddagger}$ | $18.8^{*}$ | $-2.7$ | $-3.9^{\ddagger}$ | $-2.8$ | $4.2^{\ddagger}$ | $-3.3$ | $-3.5^{\ddagger}$ | $-3.3$ | $-3.9^{\ddagger}$ | $-2.8$ | $-3.9^{\ddagger}$ | $-1.7$ | $-3.9^{\ddagger}$ | $18.8^{*}$ |
| | | $\beta_2$ | $-4.7^{\ddagger}$ | $-38.7^{*}$ | $0.0$ | $-4.7^{\ddagger}$ | $0.0$ | $3.1^{\ddagger}$ | $1.6$ | $-4.5^{\ddagger}$ | $12.0^{*}$ | $-4.8^{\ddagger}$ | $0.0$ | $-4.7^{\ddagger}$ | $1.8$ | $-4.7^{\ddagger}$ | $1.8^{*}$ |
| | | $\beta_3$ | $-3.6^{\ddagger}$ | $-72.3^{*}$ | $-5.5^{*}$ | $-3.9^{\ddagger}$ | $-5.3^{*}$ | n/a | n/a | $-3.0^{\ddagger}$ | $-50.6^{*}$ | $-3.6^{\ddagger}$ | $-5.7^{*}$ | $-3.5^{\ddagger}$ | $-5.8^{*}$ | $-3.5^{\ddagger}$ | $-5.8^{*}$ |
| | .9 | $\beta_0$ | $-4.2^{\ddagger}$ | $-57.0^{*}$ | $-5.5^{*}$ | $-4.5^{\ddagger}$ | $-4.9^{*}$ | n/a | n/a | $-4.2^{\ddagger}$ | $-40.3^{*}$ | $-4.2^{\ddagger}$ | $-5.4^{*}$ | $-4.2^{\ddagger}$ | $-5.1^{*}$ | $-4.2^{\ddagger}$ | $-4.1$ |
| | | $\beta_1$ | $-3.8^{\ddagger}$ | $1.2$ | $-2.8$ | $-3.8^{\ddagger}$ | $-2.8$ | $3.4^{\ddagger}$ | $1.4$ | $-3.9^{\ddagger}$ | $-2.0$ | $-3.8^{\ddagger}$ | $-2.8$ | $-3.8^{\ddagger}$ | $-12.0^{*}$ | $-3.8^{\ddagger}$ | $1.2$ |
| | | $\beta_2$ | $-4.5^{\ddagger}$ | $-34.1^{*}$ | $-2.4$ | $-4.5^{\ddagger}$ | $-2.4$ | $2.8^{\ddagger}$ | $3.2$ | $-4.5^{\ddagger}$ | $8.5^{*}$ | $-4.5^{\ddagger}$ | $-2.5$ | $-4.5^{\ddagger}$ | $-1.1$ | $-4.5^{\ddagger}$ | $-1.1$ |
| | | $\beta_3$ | $-3.8^{\ddagger}$ | $-71.4^{*}$ | $-7.0^{*}$ | $-4.0^{\ddagger}$ | $-5.2^{*}$ | n/a | n/a | $-3.7^{\ddagger}$ | $-53.0^{*}$ | $-3.9^{\ddagger}$ | $-5.5^{*}$ | $-3.8^{\ddagger}$ | $-5.6^{*}$ | $-3.8^{\ddagger}$ | $-5.6^{*}$ |
| | .5 | $\beta_0$ | $-3.7^{\ddagger}$ | $-56.1^{*}$ | $-7.5^{*}$ | $-5.0^{\ddagger}$ | $-7.7^{*}$ | n/a | n/a | $-4.6^{\ddagger}$ | $-46.9^{*}$ | $-3.7^{\ddagger}$ | $-7.5^{*}$ | $-3.7^{\ddagger}$ | $-7.7^{*}$ | $-4.3^{\ddagger}$ | $-7.2^{*}$ |
| | | $\beta_1$ | $-3.7^{\ddagger}$ | $-8.9^{*}$ | $-6.3^{*}$ | $-4.0^{\ddagger}$ | $-8.3^{*}$ | $3.0^{\ddagger}$ | $-3.0$ | $-3.9^{\ddagger}$ | $-6.3^{*}$ | $-3.7^{\ddagger}$ | $-6.3^{*}$ | $-3.7^{\ddagger}$ | $-14.0^{*}$ | $-3.7^{\ddagger}$ | $-8.9^{*}$ |
| | | $\beta_2$ | $-4.9^{\ddagger}$ | $-14.0^{*}$ | $-3.5$ | $-5.3^{\ddagger}$ | $-3.1$ | $1.7^{\ddagger}$ | $-2.8$ | $-5.3^{\ddagger}$ | $14.4^{*}$ | $-4.9^{\ddagger}$ | $-3.6$ | $-4.9^{\ddagger}$ | $-1.6$ | $-4.9^{\ddagger}$ | $-1.6$ |
| | | $\beta_3$ | $-4.1^{\ddagger}$ | $-70.6^{*}$ | $-5.6^{*}$ | $-4.4^{\ddagger}$ | $-6.8^{*}$ | n/a | n/a | $-4.1^{\ddagger}$ | $-61.3^{*}$ | $-4.1^{\ddagger}$ | $-7.0^{*}$ | $-4.1^{\ddagger}$ | $-7.1^{*}$ | $-4.1^{\ddagger}$ | $-7.1^{*}$ |
| 4 | 1 | $\beta_0$ | $-5.6^{\ddagger}$ | $-46.4^{*}$ | $-5.1^{*}$ | $-5.9^{\ddagger}$ | $-5.3^{*}$ | n/a | n/a | $-5.7^{\ddagger}$ | $-35.0^{*}$ | $-5.6^{\ddagger}$ | $-5.1^{*}$ | $-5.6^{\ddagger}$ | $-4.5^{*}$ | $-5.6^{\ddagger}$ | $-2.8$ |
| | | $\beta_1$ | $-5.4^{\ddagger}$ | $23.9^{*}$ | $-0.1$ | $-5.0^{\ddagger}$ | $-0.1$ | $3.1^{\ddagger}$ | $2.1$ | $-5.0^{\ddagger}$ | $2.3$ | $-5.0^{\ddagger}$ | $-0.2$ | $-5.0^{\ddagger}$ | $2.4$ | $-5.0^{\ddagger}$ | $23.9^{*}$ |
| | | $\beta_2$ | $-4.8^{\ddagger}$ | $-22.9^{*}$ | $-1.0$ | $-4.8^{\ddagger}$ | $-1.0$ | $3.3^{\ddagger}$ | $-1.7$ | $-4.7^{\ddagger}$ | $7.3^{*}$ | $-4.8^{\ddagger}$ | $-1.0$ | $-4.7^{\ddagger}$ | $1.2$ | $-4.8^{\ddagger}$ | $1.2$ |
| | | $\beta_3$ | $-5.8^{\ddagger}$ | $-61.6^{*}$ | $-5.5^{*}$ | $-6.0^{\ddagger}$ | $-5.5^{*}$ | n/a | n/a | $-5.8^{\ddagger}$ | $-46.2^{*}$ | $-5.8^{\ddagger}$ | $-5.7^{*}$ | $-5.7^{\ddagger}$ | $-5.7^{*}$ | $-5.7^{\ddagger}$ | $-5.7^{*}$ |
| | .9 | $\beta_0$ | $-5.2^{\ddagger}$ | $-45.3^{*}$ | $-4.2^{*}$ | $-5.6^{\ddagger}$ | $-4.0$ | n/a | n/a | $-5.4^{\ddagger}$ | $-34.9^{*}$ | $-5.2^{\ddagger}$ | $-4.1^{*}$ | $-5.2^{\ddagger}$ | $-4.0$ | $-5.2^{\ddagger}$ | $-2.4$ |
| | | $\beta_1$ | $-3.7^{\ddagger}$ | $11.5^{*}$ | $-4.3^{*}$ | $-3.7^{\ddagger}$ | $-4.3^{*}$ | $4.1^{\ddagger}$ | $-2.5$ | $-3.9^{\ddagger}$ | $-2.7$ | $-3.7^{\ddagger}$ | $-4.3^{*}$ | $-3.7^{\ddagger}$ | $-4.8^{*}$ | $-3.7^{\ddagger}$ | $11.5^{*}$ |
| | | $\beta_2$ | $-4.3^{\ddagger}$ | $-20.0^{*}$ | $-1.0$ | $-4.3^{\ddagger}$ | $-1.0$ | $3.7^{\ddagger}$ | $-1.8$ | $-4.2^{\ddagger}$ | $7.5^{*}$ | $-4.3^{\ddagger}$ | $-0.9$ | $-4.2^{\ddagger}$ | $0.8$ | $-4.2^{\ddagger}$ | $0.8$ |
| | | $\beta_3$ | $-5.3^{\ddagger}$ | $-61.0^{*}$ | $-4.9^{*}$ | $-5.6^{\ddagger}$ | $-5.4^{*}$ | n/a | n/a | $-5.3^{\ddagger}$ | $-47.3^{*}$ | $-5.3^{\ddagger}$ | $-5.5^{*}$ | $-5.2^{\ddagger}$ | $5.6^{*}$ | $-5.2^{\ddagger}$ | $5.6^{*}$ |
| | .5 | $\beta_0$ | $-5.8^{\ddagger}$ | $-44.0^{*}$ | $-4.6^{*}$ | $-6.2^{\ddagger}$ | $-4.4^{*}$ | n/a | n/a | $-5.8^{\ddagger}$ | $-37.4^{*}$ | $-5.8^{\ddagger}$ | $-4.6^{*}$ | $-5.8^{\ddagger}$ | $-4.7^{*}$ | $-5.8^{\ddagger}$ | $-3.8$ |
| | | $\beta_1$ | $-4.3^{\ddagger}$ | $2.9$ | $-3.4$ | $-4.3^{\ddagger}$ | $-3.4$ | $3.1^{\dagger}$ | $-0.4$ | $-4.3^{\ddagger}$ | $-0.4$ | $-4.3^{\ddagger}$ | $-3.4$ | $-4.3^{\ddagger}$ | $-4.7^{*}$ | $-4.3^{\ddagger}$ | $2.9$ |
| | | $\beta_2$ | $-3.8^{\ddagger}$ | $-6.6^{*}$ | $1.1$ | $-3.9^{\ddagger}$ | $1.1$ | $3.6^{\ddagger}$ | $0.3$ | $-3.8^{\ddagger}$ | $11.4^{*}$ | $-3.8^{\ddagger}$ | $1.1$ | $-3.8^{\ddagger}$ | $2.6$ | $-3.8^{\ddagger}$ | $2.7$ |
| | | $\beta_3$ | $-5.6^{\ddagger}$ | $-59.4^{*}$ | $-4.7^{*}$ | $-5.9^{\ddagger}$ | $-4.7^{*}$ | n/a | n/a | $-5.6^{\ddagger}$ | $-51.5^{*}$ | $-5.6^{\ddagger}$ | $-4.9^{*}$ | $-5.5^{\ddagger}$ | $-4.9^{*}$ | $-5.5^{\ddagger}$ | $-4.9^{*}$ |

$^{\dagger}$ significant bias in estimate at $P < 0.05$; $^{\ddagger}$ significant bias in estimate at $P < 0.01$; $^{*}$ significant bias in standard error at $P < 0.05$

Table 3.3: Relative bias of fixed effects parameter estimates (against marginal true values) and their standard errors obtained by eight marginal estimation procedures, based on analysis of 1000 simulated datasets per setting ($t$ = number of time points, $\rho$ = autocorrelation, parameters:( $\beta_0$ (intercept), $\beta_1$ (time coefficient), $\beta_2$ (subject level factor), $\beta_3$ (cluster level factor))). Coding for estimation procedures with **probit** link: OLR (ordinary logistic regression), GEEci (generalized estimating equations (GEE) with independence correlation at cluster level), GEEce (GEE with exchangeable correlation at cluster level), GEEf (GEE with fixed effects for cluster level and autoregressive correlation at subject level), GEEs (GEE with autoregressive correlation at subject level), ALR (alternating logistic regression).

| | | | Marginal estimation procedure (probit link) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | OLR | | GEEci | | GEEce | | GEEf | | GEEs | | ALR | |
| t | $\rho$ | par. | Est. | SE | Est. | SE | Est. | SE | Est. | SE | Est. | SE | Est. | SE |
| 8 | 1 | $\beta_0$ | $1.7^{\dagger}$ | $-64.8^{*}$ | $1.7^{\dagger}$ | $-2.0$ | 1.7 | $-2.7$ | n/a | n/a | $1.7^{\dagger}$ | $-38.8^{*}$ | $1.8^{\dagger}$ | $-1.9$ |
| | | $\beta_1$ | $1.4^{\ddagger}$ | $30.6^{*}$ | $1.4^{\ddagger}$ | $-3.2$ | $1.2^{\ddagger}$ | $-3.2$ | $11.6^{\ddagger}$ | $-1.7$ | $1.3^{\ddagger}$ | $-3.0$ | $1.2^{\ddagger}$ | $-3.4$ |
| | | $\beta_2$ | 0.8 | $-46.8^{*}$ | 0.8 | 2.2 | 0.7 | 2.4 | $11.2^{\ddagger}$ | 1.3 | 0.8 | $12.2^{*}$ | 0.8 | 2.4 |
| | | $\beta_3$ | $2.4^{\dagger}$ | $-77.6^{*}$ | $2.4^{\dagger}$ | $-8.1^{*}$ | $2.4^{\dagger}$ | $-8.4^{*}$ | n/a | n/a | $2.3^{\dagger}$ | $-52.0^{*}$ | $2.4^{\dagger}$ | $-8.2^{*}$ |
| | .9 | $\beta_0$ | $-0.6$ | $-64.3^{*}$ | $-0.6$ | $-6.8^{*}$ | $-0.3$ | $-7.5^{*}$ | n/a | n/a | $-0.6$ | $-44.2^{*}$ | $-0.5$ | $-7.0^{*}$ |
| | | $\beta_1$ | 0.6 | 1.3 | 0.6 | $-0.8$ | 0.6 | $-0.6$ | $10.3^{\ddagger}$ | 1.8 | 0.6 | $-0.5$ | 0.6 | $-0.8$ |
| | | $\beta_2$ | $1.4^{\ddagger}$ | $-44.0^{*}$ | $1.4^{\ddagger}$ | $-3.5$ | $1.4^{\ddagger}$ | $-3.5$ | $11.2^{\ddagger}$ | $-5.7^{*}$ | $1.4^{\ddagger}$ | $6.6^{*}$ | $1.4^{\ddagger}$ | $-3.6$ |
| | | $\beta_3$ | $-0.5$ | $-75.3^{*}$ | $-0.5$ | $-1.8$ | $-0.2$ | $-2.3$ | n/a | n/a | $-0.5$ | $-53.1^{*}$ | $-0.5$ | $-2.0$ |
| | .5 | $\beta_0$ | 0.3 | $-61.1^{*}$ | 0.3 | $-3.4$ | 0.2 | $-3.0$ | n/a | n/a | 0.3 | $-48.9^{*}$ | 0.4 | $-3.6$ |
| | | $\beta_1$ | $-0.5$ | $-9.7^{*}$ | $-0.5$ | $-4.0$ | $-0.5$ | $-3.9$ | $8.2^{\ddagger}$ | $-0.5$ | $-0.5$ | $-1.4$ | $-0.5$ | $-4.0$ |
| | | $\beta_2$ | 0.3 | $-20.3^{*}$ | 0.3 | $-3.1$ | 0.3 | $-3.0$ | $9.0^{\ddagger}$ | $-3.0$ | 0.3 | $16.5^{*}$ | 0.3 | $-3.2$ |
| | | $\beta_3$ | $1.9^{\dagger}$ | $-75.0^{*}$ | $1.9^{\dagger}$ | $-5.0^{*}$ | 1.8 | $-5.2^{*}$ | n/a | n/a | $1.9^{\dagger}$ | $-63.5^{*}$ | $1.9^{\dagger}$ | $-5.1^{*}$ |
| 4 | 1 | $\beta_0$ | $-0.7$ | $-54.0^{*}$ | $-0.7$ | $-5.9^{*}$ | $-0.7$ | $-6.3^{*}$ | n/a | n/a | $-0.7$ | $-38.8^{*}$ | $-0.7$ | $-5.9^{*}$ |
| | | $\beta_1$ | $-0.1$ | $45.4^{*}$ | $-0.1$ | $-0.6$ | $-0.1^{\dagger}$ | $-0.6$ | $10.9^{\ddagger}$ | 2.2 | $-0.1^{\dagger}$ | 2.1 | $-0.1$ | $-0.9$ |
| | | $\beta_2$ | 0.5 | $-32.3^{*}$ | 0.5 | $-1.5$ | 0.5 | $-1.4$ | $11.4^{\ddagger}$ | $-1.8$ | 0.7 | $7.7^{*}$ | 0.5 | $-1.4$ |
| | | $\beta_3$ | $-0.6$ | $-67.9^{*}$ | $-0.6$ | $-5.8^{*}$ | $-0.6$ | $-5.7^{*}$ | n/a | n/a | $-0.6$ | $-48.6^{*}$ | $-0.6$ | $-5.9^{*}$ |
| | .9 | $\beta_0$ | $-0.4$ | $-51.8^{*}$ | $-0.4$ | $-2.7$ | $-0.4$ | $-2.7$ | n/a | n/a | $-0.6$ | $-37.7^{*}$ | $-0.4$ | $-2.8$ |
| | | $\beta_1$ | 0.9 | $24.7^{*}$ | 0.9 | $-1.9$ | 0.9 | $-1.9$ | $11.2^{\ddagger}$ | 0.5 | 0.7 | 0.2 | 0.9 | $-1.9$ |
| | | $\beta_2$ | $0.9^{\dagger}$ | $-25.2^{*}$ | $0.9^{\dagger}$ | 2.5 | $0.9^{\dagger}$ | 2.6 | $11.3^{\ddagger}$ | 3.1 | $1.0^{\dagger}$ | $13.3^{*}$ | $0.9^{\dagger}$ | 2.6 |
| | | $\beta_3$ | $-1.0$ | $-67.6^{*}$ | $-1.0$ | $-6.8^{*}$ | $-0.9$ | $-6.4^{*}$ | n/a | n/a | $-1.0$ | $-50.8^{*}$ | $-1.0$ | $-6.9^{*}$ |
| | .5 | $\beta_0$ | 0.7 | $-51.5^{*}$ | 0.7 | $-5.7^{*}$ | 0.8 | $-5.6^{*}$ | n/a | n/a | 0.6 | $-43.1^{*}$ | 0.7 | $-5.7^{*}$ |
| | | $\beta_1$ | 1.1 | $6.7^{*}$ | 1.1 | $-0.4$ | 1.1 | $-0.4$ | $10.6^{\ddagger}$ | 2.4 | 0.9 | 2.5 | 1.1 | $-0.3$ |
| | | $\beta_2$ | 0.5 | $-13.6^{*}$ | 0.5 | $-1.9$ | 0.5 | $-1.9$ | $10.0^{\ddagger}$ | $-2.1$ | 0.6 | $10.8^{*}$ | 0.5 | $-1.9$ |
| | | $\beta_3$ | 0.8 | $-66.3^{*}$ | 0.8 | $-6.6^{*}$ | 0.9 | $-6.6^{*}$ | n/a | n/a | 0.9 | $-56.8^{*}$ | 0.9 | $-6.7^{*}$ |

$^{\dagger}$ significant bias in estimate at $P < 0.05$; $^{\ddagger}$ significant bias in estimate at $P < 0.01$; $^{*}$ significant bias in standard error at $P < 0.05$

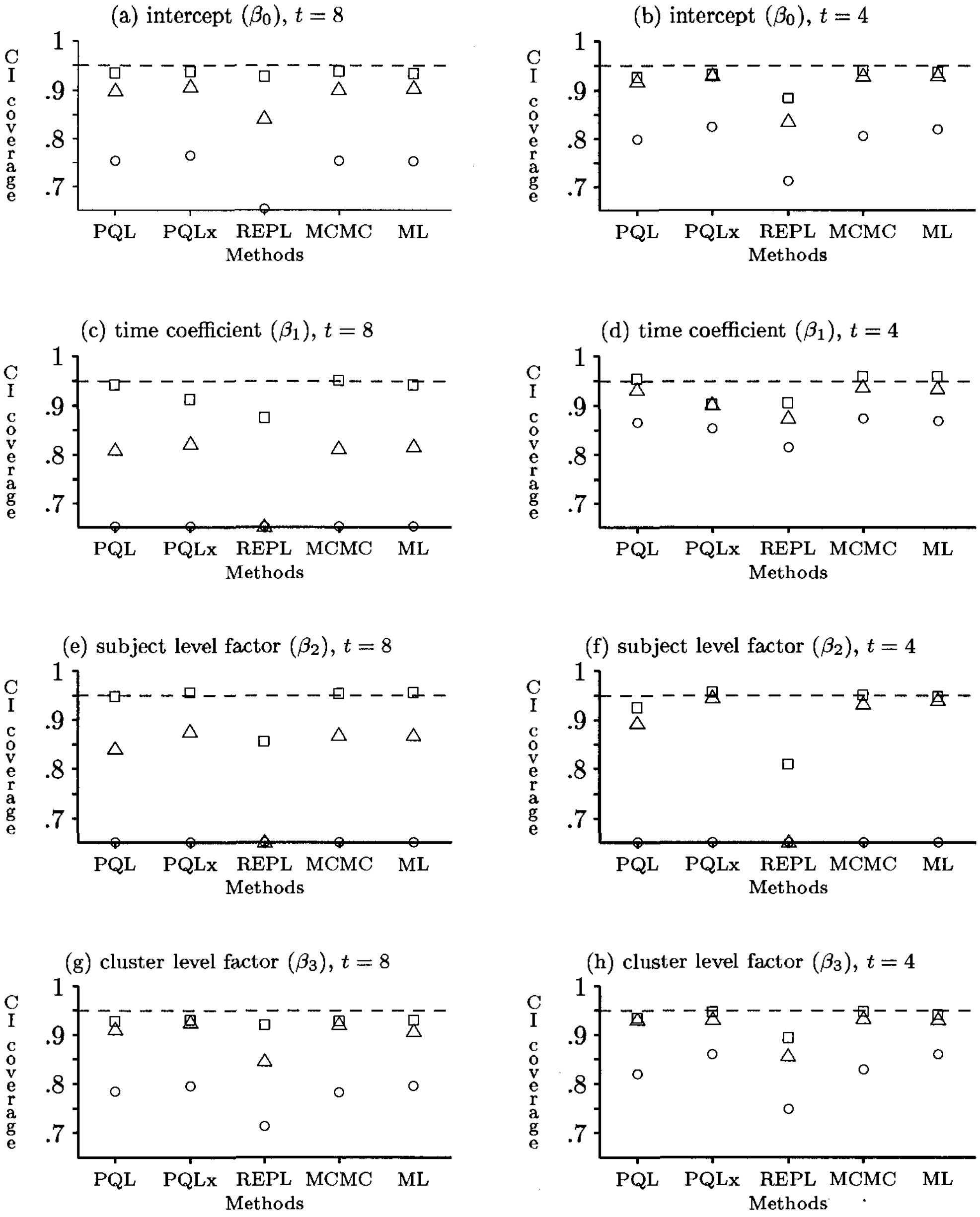

Figure 3.1: Confidence interval coverage for estimates of fixed effects parameters of different estimation procedures, based on 1000 simulated datasets per setting ($t$ = number of time points, ($\rho = (1, .9, .5) \sim (\square, \triangle, \circ)$)). Coding for estimation procedures: PQL (2nd order penalized quasi-likelihood), PQLx (2nd order penalized quasi-likelihood with extra binomial dispersion), REPL (restricted pseudo-likelihood), MCMC (Bayesian Markov chain Monte Carlo), ML (maximum likelihood).

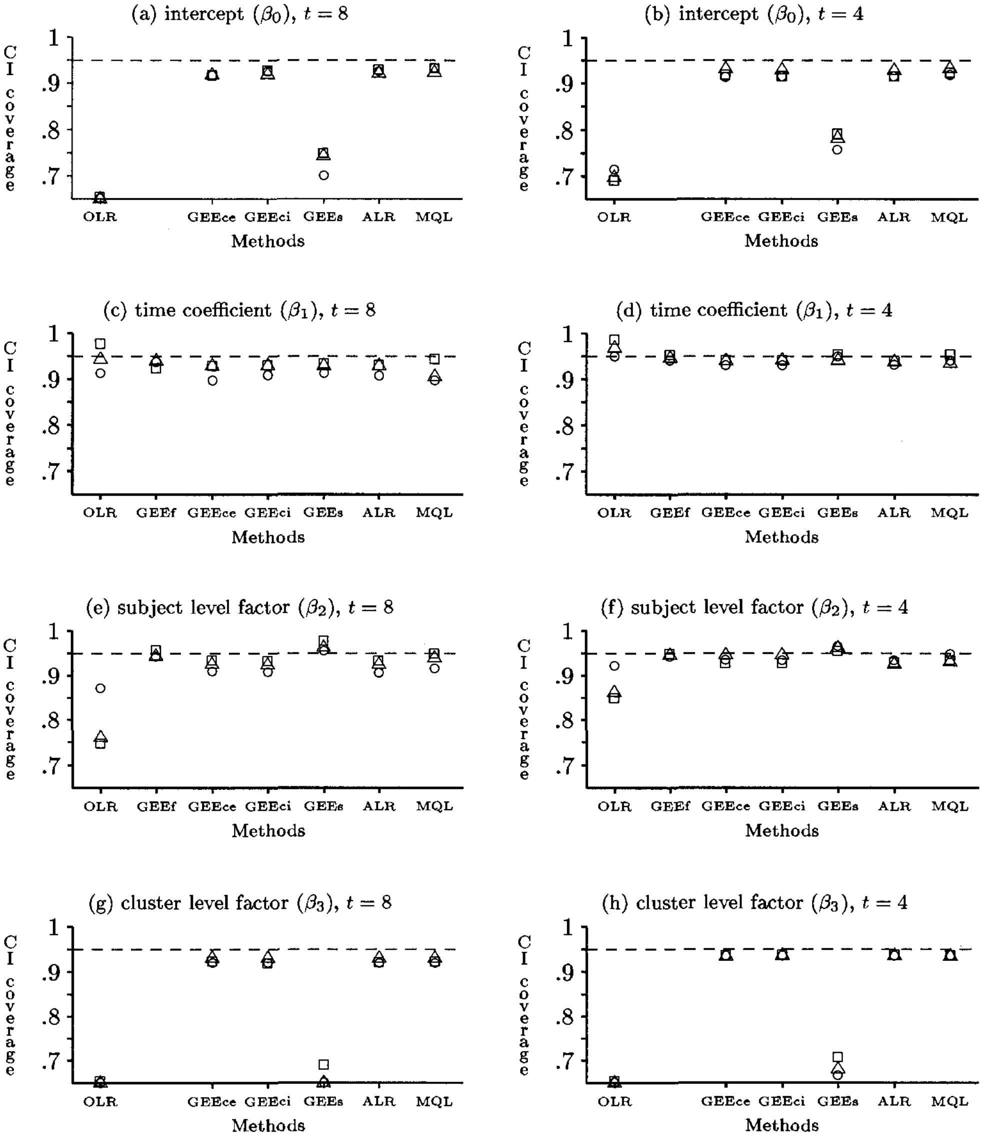

Figure 3.2: Confidence interval coverage for estimates of fixed effects parameters of seven marginal estimation procedures, based on 1000 simulated datasets per setting ($t$ = number of time points, ($\rho = (1, .9, .5) \sim (\square, \triangle, \circ)$)). Coding for estimation procedures: OLR (ordinary logistic regression) ,GEEf (generalized estimating equations (GEE) with fixed effects for cluster level and autoregressive correlation at subject level), GEEce (GEE with exchangeable correlation at cluster level), GEEci ( GEE with independence correlation at cluster level), GEEs (GEE with autoregressive correlation at subject level), ALR (alternating logistic regression), MQL (marginal quasi-likelihood).

# 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

## 4.1 Abstract

The primary objective of the study was to assess the impact of missing values on analysis of binary repeated measures data with an additional hierarchical structure. Such data are commonly encountered in veterinary epidemiological research, and 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. As the basis of a simulation study, random effects models with autocorrelated ($\rho = 1$, 0.9 or 0.5) subject-level random effects were used. 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).

Five different scenarios of simulated incomplete datasets were considered. The first scenario corresponded to a combination of three types of missingness patterns present in the scc40 dataset (scc40 scenario): delayed entry and drop-outs (where subjects enter or leave the study at some point in time, respectively), as well as intermittent missing values. 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. Diggle and Kenward's logistic model [5] was adapted to simulate the missing values.

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 level were relatively unaffected. The presence of autocorrelation in the data did not substantially alter the impact of missing values although the shrinkage of random effects estimates was marginally less pronounced than in the full datasets.

## 4.2 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., 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. Our motivating example was 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 subject.

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 [19, 16, 5, 12, 20, 9, 10] (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 [14]. Thus, additional information about the missingness mechanism is required. Missing data mechanisms have been classified into different categories [19] 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 ([25] and [6, Chapters: 8-9]). Many articles have discussed the choice between the two models (PA vs. SS) (e.g., [6, Chapters 8-9] or more recently for balanced data [21] and Chapter 3 in this thesis). 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 [1, 32]. To our knowledge no studies reported a delayed entry missing values pattern nor its impact. Fitzmaurice [11] 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 [31]. This idea can be extended to incomplete data by matching also the missing data patterns.

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 (Chapter 3) were included: maximum likelihood via numerical integration (ML), Bayesian Markov chain Monte Carlo (MCMC), penalized quasi-likelihood with binomial dispersion (PQL) and extra-binomial dispersion (PQLx), ordinary logistic regression (OLR), alternating logistic regression (ALR), and weighted generalized estimating equations (WGEE).

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.

## 4.3 Missing values

Within the context of binary repeated measures data, let $y_{ij}$ refer to complete binary records on each of $n$ subjects $(i = 1, \ldots, n)$ at $t$ time points $(j = 1, \ldots, t)$. 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 [19] (for a longitudinal data context, see e.g., [16]) 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.

### 4.3.1 Classification of missing data

Missing completely at random (MCAR) [19, 16] refers to a missingness mechanism (or missing data process) that does not depend on prior observed outcome values or an intended measurement values of the outcome (unobserved outcome values), but may depend on covariates such as time. Little and Rubin [19] showed that in the presence of a MCAR process, the estimated parameters are not biased by the absence of data, thus the missing data can be ignored. Diggle and Kenward [5] 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) [19, 16] and random drop-out (RD) [5] refer to a missing data (drop-out) process that depends on the prior observed values of the outcome only. Not missing at random (NMAR) [19, 16] and informative drop-out (ID) [5] refers to a missingness mechanism that depends on the unobserved outcome (current or future unobserved values).

### 4.3.2 Approaches to handle missing data

Several approaches have been proposed to assess and account for missing values [12], including the complete case method (also termed "listwise deletion" [22, Chapter 5]). By this method, subjects with at least one missing value are dropped from the analysis. Fitzmaurice [12] and Little and Rubin [19] 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" ([22, Chapter 5] and [19, 12])). Fitzmaurice [12] argued that WGEE falls under this approach. Kim and Curry [15] 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 [18] and Little and Rubin [19] explained that these methods assume the strong MCAR assumption. Little and Rubin [19] argued that neither complete case method nor the available case method is generally satisfactory.

Little and Rubin [19] showed that a MAR process can be ignored when using likelihood-based inference. Robins *et al.* [27] showed that ordinary GEE does not allow a 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 Janson *et al.* [13]. Hogan *et al.* [10] 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 [24, Chapter 27]. However, these approaches [28] fall beyond the present scope of this Chapter.

### 4.3.3 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" [4]) 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 [30]. 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 patterns 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 [5]:

$$\text{logit}(\Pr(r_{ij} = 1)) = \beta_0 + \beta_1 \text{time}_j + \beta_2 y_{ij-1}. \tag{4.1}$$

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.

### 4.3.4 Hierarchically structured data

The presence of missing values in multilevel data structures has been discussed in the literature [8]. In multilevel datasets, it is possible to have data missing at more than one level [8]. 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 [29] 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 [8] 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 (4.1) can be extended to clustered data by adding random effects to represent heterogeneity between clusters.

## 4.4 Example: Somatic cell count data

The scc40 dataset [7, Chapter 27] 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 9 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 [7, 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.

### 4.4.1 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 ($i = 1, \ldots, n$) distributed on $m$ herds ($k = 1, \ldots, m$) at $t$ time points ($j = 1, \ldots, t$). 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.

#### 4.4.1.1 Missing values caused by drop-outs

A matrix of binary indicators of drop-outs was constructed according to the approach described earlier (Section 4.3.3). 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 (4.1) based on an MAR process:

$$\text{logit}(\Pr(r_{ijk} = 1|v_k)) = \beta_0 + \beta_1 \text{time}_j + \beta_2 y_{ij-1k} + v_k, \tag{4.2}$$

where $(v_1, \ldots, v_m)$ are normally distributed independent random variables, say $v_k \sim N(0, \sigma_h^2)$ where $\sigma_h^2$ 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.

#### 4.4.1.2 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 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):

$$\text{logit}(\Pr(r_{ijk} = 1|v_k)) = \beta_0 + \beta_1 \text{time}_j + \beta_{12} \text{time}_j^2 + v_k, \tag{4.3}$$

with similar random effects assumptions as above. Note that the fixed and random terms in model (4.3) are different from those in model (4.2) as well as the forthcoming model (4.4); for simplicity of notation we retain the same symbols.

#### 4.4.1.3 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 (4.2) 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:

$$\text{logit}(\Pr(r_{ijk} = 1|u_{ik}, v_k)) = \beta_0 + \beta_1 \text{time}_j + \beta_2 y_{ij-1k} + \beta_3 r_{ij-1k} + u_{ik} + v_k, \tag{4.4}$$

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

## 4.5 Statistical methods

### 4.5.1 Estimation procedures

Random effects and marginal estimation procedures were selected based on their performance in the full simulated datasets (Chapter 3). For all procedures except GEE the author refers to the detailed description given therein.

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. The GEE procedure was set up with either an independence or exchangeable working correlation structure at the cluster (herd) level; results from Chapter 3 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 ([13] and [24, 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 (4.2). 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 janson *et al.* [13] to facilitate looping across the simulated datasets.

### 4.5.2 Simulation procedures

In this simulation approach, the full datasets were generated first. Then the desired missing data patterns 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. The same full datasets were used as in Chapter 3 to which the reader is referred for the details. 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 value patterns consisting exclusively of drop-outs were constructed. The drop-out patterns were modelled by either MAR or NMAR processes were adjusted to either low (L) (approx. 31%) or high (H) (approx. 52%) proportions of missing values (designated as MARL/MARH and NMARL/NMARH).

#### 4.5.2.1 Missing values: scc40 scenario

The three types of missing values were simulated in the following order: delayed entry based on model (4.3), drop-outs based on model (4.2), and intermittent missing values based on model (4.4). The parameter estimates of these models for the scc40 data (Table 4.1) were taken as true values for the simulations of the missing value patterns.

#### 4.5.2.2 Missing at random scenarios: MARL and MARH

The scc40 regression estimates (Table 4.1) for the drop-out coefficients in model (4.2) 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.

#### 4.5.2.3 Not missing at random scenarios: NMARL and NMARH

Although this study does not include methods to estimate NMAR models, data could generate 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 (4.2) 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.

### 4.5.3 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 full data (see Chapter 3). 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 ($\beta$), divided by the true value,

$$\text{relative bias to true value (RBT)} = \frac{\hat{\beta}_M - \beta}{\beta} \times 100\% \qquad (4.5)$$

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,

$$\text{relative bias to full data (RBF)} = \frac{\hat{\beta}_M - \hat{\beta}_F}{\beta} \times 100\% \qquad (4.6)$$

One could also use $\hat{\beta}_F$ in the dominator of (4.6); 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.

## 4.6 Results

After a brief review of the parameter estimates (Table 4.1) obtained for the different missingness patterns in the scc40 data, 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 4.2– 4.5, and defer relative biases and standard errors to the true values (RBT) to an appendix (Appendix B, Tables B.1–B.5). The coverages of confidence intervals are shown in Figures 4.1–4.3; these must necessarily refer to the true values. The performance of estimation procedures for the corresponding full datasets was discussed previously (Chapter 3) 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.

### 4.6.1 Missingness types for scc40 data

The strongest effects on patterns of missingness in the scc40 data were found for drop-outs (Table 4.1). 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 dropping out increase 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 4.1). The negative quadratic term ensures the likelihood of a delayed entry missing value decreases 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 4.1), in 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.

### 4.6.2 Random intercept model data ($\rho$ =1)

#### 4.6.2.1 Missing values: scc40 scenario

All estimation procedures gave estimates in fairly close agreement with those of the full datasets (Table 4.2). Small but significant negative biases for the time coefficient ($\beta_1$) 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 (Tables B.1 and 3.1 in Chapter 3).

#### 4.6.2.2 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 $\beta_1$, ranging down below -100%, and thus amounting to a sign switch in the coefficient (Table 4.2). The other coefficients showed a negative bias as well, and the confidence interval (CI) coverage was far below nominal (Figure 4.1).

The two likelihood-based procedures (ML, MCMC) were only a little affected by the missing values, the only consistent significant changes being some increased estimates for $\sigma_2^2$ (Table 4.2). Overall, the proportion of missing values had no impact, except that the MARH scenario produced an additional small positive bias for $\beta_1$ for MCMC. CI coverages were close to but mostly below nominal (Figure 4.1).

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 $\beta_1$) than for PQL (Table 4.2). The extra-binomial parameter estimates of PQLx were centered at 0.72. However, except for $\beta_1$, the coverage of fixed effects CIs was fairly close to nominal for both PQL procedures (Figure 4.1).

The ALR procedure performed well in the MARL scenario, but produced a substantially inflated estimate of $\beta_1$ for MARH. The two weighted GEE (WGEE) procedures showed minor biases for MARL and substantial biases for MARH, in particular in the estimates of $\beta_1$ (Table 4.5). 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 (Table B.5), leading to substantial to strong undercoverage of CIs (Figure 4.1).

#### 4.6.2.3 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 4.2). 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 $\sigma_2^2$ and PQLx). Confidence intervals were strongly affected for $\beta_1$ and OLR but otherwise had coverages fairly close to nominal (Figure 4.1).

### 4.6.3 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 full data (Chapter 3). 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 (Chapter 3).

#### 4.6.3.1 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 4.3–4.4). 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 (Tables B.2–B.3 and 3.1), and the CIs suffered from strong undercoverage for some parameters, in particular for $\rho = 0.5$ (Figures 4.2–4.3).

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.

#### 4.6.3.2 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 overdispersion 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 $\rho$ (Table 4.5). For example, OLR biases generally decreased, whereas ALR biases were stable around zero for MARL, but for MARH the previously observed positive bias for $\beta_1$ 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 $\rho$, but the bias in standard errors and resulting poor coverage of CIs remained (Table B.4 and Figures 4.2–4.3).

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 4.3–4.5). The CI coverages for random effects procedures dropped substantially below nominal with decreasing $\rho$ (Figures 4.2–4.3), but this was attributable to the autocorrelation itself and not a result of the missing values (compare Figure 3.1 in Chapter 3).

## 4.7 Discussion

### 4.7.1 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 by now well advanced statistical understanding of procedures to model missing data (e.g., [19] and [22, 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 [5]. 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 [23], even though they may be more difficult to fit to the missingness patterns. 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.

### 4.7.2 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 time coefficient across all procedures agrees with findings reported by Little and Rubin [19] and Laird [16] that ignoring the NMAR missing process leads to biased estimates, even when only a small proportion of the sample is missing [3]. 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.

#### 4.7.2.1 Weighted generalized estimating equations (WGEE)

A GEE procedure may allow an MAR process to be ignored if the working correlation structure is specified correctly [17, 13]; see however [26] for examples where this does not hold. 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 Janson *et al.* [13] and Molenberghs and Verbeke [24, Chapter 27]. Also small biases have been reported [26], 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 [27], but has also been reported previously for two-level data [26]. 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.

#### 4.7.2.2 Alternating logistic regression (ALR)

One might expect ALR to be affected by missing values in a similar way as GEE although to our knowledge this has not been discussed in the literature. 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.

#### 4.7.2.3 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 [2]. 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 and has the same direction as for OLR. As for ALR, a suitable weighting scheme for PQL under MAR processes could be hypothesized. Allowing for extra-binomial dispersion (PQLx) produced stronger biases and in the opposite direction, adding to the evidence from previous work (Chapter 3) that 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.

#### 4.7.2.4 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 [18]. 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.

## 4.8 References

## References

[1] Ali, M. W., Talukder, E., 2005. Analysis of longitudinal binary data with missing data due to dropouts. *Journal of Biopharmaceutical Statistics*, **15**, 993–1007.

[2] Breslow, N. E., 2003. Whither PQL?. University of Washington Biostatistics Working Paper Series **192**.

[3] Choi, S., Lu, I. L., 1995. Effect of non-random missing data mechanisms in clinical trials. *Statistics in Medicine* **14**, 2675–2684.

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

[5] Diggle P. J., Kenward M. G., 1994. Informative dropout in longitudinal data analysis. *Applied Statististics* **43**, 49–93.

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

[7] Dohoo, I. R., Martin, S. W., Stryhn, H., 2003. *Veterinary Epidemiologic Research.* AVC Inc., Charlottetown, Canada; web-site: `http://www.upei.ca/ver`.

[8] 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.

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

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

[11] Fitzmaurice G., 2004. *Applied Logitudinal analysis.* Wiley-Interscience, Hoboken, NJ.

[12] Fitzmaurice, G. M., 2003. Methods for handling dropouts in longitudinal clinical trials. *Statistica Neerlandica* **57**, 75–99.

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

[14] Kenward, M. G. Goetghebeur, J. T., Molenberghs, G. 2001. Sensitivity analysis for incomplete categorical data. *Statistical Modelling* **1**, 31–48.

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

[16] Laird, N. M., 1988. Missing data in longitudinal studies. *Statistics in Medicine* **7**, 305–315.

[17] Liang, K. Y., Zeger, S. L., 1986, Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

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

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

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

[21] Masaoud, E., Stryhn, H., 2008. A simulation study to assess statistical methods for binary repeated measures data. Submitted manuscript.

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

[23] Molenberghs, G., Kenward, M. G., Goetghebeur, E., 2001. Sensitivity analysis for incomplete contingency tables: the Slovenian plebiscite case. *Applied Statistics* **50**, 15–29.

[24] Molenberghs, G., Verbeke, G., 2005. *Models for Discrete Longitudinal Data.* Springer, New York.

[25] Neuhaus, J. M., 1992. Statistical methods for longitudinal and clustered design with binary responses. *Statistical Methods in Medical Research* **1**, 249–273.

[26] 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.

[27] 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.

[28] Roy, J., 2003. Modeling longitudinal data with non-ignorable dropouts using a latent dropout class model. *Biometrics* **59**, 829–836.

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

[30] 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.

[31] Stryhn, H., Dohoo, I. R., Tillard, E., 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, August 2000.

[32] 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.

Table 4.1: 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: $\beta_0$ = intercept, $\beta_1$ = time coefficient, $\beta_{12}$ = quadratic term for time coefficient, $\beta_2$ = previous outcome, $\beta_3$ = previous outcome missing, $\sigma_h^2$ = herd-level variance, $\sigma_c^2$ = cow-level variance.

| | Type of missing values | | | | | |
|---|---|---|---|---|---|---|
| Param- | Delayed entry | | Drop-out | | Intermittent | |
| eter | Estimate | SE | Estimate | SE | Estimate | SE |
| $\beta_0$ | −0.444 | 0.083 | −4.850 | 0.143 | −4.582 | 0.604 |
| $\beta_1$ | 0.666 | 0.055 | 0.350 | 0.019 | −0.196 | 0.075 |
| $\beta_{12}$ | −0.084 | 0.007 | | | | |
| $\beta_2$ | | | 0.224 | 0.072 | −0.698 | 0.347 |
| $\beta_3$ | | | | | 1.421 | 0.999 |
| $\sigma_h^2$ | 0.017 | 0.011 | 0.068 | 0.026 | 0.295 | 0.257 |
| $\sigma_c^2$ | | | | | 0.938 | 1.008 |

Table 4.2: 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 ($\rho = 1$) 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: $\beta_0$ (intercept), $\beta_1$ (time coefficient), $\beta_2$ (subject level factor), $\beta_3$ (cluster level factor), $\sigma_2^2$ (variance at subject level), $\sigma_3^2$ (variance at cluster level), $\phi$ (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 | $\beta_0$ | $-1.7$ (1.2) | $-0.9$ (1.2) | $-1.4$ (1.2) | 0.0 (1.3) | $-0.8$ (1.3) | $-0.6$ (1.3) |
| | $\beta_1$ | $-2.9^{\ddagger}$(0.4) | $-0.6$ (0.4) | $-1.7^{\ddagger}$(0.6) | 0.0 (0.6) | $-0.8$ (0.6) | $-0.8$ (0.6) |
| | $\beta_2$ | $-0.6$ (0.7) | $-0.6$ (0.6) | $-1.1$ (0.6) | $-0.2$ (0.7) | $-0.5$ (0.7) | $-0.4$ (0.7) |
| | $\beta_3$ | 0.1 (1.4) | 0.3 (1.3) | $-0.2$ (1.4) | 0.9 (1.4) | 0.5 (1.4) | 0.7 (1.4) |
| | $\sigma_2^2$ | | | $-2.2^{\ddagger}$(0.4) | $4.8^{\ddagger}$(0.5) | 1.0 (0.5) | $1.5^{\dagger}$(0.5) |
| | $\sigma_3^2$ | | | $-3.8^{\ddagger}$(1.4) | $-2.2$ (1.5) | $-2.9$ (1.5) | $-3.3$ (1.7) |
| | $\phi$ | | | | $-2.9^{\ddagger}$(0.1) | | |
| MARL | $\beta_0$ | $-8.0^{\ddagger}$(1.3) | $-0.3$ (1.3) | $-1.0$ (1.3) | 1.9 (1.3) | 0.2 (1.3) | 2.5 (1.7) |
| | $\beta_1$ | $-52.1^{\ddagger}$(0.5) | 0.1 (0.4) | $-2.9^{\ddagger}$(0.6) | $10.1^{\ddagger}$(0.7) | 0.1 (0.6) | 1.2 (0.7) |
| | $\beta_2$ | $-3.4^{\ddagger}$(0.7) | $-0.7$ (0.7) | $-1.4^{\dagger}$(0.7) | 0.6 (0.7) | $-1.0$ (0.7) | $-0.4$ (0.8) |
| | $\beta_3$ | $-3.6^{\ddagger}$(1.4) | $-0.9$ (1.4) | $-1.3$ (1.4) | 0.8 (1.5) | $-0.8$ (1.4) | 1.2 (1.7) |
| | $\sigma_2^2$ | | | $-3.5^{\ddagger}$(0.4) | $8.0^{\ddagger}$(0.5) | 0.8 (0.5) | $2.6^{\ddagger}$(0.6) |
| | $\sigma_3^2$ | | | $-2.4$ (1.4) | 1.5 (1.5) | $-1.4$ (1.5) | $-1.4$ (2.0) |
| | $\phi$ | | | | $-5.0^{\ddagger}$(0.1) | | |
| MARH | $\beta_0$ | $-1.5$ (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) |
| | $\beta_1$ | $-140.1^{\ddagger}$(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) |
| | $\beta_2$ | $-11.6^{\ddagger}$(0.7) | 0.0 (0.7) | $-4.1^{\ddagger}$(0.7) | $11.6^{\ddagger}$(0.8) | $-1.0$ (0.7) | $-0.4$ (0.8) |
| | $\beta_3$ | $-11.4^{\ddagger}$(1.3) | 0.0 (1.4) | $-3.6^{\ddagger}$(1.4) | $12.4^{\ddagger}$(1.6) | $-0.6$ (1.5) | 0.3 (1.5) |
| | $\sigma_2^2$ | | | $-17.4^{\ddagger}$(0.5) | $64.0^{\ddagger}$(1.0) | $1.5^{\dagger}$(0.7) | $4.0^{\ddagger}$(0.7) |
| | $\sigma_3^2$ | | | $-7.5^{\ddagger}$(1.4) | $22.8^{\ddagger}$(1.8) | $-1.9$ (1.6) | $-1.6$ (1.7) |
| | $\phi$ | | | | $-20.4^{\ddagger}$(0.2) | | |
| NMARL | $\beta_0$ | $-7.7^{\ddagger}$(1.3) | $-0.5$ (1.3) | $-2.6^{\dagger}$(1.3) | $-0.1$ (1.3) | $-0.7$ (1.3) | $-0.1$ (1.4) |
| | $\beta_1$ | $-88.2^{\ddagger}$(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) |
| | $\beta_2$ | $-2.0^{\ddagger}$(0.7) | $-0.4$ (0.7) | $-1.8^{\ddagger}$(0.7) | 0.0 (0.7) | $-1.5^{\dagger}$(0.7) | $-1.4$ (0.7) |
| | $\beta_3$ | $-2.1$ (1.4) | $-0.5$ (1.4) | $-1.4$ (1.4) | 0.5 (1.5) | $-1.2$ (1.5) | $-0.6$ (1.5) |
| | $\sigma_2^2$ | | | $-2.6^{\ddagger}$(0.4) | $8.5^{\ddagger}$(0.5) | 0.3 (0.5) | 0.9 (0.5) |
| | $\sigma_3^2$ | | | $-2.4$ (1.4) | 1.0 (1.5) | $-2.2$ (1.5) | $-2.5$ (1.7) |
| | $\phi$ | | | | $-6.5^{\ddagger}$(0.1) | | |
| NMARH | $\beta_0$ | $11.0^{\ddagger}$(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) |
| | $\beta_1$ | $-317.8^{\ddagger}$(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) |
| | $\beta_2$ | $-4.3^{\ddagger}$(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) |
| | $\beta_3$ | $-4.9^{\ddagger}$(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) |
| | $\sigma_2^2$ | | | $-14.4^{\ddagger}$(0.6) | $50.4^{\ddagger}$(1.0) | $-11.2^{\ddagger}$(0.7) | $-9.6^{\ddagger}$(0.7) |
| | $\sigma_3^2$ | | | $-8.3^{\ddagger}$(1.4) | $12.5^{\ddagger}$(1.7) | $-10.1^{\ddagger}$(1.5) | $-11.1^{\ddagger}$(1.7) |
| | $\phi$ | | | | $-21.5^{\ddagger}$(0.3) | | |

$^{\dagger}$ significant bias at $P < 0.05$; $^{\ddagger}$ significant bias at $P < 0.01$

Table 4.3: 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 ($\rho$ =**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: $\beta_0$ (intercept), $\beta_1$ (time coefficient), $\beta_2$ (subject level factor), $\beta_3$ (cluster level factor), $\sigma_2^2$ (variance at subject level), $\sigma_3^2$ (variance at cluster level), $\phi$ (extra-binomial dispersion). See Table 4.2 for coding of estimation procedures.

| Scen- | Param- | Statistical Methods | | | | | |
|---|---|---|---|---|---|---|---|
| ario | eter | OLR | ALR | PQL | PQLx | ML | MCMC |
| scc40 | $\beta_0$ | $-2.2^{\dagger}$(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) |
| | $\beta_1$ | $-3.2^{\ddagger}$(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) |
| | $\beta_2$ | 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) |
| | $\beta_3$ | $-0.9$ (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) |
| | $\sigma_2^2$ | | | $17.8^{\ddagger}$(0.3) | $25.9^{\ddagger}$(0.4) | $22.8^{\ddagger}$(0.4) | $23.4^{\ddagger}$(0.4) |
| | $\sigma_3^2$ | | | $5.4^{\ddagger}$(1.3) | $7.8^{\ddagger}$(1.3) | $8.0^{\ddagger}$(1.3) | $8.6^{\ddagger}$(1.5) |
| | $\phi$ | | | | $-6.9^{\ddagger}$(0.1) | | |
| MARL | $\beta_0$ | $-8.6^{\ddagger}$(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) |
| | $\beta_1$ | $-47.5^{\ddagger}$(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) |
| | $\beta_2$ | $-2.2^{\ddagger}$(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) |
| | $\beta_3$ | $-4.0^{\ddagger}$(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) |
| | $\sigma_2^2$ | | | $16.4^{\ddagger}$(0.4) | $28.2^{\ddagger}$(0.4) | $22.4^{\ddagger}$(0.4) | $23.1^{\ddagger}$(0.4) |
| | $\sigma_3^2$ | | | $6.1^{\ddagger}$(1.2) | $10.3^{\ddagger}$(1.3) | $8.8^{\ddagger}$(1.3) | $9.4^{\ddagger}$(1.4) |
| | $\phi$ | | | | $-8.8^{\ddagger}$(0.1) | | |
| MARH | $\beta_0$ | $-0.4$ (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) |
| | $\beta_1$ | $-113.1^{\ddagger}$(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) |
| | $\beta_2$ | $-8.8^{\ddagger}$(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) |
| | $\beta_3$ | $-10.3^{\ddagger}$(1.3) | 0.0 (1.4) | $-1.5$ (1.3) | $12.7^{\ddagger}$(1.4) | 1.8 (1.3) | 1.8 (1.3) |
| | $\sigma_2^2$ | | | $-3.4^{\ddagger}$(0.4) | $54.0^{\ddagger}$(0.9) | $11.6^{\ddagger}$(0.6) | $13.4^{\ddagger}$(0.6) |
| | $\sigma_3^2$ | | | $-1.5$ (1.2) | $24.1^{\ddagger}$(1.5) | $4.2^{\ddagger}$(1.3) | $4.8^{\ddagger}$(1.4) |
| | $\phi$ | | | | $-18.1^{\ddagger}$(0.2) | | |
| NMARL | $\beta_0$ | $-7.0^{\ddagger}$(1.2) | $-2.1$ (1.2) | $-2.1$ (1.1) | $-0.2$ (1.2) | $-1.2$ (1.1) | $-1.3$ (1.2) |
| | $\beta_1$ | $-80.6^{\ddagger}$(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) |
| | $\beta_2$ | $-0.6$ (0.6) | 0.5 (0.6) | 0.4 (0.6) | $-6.4^{\ddagger}$(0.6) | 0.7 (0.6) | 0.9 (0.6) |
| | $\beta_3$ | $-2.4$ (1.3) | $-1.3$ (1.3) | $-1.1$ (1.3) | 0.5 (1.3) | $-0.9$ (1.3) | $-1.1$ (1.3) |
| | $\sigma_2^2$ | | | $2.1^{\ddagger}$(0.3) | $9.9^{\ddagger}$(0.4) | $4.6^{\ddagger}$(0.4) | $4.9^{\ddagger}$(0.4) |
| | $\sigma_3^2$ | | | 0.5 (1.2) | $3.1^{\dagger}$(1.2) | 0.9 (1.2) | 0.9 (1.3) |
| | $\phi$ | | | | $-6.1^{\ddagger}$(0.1) | | |
| NMARH | $\beta_0$ | $13.0^{\ddagger}$(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) |
| | $\beta_1$ | $-299.6^{\ddagger}$(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) |
| | $\beta_2$ | $-3.3^{\ddagger}$(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) |
| | $\beta_3$ | $-4.4^{\ddagger}$(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) |
| | $\sigma_2^2$ | | | $-1.4^{\ddagger}$(0.5) | $46.1^{\ddagger}$(0.8) | $3.0^{\ddagger}$(0.5) | $3.9^{\ddagger}$(0.5) |
| | $\sigma_3^2$ | | | $-1.9$ (1.2) | $15.6^{\ddagger}$(1.5) | $-1.6$ (1.2) | $-1.9$ (1.4) |
| | $\phi$ | | | | $-19.6^{\ddagger}$(0.3) | | |

$^{\dagger}$ significant bias at $P < 0.05$; $^{\ddagger}$ significant bias at $P < 0.01$

Table 4.4: 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 ($\rho$ =**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: $\beta_0$ (intercept), $\beta_1$ (time coefficient), $\beta_2$ (subject level factor), $\beta_3$ (cluster level factor), $\sigma_2^2$ (variance at subject level), $\sigma_3^2$ (variance at cluster level), $\phi$ (extra-binomial dispersion). See Table 4.2 for coding of estimation procedures.

| Scen- | Param- | Statistical Methods | | | | | |
|---|---|---|---|---|---|---|---|
| ario | eter | OLR | ALR | PQL | PQLx | ML | MCMC |
| scc40 | $\beta_0$ | −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) |
| | $\beta_1$ | $-1.9^{\ddagger}$(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) |
| | $\beta_2$ | 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) |
| | $\beta_3$ | −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) |
| | $\sigma_2^2$ | | | $19.9^{\ddagger}$(0.2) | $24.5^{\ddagger}$(0.2) | $21.9^{\ddagger}$(0.2) | $22.2^{\ddagger}$(0.2) |
| | $\sigma_3^2$ | | | $9.2^{\ddagger}$(0.9) | $11.0^{\ddagger}$(0.9) | $9.6^{\ddagger}$(0.9) | $10.8^{\ddagger}$(1.0) |
| | $\phi$ | | | | $-8.5^{\ddagger}$(0.1) | | |
| MARL | $\beta_0$ | $-6.0^{\ddagger}$(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) |
| | $\beta_1$ | $-31.6^{\ddagger}$(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) |
| | $\beta_2$ | −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) |
| | $\beta_3$ | −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) |
| | $\sigma_2^2$ | | | $23.0^{\ddagger}$(0.2) | $25.6^{\ddagger}$(0.2) | $22.1^{\ddagger}$(0.2) | $22.5^{\ddagger}$(0.2) |
| | $\sigma_3^2$ | | | $10.5^{\ddagger}$(0.9) | $13.0^{\ddagger}$(0.9) | $10.8^{\ddagger}$(0.9) | $12.1^{\ddagger}$(1.0) |
| | $\phi$ | | | | $-9.7^{\ddagger}$(0.1) | | |
| MARH | $\beta_0$ | 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) |
| | $\beta_1$ | $-56.9^{\ddagger}$(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) |
| | $\beta_2$ | $-4.5^{\ddagger}$(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) |
| | $\beta_3$ | $-4.3^{\ddagger}$(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) |
| | $\sigma_2^2$ | | | $2.8^{\ddagger}$(0.2) | $17.4^{\ddagger}$(0.4) | $6.9^{\ddagger}$(0.3) | $6.2^{\ddagger}$(0.3) |
| | $\sigma_3^2$ | | | $2.2^{\ddagger}$(0.8) | $12.4^{\ddagger}$(1.0) | $4.6^{\ddagger}$(0.9) | $4.9^{\ddagger}$(1.0) |
| | $\phi$ | | | | $-10.3^{\ddagger}$(0.2) | | |
| NMARL | $\beta_0$ | $-2.8^{\dagger}$(1.2) | −1.0 (1.2) | −0.1 (0.9) | 0.7 (1.0) | 0.1 (1.0) | 0.1 (0.9) |
| | $\beta_1$ | $-65.7^{\ddagger}$(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) |
| | $\beta_2$ | 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) |
| | $\beta_3$ | 0.6 , 1.3 | 1.0 (1.3) | 1.3 (1.0) | 2.1 (1.1) | 1.4 (1.1) | 1.5 (1.1) |
| | $\sigma_2^2$ | | | $2.2^{\ddagger}$(0.1) | $4.4^{\ddagger}$(0.2) | $2.9^{\ddagger}$(0.2) | $2.8^{\ddagger}$(0.2) |
| | $\sigma_3^2$ | | | $1.7^{\ddagger}$(0.8) | $2.8^{\ddagger}$(0.8) | $1.9^{\dagger}$(0.8) | $2.2^{\dagger}$(0.9) |
| | $\phi$ | | | | $-3.4^{\ddagger}$(0.1) | | |
| NMARH | $\beta_0$ | $18.7^{\ddagger}$(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) |
| | $\beta_1$ | $-254.2^{\ddagger}$(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) |
| | $\beta_2$ | $-1.2^{\dagger}$(0.6) | 0.7 (0.6) | 0.1 (0.5) | $4.2^{\ddagger}$(0.5) | 0.6 (0.5) | 0.5 (0.5) |
| | $\beta_3$ | −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) |
| | $\sigma_2^2$ | | | $4.2^{\ddagger}$(0.2) | $17.8^{\ddagger}$(0.4) | $5.6^{\ddagger}$(0.3) | $4.4^{\ddagger}$(0.3) |
| | $\sigma_3^2$ | | | 0.9 (0.9) | $7.4^{\ddagger}$(1.0) | 1.5 (0.9) | 1.7 (1.0) |
| | $\phi$ | | | | $-11.4^{\ddagger}$(0.2) | | |

$^{\dagger}$ significant bias in estimate at $P < 0.05$; $^{\ddagger}$ significant bias in estimate at $P < 0.01$

Table 4.5: 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: $\beta_0$ (intercept), $\beta_1$ (time coefficient), $\beta_2$ (subject level factor), $\beta_3$ (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 | $\beta_0$ | | 0.1 (1.3) | $7.1^{\ddagger}$(1.5) | −1.5 (1.3) | $6.4^{\ddagger}$(1.5) | −0.8 (1.3) | $14.4^{\ddagger}$(1.4) |
| | $\beta_1$ | | $1.7^{\ddagger}$(0.6) | $3.8^{\ddagger}$(0.5) | $1.4^{\dagger}$(0.6) | $3.0^{\ddagger}$(0.6) | 0.3 (0.6) | $1.5^{\dagger}$(0.6) |
| | $\beta_2$ | | −0.1 (0.8) | 0.4 (0.8) | 0.9 (0.7) | $1.4^{\dagger}$(0.7) | 1.0 (0.6) | $2.1^{\ddagger}$(0.6) |
| | $\beta_3$ | | −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 | $\beta_0$ | | $13.5^{\ddagger}$(2.7) | $-6.8^{\dagger}$(3.3) | $7.8^{\ddagger}$(2.9) | −3.4 (3.3) | $7.7^{\ddagger}$(2.4) | $8.5^{\ddagger}$(2.7) |
| | $\beta_1$ | | $-24.0^{\ddagger}$(2.2) | $-20.3^{\ddagger}$(2.1) | $-22.9^{\ddagger}$(2.4) | $-18.2^{\ddagger}$(2.3) | $-13.7^{\ddagger}$(2.3) | $-7.6^{\ddagger}$(2.3) |
| | $\beta_2$ | | −0.3 (2.4) | $-3.9^{\dagger}$(1.7) | 1.7 (2.3) | −3.0 (1.7) | 2.9 (1.8) | 1.5 (1.4) |
| | $\beta_3$ | | −1.7 (2.7) | −4.6 (3.9) | −1.6 (2.6) | −1.5 (3.5) | 2.5 (2.2) | 2.6 (3.0) |

$^{\dagger}$ significant bias in estimate at $P < 0.05$; $^{\ddagger}$ significant bias in estimate at $P < 0.01$

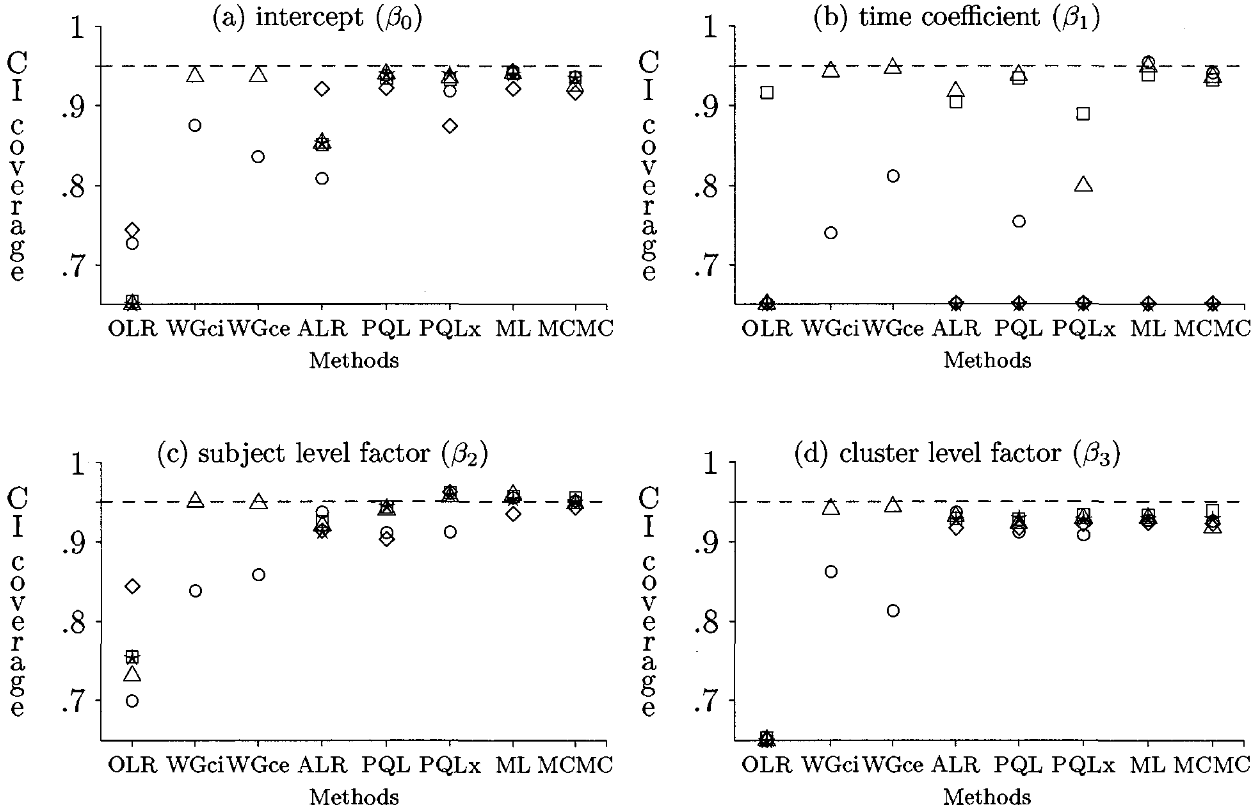

Figure 4.1: Confidence interval coverage for estimates of fixed effects parameters of eight estimation procedures, based on 1000 simulated datasets with missing values generated by random intercept model ($\rho = \mathbf{1}$) in five simulated scenarios of missing values: scc40 (missing values as in 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: $\beta_0$ (intercept), $\beta_1$ (time coefficient), $\beta_2$ (subject level factor), $\beta_3$ (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).

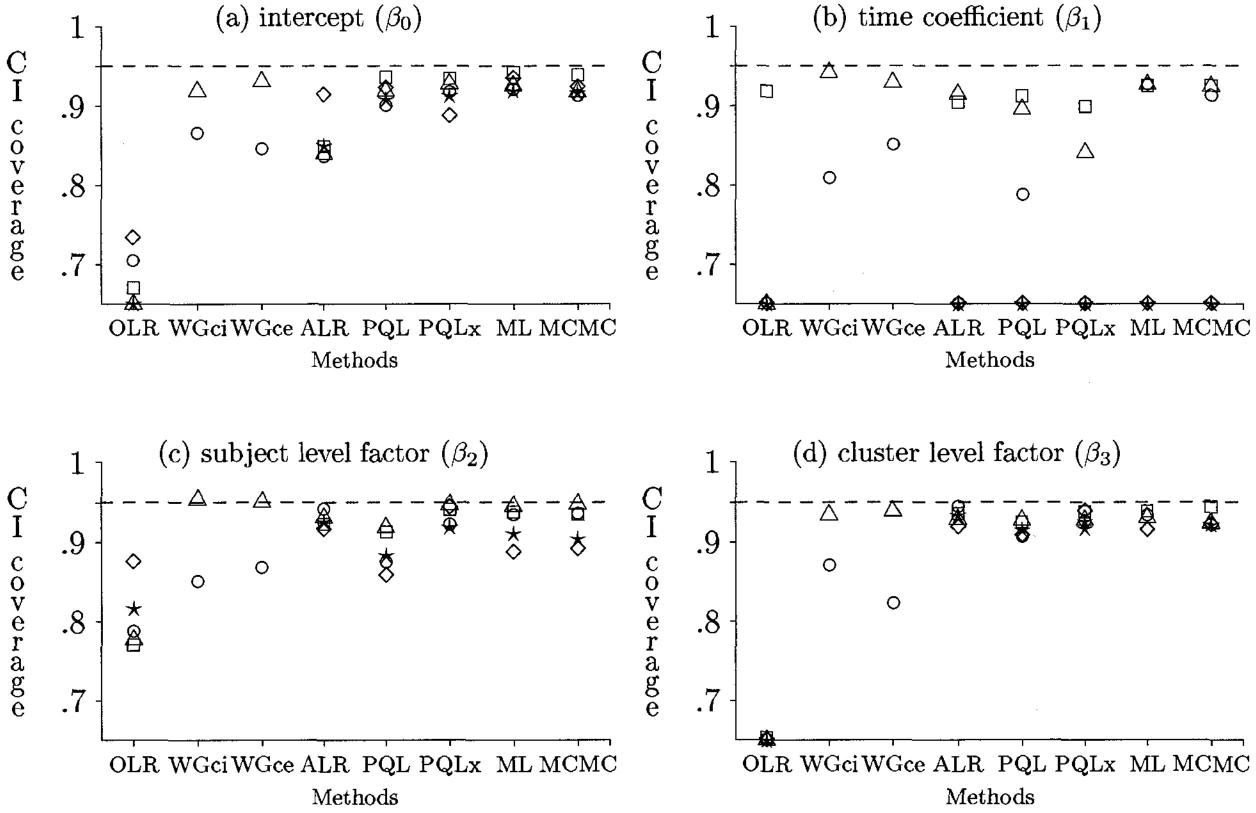

Figure 4.2: Confidence interval coverage for estimates of fixed effects parameters of eight estimation procedures, based on 1000 simulated datasets with missing values generated by random effects model ($\rho$ = **0.9**) in five simulated scenarios of missing values: scc40 (missing values as in scc40 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: $\beta_0$ (intercept), $\beta_1$ (time coefficient), $\beta_2$ (subject level factor), $\beta_3$ (cluster level factor). See Figure 4.1 for coding of estimation procedures.

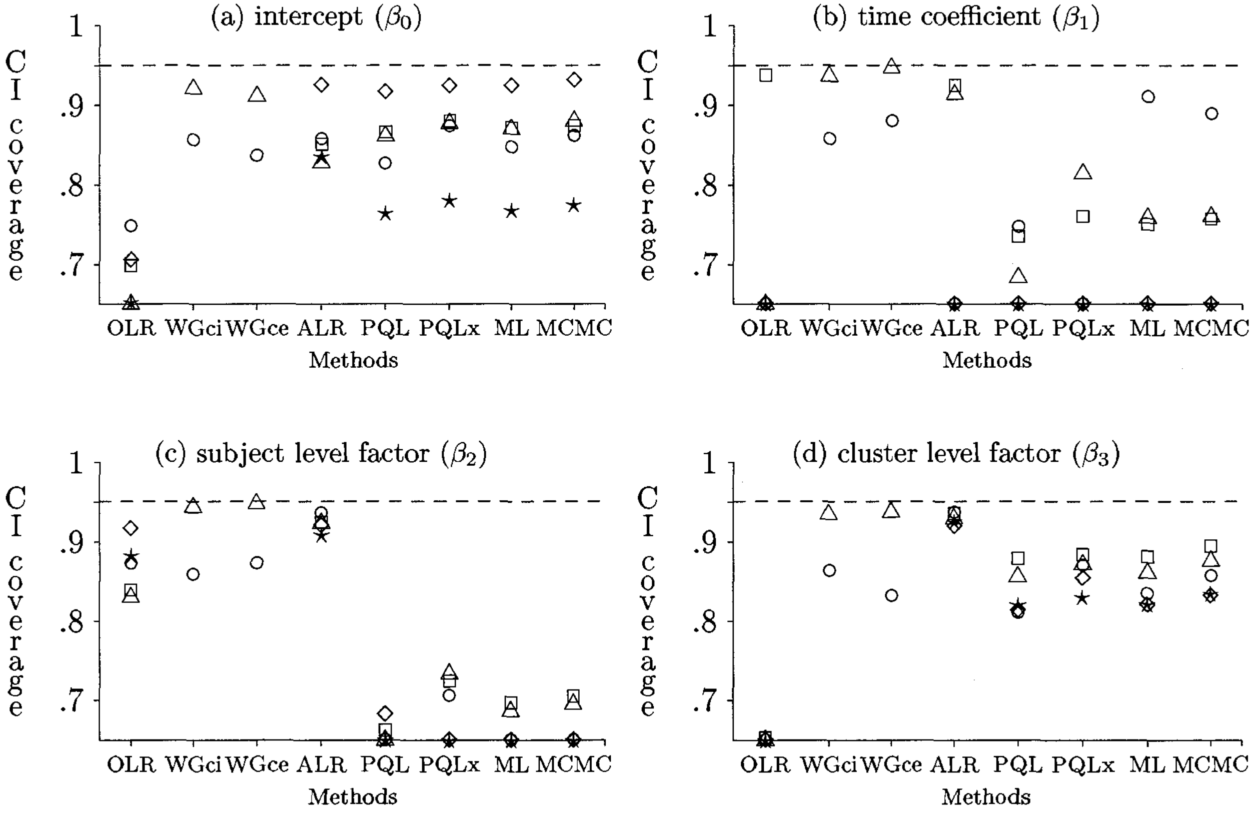

Figure 4.3: Confidence interval coverage for estimates of fixed effects parameters of eight estimation procedures, based on 1000 simulated datasets with missing values generated by random effects model ($\rho = \mathbf{0.5}$) in five simulated scenarios of missing values: scc40 (missing values as in 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: $\beta_0$ (intercept), $\beta_1$ (time coefficient), $\beta_2$ (subject level factor), $\beta_3$ (cluster level factor). See Figure 4.1 for coding of estimation procedures.

# Statistical modelling of neighbour vaccine effects in aquaculture clinical trials

## 5.1 Abstract

In the design of clinical trials involving fish observed over time in tanks, there may be advantages in housing several treatment groups within the same tank. In particular, such "within-tank" designs will be more efficient than designs with treatment groups in separate tanks when substantial between-tank variability is expected. One potential problem with within-tank designs is that it may not be possible to include all treatments in one tank; in statistical terms this means that the blocks (tanks) are incomplete. In incomplete block designs, there may be a concern that the treatments present in the same tank (denoted here as "neighbours") affect each other in their performance. Thus the need for an assessment

of neighbour effects. Two statistical approaches to assess and account for neighbour effects were proposed. The first approach was based on a non-linear mixed model and the second involved cross-classified and multiple membership models. Both approaches were illustrated on simulated data as well as a clinical ISAV (Infectious Salmon Anaemia Virus) trial carried out at the Atlantic Veterinary College. The objective of the fish trial was to investigate the effect of 14 vaccine formulations under disease challenge conditions. The outcome of interest was the mortality during a 6 week follow-up period after challenge.

The objective of the study is to explore two statistical approaches to assess and account for neighbour treatment effects in an incomplete block design setting.

The simulation studies demonstrated that both proposed models show promise in capturing neighbour treatment effects of the type assumed for the models, whenever such neighbour effects are of at least moderate magnitude. In the absence of or with low magnitudes of neighbour effects, the non-linear mixed model faced numerical challenges and produced noisy results. One version of cross-classified and multiple membership model was shown to depend strongly on prior information about variance-covariance parameters for datasets similar to the ISAV data. Analyses of the ISAV trial data by both models did not provide any evidence of substantial neighbour effects.

## 5.2 Introduction

In order to explain the meaning of "neighbour treatment effects", consider our motivating example: an experimental study on vaccination of fish against Infectious Salmon Anaemia Virus (ISAV). In this vaccine trial, fish were held in multiple tanks that each contained several but not all treatment groups. It is common in fish trials to explore and/or adjust for tank effects derived from fish sharing the same environment [25, 18]. A different effect of sharing the same environment might occur if specific treatment groups affected each other; that is, effects occur because of the co-habitation with *specific treatments* instead of the generally shared environment. An extreme example would be that the presence of an ineffective (or control) group in a tank caused the other treatments groups in the tank to perform poorly due to infection spread. In order to consider such an effect, it must be biologically plausible that a transfer of treatment characteristics can take place within the same tank. An alternative interpretation could be as a competition effect, if fish within different treatment groups compete for limited resources. Competition effects have been studied in different contexts such as plant production [14, 9]; insects [20] and fish [21, 19].

However, the present work is not based on specific designs in competition experiments and does not aim to explore the degree to which species or varieties compete, so we abstain from the using the term competition and denote simply this type of effect as a "neighbour treatment effect".

A common usage of the term "neighbour effects" can refer to datasets or experimental designs in which there is a (strong) correlation between adjacent experimental units (neighbours). Such designs have been studied for use e.g., in agriculture and forestry [1], and the analysis typically involves methods of spatial statistics. In education studies, neighbour effects can refer to the effect of the neighbourhood social interaction [17].

A necessary condition to study neighbour treatment effects, as described above, in designs with subjects in different treatment groups within clusters, is that the clusters do not contain all treatments (to the same degree) because then it will not be possible to separate neighbour treatment effects from the usual treatment and cluster effects. In terms of statistical experimental design, this means that the clusters form incomplete blocks for the treatments; we elaborate on the experimental design below. For the ISAV trial, and similar aquaculture clinical trials, this condition is met and a neighbour treatment effect seems possible or perhaps plausible; thus there is a need for statistical methodology to handle neighbour treatment effects in data analysis.

The objective of the study is to explore two statistical approaches to assess and account for neighbour treatment effects in an incomplete block design setting. The first approach is based on a non-linear model, and the second involves cross-classified and multiple membership models. Both approaches will be applied to the ISAV trial data and supplemented by simulation studies targeted at comparable parameter values.

## 5.3 Statistical design, modelling and analysis

In this section, we briefly review the concepts of an incomplete block design, thereafter introduce the two statistical models or approaches for estimation of neighbour treatment effects, and in a final section discuss the issue of model identifiability.

### 5.3.1 Incomplete block designs

Generally, one of the basic principles in experimental design is the reduction of variation between the treated units (experimental error). This is the primary motivation for introducing blocks, groups of similar experimental units, in randomized complete block designs [7] where the treatments are allocated randomly among the units within each block. In the ISAV clinical trial context, the tanks may be considered as blocks. If a blocking scheme induces a within-block variation substantially smaller than the between-block variation, there may be large gains in efficiency of the block design compared to a completely randomized design (where each block/tank covers only a single treatment). When designing an experiment, e.g., a clinical trial, the block size may however be determined or limited on physical/logistical grounds so that the blocks cannot comprise all treatments. The resulting design is called an incomplete block design [7], of which the ISAV trial is an example.

Designs for incomplete blocks range from balanced to unbalanced block designs. Multiple types of balanced and partially incomplete designs exists. The classical balanced incomplete block design (BIBD) exists for certain combinations of the number of treatments, blocks, and block sizes. This design requires that every pair of treatments occurs together within the same block an equal number of times [7, Chapter 11].

According to classical statistical theory for incomplete block designs with fixed effects of treatments and blocks [12], the analysis of incomplete block designs include both intra- and inter-block information. The intra-block analysis refers to a situation when the contrasts in the treatment effects are estimated as linear combinations of comparisons of observations in the same block. The inter-block analysis refers to the information contained in the comparison of block totals and called "recovery of inter-block information".

An alternative view of (incomplete) block designs is as a simple hierarchical structure where the experimental units are clustered within blocks [8, Chapter 20]. From this perspective it follows that block effects should be modelled as random effects; they may also be viewed as nuisance parameters (of no intrinsic interest), and their modelling by random effects may increase the precision on treatment estimates in incomplete block designs [11]. Random block effects induce a correlation between units in the same block whereas units in different blocks remain independent.

### 5.3.2 Notation and model framework

Throughout we use the following general (single index) notation. Let $y_i$ denote a continuous measurement on the $i$th experimental unit ($i = 1, \ldots, n$) located in block $\texttt{bl}(i)$ and subjected to treatment $\texttt{tx}(i)$, where $\texttt{bl}(i)$ and $\texttt{tx}(i)$ give the block and treatment number of unit $i$. Blocks are labeled $1, \ldots, b$ and treatments labeled $1, \ldots, a$; thus, $a$ is the number of treatments and $b$ is the number of blocks. For simplicity of notation, we will formulate our models in the context of blocks of size three, as in the ISAV trial data. Therefore, each experimental unit will be joined by two other treatments in its block; we call these neighbour treatments and denote their treatment numbers as $\texttt{n1}(i)$ and $\texttt{n2}(i)$. Both models can be applied to other block sizes in a straightforward manner.

### 5.3.3 Non-linear mixed model (NLM)

The idea of this model is to capture a simple, possibly the simplest, way in which a treatment may affect its neighbour treatments in the same block: by an additive effect determined by scaling of the treatment effect itself. As an illustration, it might be conceivable that a treatment in addition to its effect on the treated unit contributes 20% of this effect to the neighbour units as well. In the context of competition for limited resources one might expect the neighbour effect to be negative (say -20%), so that a high-performing treatment reduces the neighbour treatments by 20% of its own effect. In a model equation including also the previously discussed random block effects, this idea takes the following form:

$$y_i = \mu + \beta_{tx(i)} + \delta(\beta_{n1(i)} + \beta_{n2(i)}) + u_{bl(i)} + e_i, \tag{5.1}$$

where $\mu$ is the overall mean; $\beta_1, \ldots, \beta_a$ are fixed treatment parameters normalized by the restriction $\sum \beta_j = 0$ where $(j = 1, \ldots, a)$; $\delta$ is a fixed neighbour treatment parameter; $u_1, \ldots, u_b$ are random block effects assumed to follow the Gaussian (normal) distribution $N(0, \sigma_b^2)$; and $e_1, \ldots, e_n$ are error terms assumed $\sim N(0, \sigma_e^2)$. The equation shows how the neighbour treatment effects enter as additive terms formed by multiplying the respective treatment effects by the scaling parameter $\delta$. Obviously, $\delta = 0$ corresponds to no neighbour treatment effects, and $\delta$ may take both positive and negative values. As $\delta$ and the $\beta_j$'s enter into the equation in a non-linear manner (by the multiplication), the model is non-linear in its parameters and therefore a non-linear mixed model [27].

The fixed part of the model takes a non-standard form, but the random part of the model is very simple and allowed the model to be fit using the (`nlmixed`) procedure in SAS software [27, 23]. The procedure employed adaptive Gaussian quadrature to approximate the likelihood function and a quasi-Newton search algorithm to locate the maximum of the (approximate) log-likelihood function. The restriction on the treatment parameters that their sum be zero is equivalent to setting a baseline treatment, but avoided choosing an arbitrary baseline treatment and improved the performance of the search algorithm. Sensible starting values were given for all parameters (for $\delta$ a range of values were offered), and $\delta$ was restricted in range to values within $(-3,3)$ to prevent the algorithm from diverging into nonsensical domains of the parameter values. Variances were bounded below at $(0.001)^2$ to avoid problems for the search algorithm resulting from zero variances.

A change of the model (5.1) to incorporate random instead of fixed treatment effects, in the spirit of the models to be described in the next section, was not possible within the estimation framework of the SAS procedure (or any other software), and was therefore not investigated further.

### 5.3.4 Cross-classified and multiple membership models (CC, MMI and MMCP)

Cross-classified and multiple membership models extend multilevel mixed models to non-hierarchical data structures, in two different ways. A cross-classified data structure exists if each experimental unit is a member of two separate hiercharchies instead of a single hierarchy with multiple levels. Models for randomized block designs with random effects of both treatment and block factors can be viewed as the simplest example of a cross-classified data structure [7, 24]. The models in this section represent the treatment effects by random effects instead of fixed effects in order to model neighbour treatment effects in terms of correlation structure instead of fixed effects. For reference, we formulate first the cross-classified model (CC) without neighbour treatment effects,

$$y_i = \mu + u_{bl(i)} + v_{tx(i)} + e_i, \tag{5.2}$$

where $\mu$ is the overall mean, and where $(u_1, \ldots, u_b)$, $(v_1, \ldots, v_a)$ and $(e_1, \ldots, e_n)$ are sets of independent random variables representing the random block, treatment and error terms, respectively, and assumed to follow zero-mean Gaussian distributions with variances $\sigma_b^2$, $\sigma_t^2$ and $\sigma_e^2$.

Multiple membership (MM) models allow a lowest level unit to be a member of more than one higher classification unit [15, 6]. Examples of the use of this model are: students (pupils) changing schools during a term which therefore have contributions from two schools; patients in a hospital attended by more than one doctor or nurse; and populations of production animals (fish, chicken) that originate from several different sources (hatcheries). The multiple membership model has also been proposed as a model for spatial dependence as an alternative to e.g., the commonly used CAR models [3]; the idea is that the neighbours of a given unit (e.g., region) has its neighbouring regions included in a multiple membership classification. This is the closest analogy to our use of multiple membership here; the neighbour treatments within a block are treated as (spatial) neighbours in a (spatial) MM model. Adding these terms to the CC model yields our MMCP (multiple membership with correlated pairs) model,

$$y_i = \mu + u_{bl(i)} + v_{tx(i)} + 0.5v^*_{n1(i)} + 0.5v^*_{n2(i)} + e_i, \tag{5.3}$$

where the variables $v_1^*, \ldots, v_a^*$ represent the neighbour effects of each of the treatments. The assumptions for the $u_j$ and $e_i$ are unchanged from above, and the pairs $(v_1, v_1^*), \ldots, (v_a, v_a^*)$ are independent and assumed to follow a two-dimensional normal distribution $N(0, 0, \sigma_t^2, \sigma_n^2, \rho)$. The variances represent the heterogeneity between treatments and neighbours respectively, whereas the correlation is between the treatments effects and its associated neighbour effects. A simpler version of model (5.3) assumes independence between treatment and neighbour effects, i.e. $\rho = 0$ (MMI model).

The MMI model corresponds to the MM model previously used in the literature, but the assumption that treatment and neighbour effects are unrelated may seem unnatural in the present context. The MMCP extension is designed to quantify a correlation between the treatment and neighbour effects, and is similar to an extension of the multiple membership model for spatial applications proposed by Langford *et al.* [16]. If the MM variance component is substantial, one could furthermore plot the estimated random effects for treatments and neighbour effects to study their dependence pattern. A positive correlation would mean that being together with a "good" treatment tends to produce a "good" performance, a negative correlation produces the converse. In Equation (5.3), the neighbour effects enter with weights of 0.5; generally the weights are assumed to sum to 1 for each lowest level unit. Criteria for choosing appropriate weights may depend on the information available, however we have followed a simple approach and assigned equal weights [6].

MM models are generally fit in a Bayesian setting using MCMC estimation, but if the model is specified with vague ("non-informative") prior distributions the approach effectively uses the Bayesian framework as an estimation algorithm for an otherwise untractable model. The CC and MMI models can be fit using the MLwiN software [3], but the MMCP extension was programmed in the WinBUGS software; for convenience, WinBUGS version 1.4 was used for all analyses. Prior distributions were generally vague: $N(0, 10^6)$ for $\mu$; the classical gamma distribution $(10^{-3}, 10^{-3})$ for the inverse variances $\sigma_b^{-2}$ and $\sigma_e^{-2}$ [4]; and a Wishart distribution with a diagonal variance covariance matrix of 0.1 and degrees of freedom of 2 for the inverse covariance matrix of two-dimensional normal distribution. Browne and Draper [5] reported in a simulation study that using Wishart priors may result in biases especially in small datasets (see also the discussion on Wishart priors by Browne [2]). Thus a sensitivity analysis based on a range of values (0.01, 0.25) for the diagonal variance covariance matrix was carried out. Given the range of the treatment and neighbour effects standard deviations (0.1-0.5), the choice of 0.1 in the diagonal variance covariance matrix seemed a reasonable overall choice. For the simulated data, Markov chains were run with 10 000 burn-in samples, and the subsequent estimates (posterior distribution medians) were based on 100 000 samples, whereas for the ISAV dataset the posterior distribution medians were based on 1 000 000 samples.

### 5.3.5 Model identifiability

A model is defined as identifiable in a situation where the model parameters are uniquely determined from the distribution of the observed random variables [22]. For estimation procedures based on maximization of a target function (e.g., the log-likelihood function), non-identifiability of parameters usually manifests itself as non-convergence of the search algorithm or extreme sensitivity of the final estimates to initial values provided to the algorithm. Such deficiencies will often appear more clearly in data with low residual (error) variation, and the identifiability of a model may therefore be determined from simulated data with low residual variation [10]. Preliminary analyses of the NLM model established that the model could be non-identifiable in some designs with small number of treatments and block sizes (e.g., in the smallest possible BIBD), and that its non-linear model counterpart with fixed block effects could be non-identifiable even for larger designs.

Non-identifiability of Bayesian models estimated by MCMC often manifests itself by poor convergence of the Markov chains to a stationary distribution, although a nicely converged chain may still contain non-identifiable parameters or combinations of parameters [26]. We used the BGR diagnostic based on multiple chains [13] to assess convergence from different starting values and examined the correlations of model parameters to ascertain that converged chains did not mask non-identifiable parameters.

## 5.4 Infectious Salmon Anaemia Virus (ISAV) trial

The objective of the ISAV clinical trial was to evaluate the immune response to Infectious Salmon Anaemia Virus in Atlantic Salmon after vaccination with different vaccine formulations. Fourteen vaccine formulations were investigated. Fourteen tanks were used in the trial, each containing three different randomly allocated vaccinated groups composed of 50 fish each. Each vaccine formulation was replicated three times among the study tanks, and each tank held a unique combination of three applied vaccination formulations, for a total of 42 vaccine groups. The fish were tagged with a unique colour-coded tag to identify the vaccine group. The outcome of interest here is the mortality during a 6-week follow-up period after challenge with the virus. The trial was carried out during March–August 2000 by Dr. Shona White at the Atlantic Veterinary College, University of Prince Edward Island, Canada.

The design of the ISAV trial is shown in Table 1, with mortality rates for each treatment–tank combination. The incomplete nature of the blocks is evident from the table. It is also clear that all pairs of treatments do not occur equally often in the same tank; e.g., treatments 6 and 9 meet once in tank 14, whereas treatments 1 and 2 do not meet at all. Therefore, the design is not balanced in the sense of a BIBD, and nor does it correspond to any other specialized incomplete block design [7, Chapter 11].

The ISAV data was modelled by the models of Section 5.3 by defining a continuous outcome $y_i$ for Models (5.1)–(5.3) for the fish group $i$ as the mortality rate at 6 weeks. This approach involved two data reductions and approximations. First, the proportions were really scaled binomial outcomes with denominator 50 and not truly continuous outcomes. With equal and large denominators as well as proportions well away from the extremes of the unit interval, the normal distribution model should provide a fair approximation to the binomial distribution with fairly homoscedastic variances. A constant variance may be a better approximation than the binomial variance ($\mathrm{Var}(y_i) \propto p_i(1 - p_i)$) if there is clustering at the fish group level (i.e., the 50 identically treated fish are more alike than expected from a binomial distribution). Models (5.1)–(5.3) can easily be extended to logistic regression models but this was considered an unnecessary complication for the purpose of studying the models. Second, a substantial data reduction was implied by ignoring the survival curve up till 6 weeks and focusing only on the resulting mortality. We believe the models have the potential to be extended to survival data, but this was considered beyond the scope of the present study. It could also be argued that from a practical perspective, the overall mortality at the end of the study is a more direct measure of vaccine efficacy than e.g., hazard rates based on the entire follow-up period.

## 5.5 Simulation study

A targeted simulation study based around the ISAV clinical trial dataset was carried out. All simulated data were generated within the experimental design of the ISAV data (Table 5.1); in particular, $a = 14$, $b = 14$ and $n = 42$. Both the NLM model (5.1) and the MMCP model (5.3) were used as true models for the simulated datasets. A total of six scenarios of simulated datasets were included: three scenarios (A.1–A.3, Table 5.2) corresponding to model (5.1), and three scenarios (B.1–B.3, Table 5.3) corresponding to model (5.3). The first NLM scenario (A.1, Table 5.2) and the second MMCP scenario (B.2, Table 5.3) were linked to the ISAV dataset by having true parameters close to those obtained for the ISAV data (Table 5.4). The first MMCP scenario (B.1, Table 5.3) had $\rho = 0$ to yield the simpler MMI model. The correlated pair of random effects $(v_j, v_j^*)$ was generated from uncorrelated, standard normally distributed random variables $(v_j, x_j)$ by defining $v_j^*$ as $v_j^* = \rho v_j + (\sqrt{1-\rho^2})x_j$, where $x_j \sim \mathrm{N}(0,1)$, and scaling with respective standard deviations.

### 5.5.1 Analysis of results for simulated data

For the analyses by the NLM model, the means and standard deviations of estimates across simulations, as well as the mean standard errors were reported (Table 5.2). The treatment estimates were converted to a standard deviation $\sigma_t$ between treatments (without associated standard error), and the individual treatment estimates were omitted. For the multiple membership models analyzed within a Bayesian framework, the means and standard deviations of the estimate's posterior medians across simulations, as well as the mean posterior standard deviations were reported. In addition, the difference in the deviance information criterion (DIC) between the MMCP and MMI models were calculated for each simulated dataset and reported (Table 5.3). For the ISAV dataset, both the actual treatment estimates and DIC values were presented (Table 5.4).

## 5.6 Results

The results of the simulation studies are presented according to the true model behind the data (NLM: Table 5.2; MMCP: Table 5.3), then followed by the analysis of the ISAV trial dataset (Table 5.4). Our main focus will be on the neighbour effects $\delta$ and $(\sigma_n, \rho)$ . The mean estimates across simulations will be compared to the true values, and the mean standard errors (SE), or posterior standard deviations for the Bayesian models, will be compared to the standard deviations across simulations (SD).

### 5.6.1 Non-linear mixed (NLM) model data

Across all scenarios and estimation procedures, the overall mean ($\mu$) and error standard deviation ($\sigma_e$) were estimated consistently close to the true values (Table 5.2). The performance for the other parameters varied across both the data scenarios and estimation procedures. A general pattern observed for the NLM estimates was that the SD was always, and at times much, larger than the corresponding SE. NLM searches for the ML estimates frequently lead to the block variance reaching its lower boundary; this could indeed lead to underestimated standard errors when calculated from the observed Hessian matrix at the parameter estimates.

In scenario A.1 with low variances and neighbour effect, the mean NLM estimates for $\sigma_t$, $\sigma_b$ and $\delta$ were far from their true values. The SD for $\delta$ was large, with a wide 95% range across simulations of (−.467, 1.24), and about 2.5 times the size of the SE. This signals that NLM estimates in this scenario were very noisy and associated with grossly incorrect inference. In scenarios A.2–3, all estimates were closer to their true value and associated with smaller SD; these scenarios may be considered to show an acceptable performance of the procedure although some biases still exist. For example, the 95% range for $\delta$ across simulations for scenario A.3 was (.084,.788), which seems quite reasonable. Simulations of additional scenarios with large between-block variances also showed large SDs and wide ranges in the estimates of $\delta$ (results not shown).

As the MMCP models do not match the true models, a close agreement of the estimates with the true values cannot be expected. For scenario A.1, the MMI model with independent (and fairly small) neighbour effects performed better than the MMCP model in terms of parameter estimates. This could be due the sensitivity of the Wishart prior in the MMCP model [2]. Using a value of 0.01 in the diagonal variance covariance matrix resulted in a better agreement between the estimates and their corresponds true values, for example the mean estimates of $\sigma_t^2$, $\sigma_n^2$ and their associated SD and SE were equal to .092 (.023, .032) and .089 (.017, .042), respectively. In scenarios A.2-3 (Table 5.2), the MMI model showed some discrepancy in estimating its parameters, resulting in estimates of $\sigma_t$ and $\sigma_n$ much smaller than their true values, and for $\sigma_n$ a pronounced difference between the SD and SE. However, the MMCP model seemed to gain some improvement over the MMI model in those scenarios and resulted in closer estimates of $\sigma_t$ and $\sigma_n$ to their true values especially in scenario A.3 (Table 5.2). Note that the linear relationship between treatment and neighbour effects in the NLM true model of the data should correspond to a perfect correlation $\rho$ (irrespective of the value of $\delta$). This explains why $\rho$ increases in the stronger scenarios, although the values are still far from 1.

### 5.6.2 Cross-classified and multiple membership (MMCP) model data

Similarly to the NLM model data, the overall mean ($\mu$) and error standard deviation ($\sigma_e$) were estimated consistently close to the true values (Table 5.3). A general pattern observed for the MMI and MMCP estimates was that the SD was almost always, and at times substantially, smaller than the corresponding SE.

In scenario B.1 with $\rho = 0$, low variances and neighbour effect, the MMI estimates were close to their true values, with a reasonable agreement between SD and SE. On the other hand, the MMCP estimates were further from their true values and associated with a pronounced disagreement between SD and SE, especially for treatment and neighbour effect standard deviations (Table 5.3). The NLM estimates showed similar patterns as the MMCP model estimates. The mean estimate of $\delta$ was small (0.11) with a wide 95% range across simulations of (-2.409, 2.822), as indicated by the large SD. Allowing for a weak dependence between treatment and neighbour effects in scenario B.2 produced only minor changes in the MMI and MMCP estimates. The sensitivity analysis for the MMCP model showed a strong impact of the Wishart prior. For example, in scenario B.2 and with a value of .01 in the diagonal, the estimates with (SD, SE) for $\sigma_t$ and $\sigma_n$ were: .093 (.022, .032) and .096 (.020, .046), respectively.

With stronger dependence between treatment and neighbour effects and larger variation between treatments and neighbours (scenarios B.3), the MMCP parameter estimates improved further, and were closer to their true values than those from the MMI model. The sensitivity analysis showed that a larger value (.25) in the diagonal variance covariance matrix of the Whishart prior gave estimates very close to the true values (.482 and .505 for $\sigma_t$ and $\sigma_n$, respectively). On the other hand, a value of .01 in the diagonal resulted in shrunk estimates. The main problem for the MMI model was a far too low estimate of $\sigma_n$.

### 5.6.3 Infectious Salmon Anaemia Virus (ISAV) data

The mortality rates of the ISAV trial were in the range of 0.2–0.7 (Table 5.1). The data were modelled by four different models (NLM, CC, MMI, MMCP) all assuming normally distributed random effects and errors. Two additional approaches were briefly explored, one based on binomial data and rephrasing equations (5.1) and (5.3) on logistic scale, and one based on restricting the variance of normally distributed proportions to follow the binomial distribution. Both approaches faced numerical challenges.

The results (Table 5.4) showed close agreement between the four models in estimates of $\mu$ and $\sigma_e$; however, the standard errors from NLM were of smaller magnitude than those from the other models. The estimated neighbour effect of the NLM model was moderate in magnitude and statistically significant, as assessed by a $z$-test. The MMCP model showed a bit larger estimates (with small posterior standard deviation) for the treatment and neighbour effect variances than those from the MMI and CC models. The correlation estimate was reasonably low ($\rho =$.22) however with a large standard error (Table 5.4). All models MMCP, CC and MMI models showed almost identical DIC, probably leading one to choose the simpler CC model. Even the MMI model indicated only minor variance in neighbour treatment effects, and these two models showed good agreement on all other parameter estimates, contrasting the values of the MMCP model for the treatment and neighbour effect variances.

## 5.7 Discussion

In this study, we explored two statistical approaches to assess and account for the neighbour treatment effects in an incomplete block design, while accounting for block effects. Despite the relatively small dataset relative to the number of model effects/parameters, the simulation studies demonstrated a potential utility of these models in the investigated settings. As mentioned in Section 5.4, the models can easily be extended to other block sizes. They can also easily be programmed as logistic models for proportion data, although some numerical issues were experienced in fitting such models.

### 5.7.1 Non-linear mixed model

Results from the NLM model data indicate that the specific non-linear relationship between the outcome and the treatments and neighbours could be captured well enough by the neighbour treatment effect. However, some restrictions on the variation between treatments and tanks seemed to apply. Our results showed noisiness in NLM estimates especially in scenarios A.1 and B.1-2. These finding probably signal identifiability problems for such data. In the absence of treatment effects (i.e., if all $\beta_j = 0$ in (5.1)), the parameter $\delta$ is clearly not identifiable. Thus it is suggested that there should be a minimum variation between the treatments in order to be able to see and estimate the neighbour effect.

### 5.7.2 Cross-classified and multiple membership models

The MMI model performed fairly well in scenarios A.1 and B.1–2 with a weak link (or low correlation) between treatment and neighbour effects. In datasets with stronger dependence, the estimates of between-treatment and between-neighbour variances were substantially shrunk towards zero. It was also noted that the estimates were less variable than expected from the posterior standard devations for the NLM true model data, but this may be a result of the model misspecification. For MMI data (scenario B.1), the discrepancy was less marked and may be attributable to the fairly low sample size or sensitivity to prior distributions.

The apparent better performance of the MMI model in scenarios A.1 and B.1-2 over the MMCP model was shown to be caused by a sensitivity to the Wishart prior distribution (matrix) of the MMCP model. With a suitable prior, the MMCP model could reproduce the true values well in all scenarios. However, as the scenarios covered a range of variances, the same prior could not reproduce the true values exactly across all scenarios. This observed sensitivity to the Wishart prior is in agreement with previous work [5, 2] that biases in the estimates may arise from the prior especially in small datasets. The dependence on prior information raises the question how one should choose the prior in a practical application. The best answer we can offer is that one should always carry out a sensitivity analysis, and that it is often useful to try to center the prior distribution on values close to the estimates (possibly in an iterative fashion).

We also noted that the MMCP model requires a substantial correlation between treatment and neighbour effects to present an improvement over the MMI model. Given the fairly small dataset, this is not surprising. In conclusion, our results seem to demonstrate the utility of the MMCP extension of the standard multiple membership model.

### 5.7.3 ISAV data

The finding of an apparent significant neighbour effect in the NLM model was compromised by the results of the simulation studies at low levels of between-treatment and between-tank variation, in two ways. First, the NLM estimate of the neighbour effect tended to be very variable and on the average inflated (too large); second, its standard error tended to be substantially underestimated. These two findings cast so much doubt on the significant neighbour effect that it should probably be disregarded as a spurious effect. Moreover, the results showed only minor treatment effects and a between-tank variation that was wholly consumed by the neighbour effect. Even without any neighbour effect all variances were fairly small. The natural conclusion seems to be that the actual data did not exhibit values within a range where the NLM model could provide evidence of neighbour treatment effect.

This conclusion is supported by the results of analysis by the cross-classified and multiple membership models. The DIC model selection criterion pointed towards the cross-classified model with no neighbour treatment effect, and there was absolutely no evidence of the existence of a neighbour treatment correlated to the treatment effect itself. When faced with a negative (non-significant) finding, the question arises whether there was sufficient power in the data to detect any neighbour effect. The simulations from scenarios including moderate neighbour effects indicate that a moderate neighbour effect could have been detected from the data. Apparently, such an effect was just not present.

## 5.8 References

## References

[1] Azais, J. M., Bailey, R. A., Monod, H., 1993. A catalogue of efficient neighbour designs with border plots. *Biometrics* **49**, 1253–1261.

[2] Browne, W. J., 2006. Discussion on the paper by Pardoe and Weidner. *Journal Statistical Planning and inference* **136**, 1462–1465.

[3] Browne, W. J., 2003. *MCMC Estimation in MLwiN, version 2.0.* Centre of Multilevel Modelling, Institute of Education, University of London.

[4] Browne, W. J., Draper, D., 2006. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis* **3**, 473–514.

[5] Browne, W. J., Draper, D., 2000. Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics* **15**, 391–420.

[6] Browne, W. J., Goldstein, H., Rasbash, J., 2001. Multiple membership multiple classification (MMMC) models. *Statistical Modelling* **1**, 103–124.

[7] Dean, A., Voss, D., 1999. *Design and Analysis of Experiments.* Springer-Verlag New York, Inc.

[8] Dohoo, I. R., Martin, S. W., Stryhn, H., 2003. *Veterinary Epidemiologic Research.* AVC Inc., Charlottetown, Canada; web-site: `http://www.upei.ca/ver`.

[9] Durban, M., Currie, I. D., Kempton, R. A., 2001. Adjusting for fertility and competition in variety trials. *Journal of Agricultural Science* **136**, 129–140.

[10] Formann, A. K., 1993. Latent class model diagnosis from a frequentist point of view. *Biometrics* **59**, 189–196.

[11] Ganju, J., 2000. On choosing between fixed and random block effects in some no-interaction models. *Journal of Statistical Planning and Inference* **90**, 323–334.

[12] Giesbrecht F. G., 1986. Analysis of data from incomplete block designs. *Biometrics* **42**, 437–448.

[13] Gill, J., 2008. Is partial-dimension convergence a problem for inferences from MCMC algorithms?. *Political Analysis* **16**, 153–178.

[14] Golaszewski, J., Stawiana-Kosiorek, A., Zaluski, D., Zaręba, A., 2005. Competition effects in plant breeding field trials with pea (*Pisum sativum L.*), faba bean (*Vicia faba L.*) and yellow lupin

(*Lupinus luteus L.*). *Electronic Journal of Polish Agricultural Universities* **8**, 29.

[15] Hill, P. W., Goldstein, H., 1998. Multilevel modelling of educational data with cross-classification and missing identification of units. *Journal of Educational and Behavioral Statistics* **23**, 117–28.

[16] Langford, I., Leyland, A., Rasbash, J., Goldstein, H., 1999. Multilevel modelling of the geographical distributions of diseases. *Applied Statistics* **48**, 253–268.

[17] Leckie, G., 2008. Modelling the effects of pupil mobility and neighbourhood on school differences in educational achievement. University of Bristol Working Paper Series **189**.

[18] Lovy, J., Speare, D. J., Stryhn, H., Wright, GM., 2008. Effects of dexamethasone on host innate and adaptive immune responses and parasite development in rainbow trout On corhynchus mykiss infected with Loma salmonae. *Fish and Shellfish Immunology* **24**, 649–658.

[19] Jansen, P. A., Slettvold, H., Finstad, A. G., Langeland, A., 2002. Niche segregation between Arctic char (*Salvelinus alpinus*) and brown trout (*Salmo trutta*): an experimental study of mechanisms. *Canadian Journal of Fisheries and Aquatic Sciences* **1**, 6–11.

[20] Juliano, S., 1998. Adjusting for species introduction and replacement amongst mosquitoes: interspecific resource competition or apparent competition?. *Ecology* **79**, 255–268.

[21] McMahon, T. E., Zale, A. V., Barrows, F. T., Selong, J. H., Danehy, R. J., 2007. Temperature and competition between bull trout and brook trout: A test of the elevation refuge hypothesis. *Transactions of the American Fisheries Society* **136**, 1313–1326.

[22] Paulino, C. D. M., Pereira, C. A. B., 1994. On identifiability of parametric statistical models. *Journal of the Italian Statistical Society* **3**, 125–151.

[23] Pinheiro, J. C., Bates, D. M., 1995. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics* **4**, 12–35.

[24] Rasbash, J., Goldstein, H., 1994. Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model. *Journal of Educational and Behavioral Statistics* **19**, 337–350.

[25] Speare, D. J., MacNair, N., Hammell, K. L., 1995. Demonstration of tank effect on growth indices of juvenile rainbow trout (*Oncorhynchus mykiss*) during an ad libitum feeding trial. *American Journal of Veterinary Research* **56**, 1372–9.

[26] Tjur, T., 2003. A warning concerning random effects and random coefficients in logistic regression models for binary data. Dept. of Management Science and Statistics, Copenhagen Business School. Manuscript, available from `http://www.cbs.dk/staff/tuetjur`.

[27] Wolfinger R. D., 1999. Fitting nonlinear mixed models with the new `NLMIXED` procedure. Cary, NC: SAS Institute, Paper **287**.

Table 5.1: Study design and Mortality proportions (based on 50 fish per group) at the end of the follow-up period in a vaccine (ISAV) trial on Atlantic salmon carried out at the Atlantic Veterinary College.

| | Tank | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Treatment | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 18 | 19 | 22 |
| 1 | | | | | | | | | .44 | .34 | | | | .56 |
| 2 | | | .56 | .60 | | | | .60 | | | | | | |
| 3 | | | | | | | .54 | .46 | | | | .58 | | |
| 4 | .48 | .48 | | | | .42 | | | | | | | | |
| 5 | | | | | .34 | .52 | | | | .32 | | | | |
| 6 | | | | | | | | .30 | .42 | | | | .20 | |
| 7 | | .40 | .34 | | | | .52 | | | | | | | |
| 8 | | | | .46 | | | | | | | | | .30 | .48 |
| 9 | | | | .50 | .36 | | | | .42 | | | | | |
| 10 | .52 | | | | | | | | | .42 | .62 | | | |
| 11 | | | .62 | | | | | | | | | .66 | .56 | |
| 12 | .66 | | | | .50 | | | | | | | | | .50 |
| 13 | | | | | | .64 | .48 | | | | .60 | | | |
| 14 | | .72 | | | | | | | | | .72 | .70 | | |

Table 5.2: Mean parameter estimates followed in parenthesis by standard deviation (SD) among simulations and mean standard error/posterior standard deviation (SE) of non-linear mixed model (NLM) and two cross-classified and multiple membership models (MMI and MMCP the former with $\rho = 0$), based on analyses of 1000 simulated datasets generated by **non-linear mixed model (NLM )**. Parameters: $\mu$ (overall mean), $\sigma_b$, $\sigma_e$, $\sigma_t$, $\sigma_n$ (standard deviation between blocks, observations, treatments, neighbour treatments), $\delta$ (neighbour treatment effect), $\rho$ (correlation between treatment and neighbour effects), $\Delta$DIC (DIC from MMCP model - DIC from MMI model).

| Scen- | Param- | True | NLM model | MMI model | MMCP model |
|---|---|---|---|---|---|
| ario | eter | Values | Mean (SD, SE) | Mean (SD, SE) | Mean (SD, SE) |
| A.1 | $\mu$ | .50 | .500 (.029, .017) | .500 (.029, .048) | .499 ( .029, .079) |
| | $\sigma_b$ | .10 | .021 (.042, .026) | .106 (.032, .037) | .071 ( .021, .037) |
| | $\sigma_e$ | .10 | .082 (.015, .009) | .097 (.016, .017) | .094 ( .016, .016) |
| | $\sigma_t$ | .10 | .229 (.029, –[a] ) | .076 (.023, .030) | .145 ( .016, .036) |
| | $\delta$ | .30 | .513 (.400, .161) | | |
| | $\sigma_n$ | (.06) | | .065 (.014, .038) | .166 ( .018, .049) |
| | $\rho$ | | | 0 | .367 ( .125, .280) |
| | $\Delta$DIC | | | | .466 (2.481) |
| A.2 | $\mu$ | .50 | .502 (.131, .111) | .506 (.131, .193) | .502 ( .131, .266) |
| | $\sigma_b$ | .50 | .399 (.163, .087) | .546 (.122, .138) | .411 ( .142, .162) |
| | $\sigma_e$ | .10 | .087 (.050, .011) | .100 (.018, .020) | .101 ( .019, .020) |
| | $\sigma_t$ | .50 | .527 (.126, –[a] ) | .350 (.082, .091) | .463 ( .127, .137) |
| | $\delta$ | .30 | .296 (.239, .145) | | |
| | $\sigma_n$ | (.30) | | .126 (.046, .130) | .401 ( .143, .195) |
| | $\rho$ | | | 0 | .643 ( .254, .360) |
| | $\Delta$DIC | | | | .188 ( .559) |
| A.3 | $\mu$ | .50 | .502 (.131, .109) | .508 (.131, .200) | .506 ( .132, .301) |
| | $\sigma_b$ | .50 | .400 (.164, .092) | .651 (.140, .156) | .394 ( .171, .185) |
| | $\sigma_e$ | .10 | .080 (.030, .012) | .100 (.018, .020) | .100 ( .018, .020) |
| | $\sigma_t$ | .50 | .527 (.129, –[a] ) | .246 (.061, .070) | .448 ( .143, .146) |
| | $\delta$ | .50 | .480 (.180, .112) | | |
| | $\sigma_n$ | (.50) | | .106 (.020, .108) | .517 ( .209, .236) |
| | $\rho$ | | | 0 | .804 ( .166, .283) |
| | $\Delta$DIC | | | | –.830 (1.020) |

[a] Not estimated because treatments modelled by fixed effects

Table 5.3: Mean parameter estimates followed in parenthesis by standard deviation (SD) among simulations and mean standard error/posterior standard deviation (SE) of non-linear mixed model (NLM) and two cross-classified and multiple membership models (MMI and MMCP the former with $\rho = 0$), based on analyses of 1000 simulated datasets generated by **cross-classified and multiple membership models (MMCP, MMI)**. Parameters: $\mu$ (overall mean), $\sigma_b$, $\sigma_e$, $\sigma_t$, $\sigma_n$ (standard deviation between blocks, observations, treatments, neighbour treatments), $\delta$ (neighbour treatment effect), $\rho$ (correlation between treatment and neighbour effects), ΔDIC (DIC from MMCP model - DIC from MMI model).

| Scen- | Param- | True | NLM model | MMI model | MMCP model |
|---|---|---|---|---|---|
| ario | eter | Values | Mean (SD, SE) | Mean (SD, SE) | Mean (SD, SE) |
| B.1 | $\mu$ | .50 | .496 (.049, .024) | .497 (.049, .055) | .496 ( .049, .078) |
| | $\sigma_b$ | .10 | .060 (.058, .026) | .095 (.031, .038) | .073 ( .022, .038) |
| | $\sigma_e$ | .10 | .086 (.023, .010) | .102 (.018, .020) | .097 ( .016, .017) |
| | $\sigma_t$ | .10 | .221 (.032, –[a] ) | .093 (.029, .034) | .148 ( .017, .036) |
| | $\delta$ | | .106 (.968, .326) | | |
| | $\sigma_n$ | .10 | | .101 (.037, .053) | .177 ( .021, .052) |
| | $\rho$ | .00 | | 0 | .193 ( .155, .302) |
| | ΔDIC | | | | −1.462 (1.766) |
| B.2 | $\mu$ | .50 | .496 (.052, .023) | .497 (.052, .054) | .495 ( .052, .079) |
| | $\sigma_b$ | .10 | .053 (.059, .026) | .102 (.033, .039) | .074 ( .022, .038) |
| | $\sigma_e$ | .10 | .016 (.087, .010) | .101 (.018, .019) | .096 ( .016, .017) |
| | $\sigma_t$ | .10 | .223 (.034, –[a] ) | .087 (.027, .033) | .147 ( .016, .036) |
| | $\delta$ | | .201 (.903, .282) | | |
| | $\sigma_n$ | .10 | | .093 (.032, .051) | .176 ( .021, .052) |
| | $\rho$ | .25 | | 0 | .255 ( .148, .296) |
| | ΔDIC | | | | −1.077 (1.900) |
| B.3 | $\mu$ | .50 | .481 (.263, .135) | .486 (.263, .227) | .488 ( .263, .257) |
| | $\sigma_b$ | .50 | .475 (.240, .109) | .584 (.155, .168) | .538 ( .156, .192) |
| | $\sigma_e$ | .10 | .074 (.128, .013) | .101 (.019, .020) | .101 ( .019, .021) |
| | $\sigma_t$ | .50 | .495 (.161, –[a] ) | .410 (.106, .115) | .436 ( .111, .137) |
| | $\delta$ | | .004 (.688, .332) | | |
| | $\sigma_n$ | .50 | | .262 (.162, .208) | .375 ( .113, .214) |
| | $\rho$ | .50 | | 0 | .193 ( .400, .470) |
| | ΔDIC | | | | .360 ( .298) |

[a] Not estimated because treatments modelled by fixed effects

Table 5.4: Parameter estimates and associated standard errors (SE) or posterior distribution standard deviation (SD) for non-linear mixed model (**NLM**) and three cross-classified and multiple membership models (**CC, MMI, MMCP** with increasing level of neighbouring effects), from analysis of the **ISAV dataset**. Parameters: $\mu$ (overall mean), $\sigma_b$, $\sigma_e$, $\sigma_t$, $\sigma_n$ (standard deviation between blocks, observations, treatments, neighbour treatments), $\delta$ (neighbour treatment effect), $\rho$ (correlation between treatment and neighbour effects) and $\beta$'s

| Param- | NLM model | CC model | MMI model | MMCP model |
|---|---|---|---|---|
| eter | Est. (SE) | Est. (SD) | Est. (SD) | Est. (SD) |
| $\mu$ | .497 (.010) | .497 (.033) | .497 (.036) | .498 (.062) |
| $\sigma_b$ | .001 ( –[a] ) | .064 (.020) | .058 (.021) | .055 (.024) |
| $\sigma_e$ | .061 (.007) | .063 (.012) | .060 (.011) | .060 (.011) |
| $\sigma_t$ | .071 ( –[b] ) | .095 (.025) | .091 (.026) | .140 (.032) |
| $\beta_1$ | −.006 (.042) | | | |
| $\beta_2$ | .117 (.037) | | | |
| $\beta_3$ | .012 (.036) | | | |
| $\beta_4$ | −.016 (.038) | | | |
| $\beta_5$ | −.136 (.038) | | | |
| $\beta_6$ | −.197 (.035) | | | |
| $\beta_7$ | −.128 (.039) | | | |
| $\beta_8$ | −.086 (.035) | | | |
| $\beta_9$ | −.037 (.037) | | | |
| $\beta_{10}$ | −.013 (.040) | | | |
| $\beta_{11}$ | .091 (.036) | | | |
| $\beta_{12}$ | .084 (.036) | | | |
| $\beta_{13}$ | .109 (.039) | | | |
| $\beta_{14}$ | .206 (.037) | | | |
| $\delta$ | .280 (.110) | | | |
| $\sigma_n$ | | 0 | .049 (.025) | .134 (.033) |
| $\rho$ | | 0 | 0 | .223 (.286) |
| DIC | | −92.603 | −92.584 | −92.281 |

[a] No standard error available because estimate is on the boundary.

[b] Not estimated because treatments modelled by fixed effects

# Conclusion

## 6.1 Introduction

The objective of this research project was to assess the performance of statistical procedures for the analysis of binary longitudinal data in veterinary science, specifically, to describe and quantify their performance in terms of statistical properties such as unbiasedness, confidence interval coverage and efficiency. We identified procedures belonging to two model types for the assessment: marginal and random effects models. These models handle the within-subject dependence differently, and they offer different interpretations of regression estimates for binary longitudinal data. In order to achieve the objective, we set up a general structure for studies to examine the characteristics of these procedures. A statistical simulation approach was used as the tool for the assessment.

The objective of the first study was to give a detailed description of the choice between marginal and random effects models and procedures in a full binary repeated measures data setting (Chapter 2). The second study objective was to compare statistical procedures in a full binary repeated measures data setting with additional hierarchical data structure (Chapter 3). The objective of the third study was to assess the impact of a combination of different missing data patterns on selected statistical procedures described in the second study (Chapter 4). Finally, the objective of the last study was to develop two statistical approaches to model neighbour effects in an aquaculture clinical trials setting (Chapter 5).

In this final chapter, we summarize the current knowledge in modelling of binary longitudinal data. Specifically, to provide some general guidelines for the choice between marginal and random effects models. We also highlight some innovation and limitations of this research project, and finally, identify areas of potential future research and some possible directions in this area.

## 6.2 State of knowledge in modelling binary longitudinal data

The following discussion reflects the knowledge and experience of modelling binary longitudinal data, based on the findings extracted from the simulation studies carried out in this project. Both marginal and random effects estimation procedures were assessed, and among the procedures included were: ordinary logistic regression (OLR), generalized estimating equations (GEE), Weighted Generalized Estimating Equations (WGEE), alternating logistic regression (ALR), procedures based on pseudo- or quasi-likelihood (REPL, MQL and PQL respectively), Markov chain Monte Carlo (MCMC) and maximum likelihood estimation (ML). These procedures were examined in a fairly wide range of correlated binary data settings including a two-level balanced longitudinal design, a three-level balanced setting of binary repeated measures data, and repeated measures data with missing values. Three types of missing values patterns were considered; missing completely at random (MCAR); missing at random (MAR); not missing at random (NMAR).

The following sections will discuss some issues for the choice between models and procedures. The issues will be converted into a set of practical guidelines, based, in part, on the literature but primarily on the findings of the thesis.

### 6.2.1 Guidelines for the choice between marginal and random effects models

The random effects model was used to create the simulated datasets (Chapters: 2, 3, 4), additionally, a marginal model was used to create some of the simulated datasets in the first study (Chapter 2). A between-subjects design was considered throughout the simulation studies, additionally in Chapter 2 the dichotomous treatment was modelled either within subjects, or by a time interaction. All random effects procedures under study here, excluding REPL procedure, make the conceptually unreasonable assumption that residual correlations are constant over time, the question for application of such random effects procedures is the sensitivity of the results to that assumption.

1: For the marginal model data with either the within-subject design or interaction design (Chapter 2), the random effects procedures displayed severe deficiencies in terms of both efficiency and CI coverage, which increased with the size of the dataset and the true autocorrelation. For the between-subject design with a small data size, all marginal estimation procedures experienced problems with CI undercoverage and biased estimates, whereas the random effects procedures showed a minor loss of efficiency.

2: For the random effects model data with additional hierarchical structure (Chapter 3), the quasi-likelihood random effects procedures showed some attenuation of regression and variance parameters, the inclusion of an extra-binomial parameter in these methods did not clearly improve their performance. For autoregressive data, the random effects procedures performed poorly, therefore marginal procedures may seem more attractive.

3: For the random effects model data with additional hierarchical structure and missing values (Chapter 4), although the focus of this study was the impact on missing values, we concluded here with some findings that could be of help in the choice between marginal and random effects models. For autoregressive data with drop-outs missing at random, marginal estimation procedures performed well, with up to moderate percentages of missing values. The likelihood-based procedures performed well only for the random intercept models data, whereas, the quasi-likelihood method resulted in substantially biased estimates (Chapter 4).

4: The size of the data is controlled by the length of the time series and the number of replicated subjects. Results demonstrated that a small number of subjects with a short time series proved to be a challenge for both marginal and random effects methods. However, random effects procedures may be acceptable for some small datasets that do not guarantee asymptotic properties for marginal methods (Chapter 2).

5: The relationship between random effects and marginal estimates has been discussed and described previously [8, 5]; see also the summary by Diggle *et al.* [2]. For logistic regression, it has resulted in an approximate conversion formula proposed by Zeger *et al.* [8]. A simulation study (Chapter 3) carried out in this thesis showed that this formula is the possible source of a small general bias. Therefore, based on the findings of the thesis, for a marginal (population-averaged) estimates/interpretation, the marginal procedures may seem more attractive, especially in a situation with decaying correlation over time. However, for a random effects (subject-specific) estimate/interpretation, the marginal estimation procedures are of little use (between-subjects variance is not known).

### 6.2.2 Guidelines for the choice among marginal procedures

Generally, the semi-parametric marginal estimation procedures have to their credit the robustness implicit in making no specific assumptions about random effects and correlation structure. However, the choice between the procedures included in the current study could be highlighted in the view of the results.

1: For the classical two-level settings in repeated measures data (Chapter 2), the autoregressive GEE remained highly efficient in all settings. The estimates of GEE with either exchangeable or independence correlation structures, ALR and MQL procedures agreed closely; however, in the within-subject design, their relative efficiency dropped down dramatically for the longer series with high correlation. Additionally, the MQL procedure suffered from substantial undercoverage for longer series with moderate to high correlation.

2: For repeated measures data with additional hierarchical structure (Chapter 3), a version of GEE with either independence or exchangeable correlation at the cluster-level was evaluated and showed to perform similarly to ALR procedure and generally well across the range of settings covered. All other attempts to incorporate the additional hierarchical level into the GEE framework produced estimates with serious deficiencies for some of the fixed effects parameters. The MQL method showed some fluctuation in the standard error for the time coefficient, but generally performed on par with ALR method.

3: For repeated measures incomplete data with additional hierarchical structure (Chapter 4), both ALR and WGEE with either independent or exchangeable correlation at the cluster-level, performed well at a low (31%) proportion of missing values at random regardless of the correlation structure in the data. Additionally, ALR showed some robustness against the combination of patterns of the missing values and for missing values not at random (except for the time coefficient) regardless of the correlation structure in the data.

### 6.2.3 Guidelines for the choice among random effects procedures

Generally, one advantage of using random effects procedures is the ability to model and predict effects at the individual level. However, in situations with decaying correlation over time, the random effects procedures failed to reproduce the subject-specific value. For this situation we cannot point to any procedures among those covered in the study to obtain subject-specific estimates with acceptable performance. Here the focus is on and highlights some issues that might help with the choice between random effects procedures in view of the results.

1: For the classical two-level settings in repeated measures (Chapter 2), all the random effects procedures (except REPL) performed well in the data generated from random intercept models. The REPL method performed mostly as a marginal estimation procedure, and showed no promise for estimation of the variance and autoregressive parameter in the autoregressive random effects data.

2: For random intercept models data with additional hierarchical structure (Chapter 3), the likelihood-based random effects procedures performed better than methods based on quasi- or pseudo-likelihood.

3: The REPL procedure demonstrated poor performance for repeated measures with additional hierarchical structure data performed poorly in such settings.

4: The additional hierarchical structure challenges some statistical procedures, for example, one procedure (ML) may involve an extensive and time consuming computation for estimating the model parameters.

5: For repeated measures incomplete data with additional hierarchical structure (Chapter 4), both likelihood-based approximations methods (ML, MCMC) demonstrated that the accuracy of the approximations were sufficient to, by and large, ensure the ignorability of missing completely at random and drop-out missing values at random. The penalized quasi-likelihood procedures demonstrated a bias in the estimates for drop-out missing values either at random or not at random.

## 6.3 Simulation as a tool for modelling repeated measures and hierarchical data

Statistical simulation showed itself to be an effective approach for studying repeated measures and hierarchical data. By this approach we were able to explore the properties of different models/procedures and their ability to hold these properties under the study settings. Here we point to the additional hierarchical structure and the missing data that commonly arise in longitudinal data.

1: By the simulation approach we demonstrated the ability of the autoregressive random effects model to simulate binary repeated measures data with additional hierarchical structure. A marginal model for the same data structure was complicated and not easy to set up.

2: Similarly, by simulation we illustrated the repeated measures random effects model to simulate different patterns of missing data. The finding from this thesis showed the ability of this model to study the impact of missing values, especially for combination of missing data patterns within the same dataset (Chapter 4).

3: Through a targeted simulation study to a specific dataset, we were able to explore the two proposed statistical approaches for modelling neighbour treatment effects in aquaculture clinical trials (Chapter 5).

## 6.4 Study innovation

Statistical simulation was known to provide solid evidence for the statistical assessment of the properties of statistical models. In the current thesis, the simulation approach was used to highlight and assess some properties of marginal and random effects models for analysis of binary longitudinal data. The approach of matching the simulated data structure to the data at hand, as closely as possible can be helpful to provide much insight into which procedures provide the accurate answers. By simulation studies, we illustrated the use of autoregressive random effects model (Chapters: 2, 3, 4, 5) for simulating binary autocorrelated data in a wide range of settings including balanced and incomplete binary longitudinal data.

By simulation, two approaches based on random effects model, were explored and showed to be sufficient for modelling and estimating the neighbour treatments effects(Chapter 5). In this thesis we demonstrated a simple simulation approach to study the impact of combination of different types of missing values within the same dataset. By this approach we were able to show some of the limitations of marginal and random effects estimation procedures for analysis of incomplete binary longitudinal data.

Among the specific findings of the current thesis are:

1: Throughout the thesis we demonstrated that for autocorrelated data, the random effects procedures performed poorly and failed to reproduce the subject-specific value (Chapters: 2, 3, 4).

2: MQL method performed as a marginal procedure but tended to underestimate the standard errors of fixed effects (Chapter 2).

3: In the classical two-level repeated measures data with within-subject design, the relative efficiency of ALR and MQL procedures decreased dramatically for the longer series with high correlation (Chapter 2).

4: The REPL procedure for repeated measures with additional hierarchical structure data demonstrated a bias for the estimates of both the regression and the variance parameters (Chapter 3).

5: The logistic conversion formula from subject-specific parameters may be a possible source of a small general bias (Chapter 3).

6: Two simulation studies showed that the extra-binomial parameter estimates in quasi- or pseudo-likelihood were associated with inflated standard errors (Chapters 3, 4)

7: In a 3-level data structure, the GEE handling of correlation structure must be shifted from the subject to the cluster to achieve correct inference at the cluster level (Chapter 3).

8: The ALR procedure was shown to be robust against the combinations of patterns of missing values, moderate proportion of missing at random and missing values not at random (except for the time coefficient) regardless of the correlation structure in the data (Chapter 4).

9: The weighted GEE procedure with either independence or exchangeable correlation at the cluster level was shown to be robust for a moderate proportion of missing values at random (Chapter 4).

10: One study demonstrated a potential bias in using penalized quasi-likelihood procedures for the analysis of an incomplete dataset with drop-out missing values at random (Chapter 4).

11: A targeted simulation study demonstrated a potential usage of the non-linear mixed model and the cross-classified and multiple membership models in modelling neighbour treatment effects (Chapter 5).

## 6.5 Study limitations

One limitation of the study was the absence of a real dataset, especially for the classical two-level settings in repeated measures. However, the simulation study settings (Chapter 2) covered a wide range of binary longitudinal data settings in veterinary science. These limitations might give the impression that the statistical procedures were not matched closely enough to data arising from veterinary science. However, this can also be seen as an advantage, because longitudinal binary data occur in many fields. Another limitation may be that the simulation studies were mainly set up for experimental data, whereas in practice observational studies are common.

The limitation of the study to only include procedures implemented in broadly accessible statistical software, could be taken as a disadvantage as some statistical models/methods were excluded, such as the pattern-mixture models for incomplete data [6]; the marginalized models [3]; the transition model [2]; the multivariate approach [7]; the approach proposed by Barbosa and Goldstein [1] to model correlations between lowest level residuals, conditional upon the random effects, by an autoregressive function of time. However, the argument was made in Chapter 1 that the range of procedures included should reflect the choice an applied researcher faces when it comes to data analysis.

Another limitation was the lack of a reference estimates for the autoregressive model. We tried to fit the model by MCMC estimation but could not achieve acceptable trajectories of the resulting Markov chains. This had two consequences: First, we could not point to any acceptable random effects estimation procedure in the presence of autocorrelation. Second, we were unable to compute efficiencies (relative to reference estimates) for the random effects data (Chapters 2,3,4).

## 6.6 Future directions for additional research

### 6.6.1 Confirmation/expansion of findings

Based on Finding 5 (Section 6.4) we suggest that future research should include ways to assess and improve the relationship between marginal and random effects models. One idea is through theoretical research to confirm and improve the logistic conversion formula. Another idea is through new statistical modelling, where the marginalized models [3] showed to be a promising approach to overcome some of the limitations experienced by marginal and random effects models.

Finding 4 indicates that further research may be needed to assess the accuracy and validity of the REPL procedure. Regarding Finding 6, we recommend more research to confirm and justify the usefulness of the extra-binomial parameter as a diagnostic tool.

### 6.6.2 New ideas or suggestions

Based on Finding 2, we propose to add robust ("sandwich") variance estimation to the MQL procedure (Chapter 2). Based on Finding 10, we suggest that theoretical research could be needed to explain the poor performance of the penalized quasi-likelihood procedures in data with values missing at random. One idea could be based on the similarity of the PQL procedure to GEE and may indicate a potential weighting scheme for missingness at random. However, this step requires additional theoretical and applied research.

Research into methods to account for a combination of missing data patterns within the same dataset is proposed, because this situation is a challenge for many of the simple approaches. Some statistical approaches such as the available case method and imputations may be limited to a strong MCAR missingness assumption. Other approaches (WGEE) are in their current implementation available to only the MAR missingness mechanism, whereas likelihood inference based on the available data may accommodate MCAR and MAR but not the NMAR missingness mechanism.

### 6.6.3 Research into limitations

We recommend the implementation of the autoregressive repeated measures random effects model through the Bayesian framework using MCMC methods. We, also recommend further exploration of the following models for binary longitudinal data: the marginalized model [3] and the transition model [2].

## 6.7 References

## References

[1] Barbosa, B., Goldstein, H., 2000. Discrete multilevel response models. *Quality and Quantity* **34**, 323–330.

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

[3] Griswold, M. E., Zeger, S. L., 2004. On marginalized multilevel models and their computation. *Johns Hopkins University, Dept. of Biostatistics, Working Papers No.* **99**.

[4] Heagerty, P. J., Zeger, S. L., 2000. Marginalized multilevel models and likelihood inference. *Statistical Science* **15**, 1–19.

[5] Neuhaus, J. M., 1992. Statistical methods for longitudinal and clustered design with binary responses. *Statistical Methods in Medical Research* **1**, 249–273.

[6] Roderick, J. A., Little., 1993. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* **88**, 125–134.

[7] Yang, M., Goldstein, H., Heath, A., 2000. Multilevel models for repeated binary outcomes: attitudes and voting over the electoral cycle. *Journal of the Royal Statistical Society, Series A* **163**, 49–62.

[8] Zeger, S. L., Liang, K. Y., Albert, P. S., 1988. Models for longitudinal data - a generalized estimating equation approach. *Biometrics* **44**, 1049–1060.

# Appendix A

# Additional tables of Chapter 2

Table A.1: Mean estimate of between-subjects (**BS**) treatment effect (true value = 0.35), followed in parenthesis by standard deviation among simulations, mean standard error, confidence interval coverage and relative efficiency, based on analyses of 1000 simulated marginal (**PA**) datasets per setting ($n$ = number of subjects, $t$ = number of time points, $\rho$ = autocorrelation). Analysis by procedure B of type $A$ is designated by $\hat{\beta}^{A}_{\mathrm{B}}$, where $A = PA$ (population-averaged) or $SS$ (subject-specific), and B = IND (generalized estimating equations (GEE) with independence correlation), EXCH (GEE with exchangeable correlation), REPL (marginal restricted pseudo-likelihood) MLa (maximum likelihood based on Gauss Hermite-quadrature in R), MLb (maximum likelihood based on adaptive quadrature in Stata).

| | | | Statistical Methods[T] | | | | |
|---|---|---|---|---|---|---|---|
| $n$ | $t$ | $\rho$ | $\hat{\beta}^{PA}_{\mathrm{IND}}$ | $\hat{\beta}^{PA}_{\mathrm{EXCH}}$ | $\hat{\beta}^{PA}_{\mathrm{REPL}}$ | $\hat{\beta}^{SS}_{\mathrm{MLa}}$ | $\hat{\beta}^{SS}_{\mathrm{MLb}}$ |
| 100 | 16 | .7 | .360 (.23,.23) .95 .95 | .360 (.24,.24) .96 .87 | .359 (.22,.23) .96 1.00 | .379 (.24,.24) .95 .83 | .381 (.24,.24) .95 .83 |
| | | .5 | .360 (.17,.18) .96 .98 | .359 (.18,.18) .96 .96 | .359 (.17,.18) .96 1.00 | .370 (.18,.18) .95 .91 | .370 (.18,.18) .95 .91 |
| | | .2 | .352 (.13,.13) .95 1.00 | .352 (.13,.13) .95 1.00 | .352 (.13,.13) .96 1.00 | .355 (.13,.28) .95 .98 | .355 (.13,.13) .95 .98 |
| | 8 | .7 | .342 (.28,.28) .96 .93 | .341 (.28,.28) .96 .92 | .344 (.27,.28) .96 1.00 | .376 (.31,.30) .94 .75 | .378 (.31,.30) .94 .76 |
| | | .5 | .348 (.23,.23) .95 .96 | .348 (.23,.23) .95 .95 | .307 (.21,.23) .96 1.00 | .366 (.24,.24) .95 .87 | .366 (.24,.24) .95 .87 |
| | | .2 | .349 (.17,.17) .95 .96 | .349 (.17,.17) .95 .95 | .337 (.17,.17) .96 1.00 | .355 (.17,.18) .95 .96 | .355 (.17,.17) .95 .96 |
| | 4 | .7 | .341 (.35,.33) .94 .95 | .341 (.35,.33) .94 .94 | .348 (.34,.33) .95 1.00 | .381 (.41,.37) .93 .69 | .380 (.40,.36) .93 .71 |
| | | .5 | .343 (.30,.29) .94 .95 | .344 (.30,.29) .94 .95 | .340 (.30,.29) .94 1.00 | .366 (.32,.31) .95 .85 | .366 (.32,.31) .95 .85 |
| | | .2 | .340 (.24,.23) .95 .99 | .340 (.24,.23) .95 .99 | .339 (.24,.24) .95 1.00 | .350 (.25,.24) .95 .94 | .350 (.25,.24) .95 .94 |
| 20 | 16 | .7 | .351 (.55,.51) .92 .94 | .351 (.58,.53) .92 .85 | .351 (.54,.52) .94 1.00 | .351 (.57,.49) .89 .87 | .371 (.58,.53) .92 .86 |
| | | .5 | .348 (.41,.38) .92 .97 | .347 (.41,.38) .92 .95 | .347 (.40,.40) .95 1.00 | .357 (.42,.39) .92 .93 | .357 (.42,.39) .92 .93 |
| | | .2 | .347 (.29,.27) .92 .99 | .347 (.29,.27) .92 .99 | .347 (.29,.29) .95 1.00 | .349 (.30,.30) .93 .98 | .349 (.30,.28) .93 .98 |
| | 8 | .7 | .381 (.66,.63) .93 .91 | .381 (.68,.64) .93 .88 | .378 (.63,.65) .96 1.00 | .341 (.72,.54) .82 .78 | .408 (.70,.60) .90 .82 |
| | | .5 | .373 (.53,.50) .92 .94 | .374 (.53,.50) .93 .93 | .372 (.51,.52) .95 1.00 | .387 (.55,.52) .93 .87 | .387 (.55,.52) .93 .87 |
| | | .2 | .364 (.40,.37) .92 .99 | .365 (.40,.37) .92 .98 | .364 (.39,.40) .95 1.00 | .369 (.40,.38) .93 .96 | .369 (.40,.38) .93 .96 |
| | 4 | .7 | .406 (.83,.76) .94 .98 | .401 (.81,.77) .95 .98 | .412 (.80,.80) .96 1.00 | .369 (.87,.97) .84 .88 | .401 (.86,.73) .90 .92 |
| | | .5 | .394 (.71,.66) .93 .98 | .395 (.71,.66) .93 .97 | .402 (.71,.69) .95 1.00 | .409 (.76,.85) .94 .85 | .409 (.74,.70) .94 .91 |
| | | .2 | .379 (.57,.52) .91 .99 | .379 (.57,.52) .91 .99 | .381 (.57,.55) .94 1.00 | .388 (.58,.55) .93 .95 | .388 (.58,.55) .93 .95 |

[T] Note that SS estimates were converted to PA value (see text).

Table A.2: Mean estimate of within-subjects (**WS**) treatment effect (true value = 0.35), followed in parenthesis by standard deviation among simulations, mean standard error, confidence interval coverage and relative efficiency, based on analyses of 1000 simulated marginal (**PA**) datasets per setting ($n$ = number of subjects, $t$ = number of time points, $\rho$ = autocorrelation). See Table A.1 for coding of statistical methods.

| | | | Statistical Methods$^{\top}$ | | | | |
|---|---|---|---|---|---|---|---|
| $n$ | $t$ | $\rho$ | $\hat{\beta}^{PA}_{\text{IND}}$ | $\hat{\beta}^{PA}_{\text{EXCH}}$ | $\hat{\beta}^{PA}_{\text{REPL}}$ | $\hat{\beta}^{SS}_{\text{MLa}}$ | $\hat{\beta}^{SS}_{\text{MLb}}$ |
| 100 | 16 | .7 | .355 (.18,.18) .95 .55 | .356 (.18,.18) .95 .55 | .352 (.14,.14) .95 1.00 | .368 (.19,.10) .66 .51 | .369 (.19,.10) .66 .51 |
| | | .5 | .359 (.16,.16) .95 .79 | .359 (.16,.16) .95 .79 | .358 (.14,.14) .96 1.00 | .368 (.16,.10) .77 .74 | .368 (.16,.10) .77 .74 |
| | | .2 | .355 (.13,.12) .94 .96 | .355 (.13,.12) .94 .96 | .355 (.12,.12) .95 1.00 | .357 (.13,.11) .89 .94 | .357 (.13,.11) .89 .94 |
| | 8 | .7 | .349 (.19,.18) .94 .63 | .349 (.19,.18) .94 .63 | .350 (.15,.15) .95 1.00 | .371 (.20,.12) .76 .56 | .371 (.20,.12) .76 .56 |
| | | .5 | .353 (.19,.18) .94 .81 | .353 (.19,.18) .94 .81 | .326 (.17,.16) .95 1.00 | .370 (.20,.13) .80 .73 | .370 (.20,.13) .80 .73 |
| | | .2 | .355 (.16,.16) .95 .97 | .355 (.16,.16) .95 .96 | .339 (.15,.16) .97 1.00 | .361 (.17,.14) .93 .93 | .361 (.17,.14) .93 .93 |
| | 4 | .7 | .352 (.17,.17) .95 .86 | .352 (.17,.17) .95 .86 | .358 (.15,.15) .95 1.00 | .379 (.18,.15) .90 .74 | .379 (.18,.15) .90 .74 |
| | | .5 | .355 (.20,.19) .95 .92 | .355 (.20,.19) .95 .92 | .359 (.19,.19) .95 1.00 | .378 (.21,.18) .90 .81 | .378 (.21,.18) .90 .81 |
| | | .2 | .355 (.20,.21) .96 .98 | .355 (.20,.21) .96 .98 | .356 (.20,.21) .97 1.00 | .365 (.21,.20) .94 .92 | .365 (.21,.20) .94 .92 |
| 20 | 16 | .7 | .388 (.43,.40) .93 .53 | .387 (.43,.40) .93 .53 | .373 (.32,.32) .96 1.00 | .396 (.44,.22) .68 .50 | .401 (.44,.22) .68 .50 |
| | | .5 | .383 (.36,.35) .92 .81 | .383 (.36,.35) .92 .81 | .382 (.33,.33) .95 1.00 | .392 (.37,.23) .78 .78 | .392 (.37,.23) .78 .78 |
| | | .2 | .372 (.28,.27) .94 .97 | .372 (.28,.27) .94 .98 | .373 (.28,.28) .96 1.00 | .374 (.28,.24) .91 .96 | .374 (.28,.24) .91 .96 |
| | 8 | .7 | .358 (.42,.40) .93 .62 | .358 (.42,.40) .92 .62 | .356 (.33,.33) .95 .99 | .370 (.44,.27) .78 .57 | .375 (.44,.27) .78 .57 |
| | | .5 | .355 (.42,.40) .93 .81 | .355 (.42,.40) .93 .81 | .347 (.38,.38) .96 .99 | .371 (.44,.31) .83 .74 | .371 (.44,.31) .83 .74 |
| | | .2 | .350 (.37,.35) .92 .97 | .350 (.37,.35) .92 .97 | .351 (.36,.36) .95 1.00 | .356 (.37,.32) .91 .94 | .356 (.37,.32) .91 .94 |
| | 4 | .7 | .358 (.38,.37) .93 .82 | .357 (.38,.37) .93 .82 | .354 (.34,.35) .95 1.00 | .386 (.44,.61) .91 .62 | .360 (.40,.35) .94 .77 |
| | | .5 | .362 (.45,.43) .94 .94 | .362 (.45,.43) .94 .94 | .365 (.43,.43) .95 1.01 | .379 (.47,.41) .93 .84 | .379 (.47,.40) .93 .85 |
| | | .2 | .356 (.47,.46) .94 .98 | .356 (.47,.46) .94 .98 | .354 (.46,.47) .95 1.01 | .366 (.48,.45) .94 .93 | .366 (.48,.45) .94 .93 |

$^{\top}$ Note that SS estimates were converted to PA value (see text).

Table A.3: Mean estimate of between-subjects (**BS**) treatment effect (true value = 0.35, marginal true value = 0.302), followed in parenthesis by standard deviation among simulations, mean standard error and confidence interval coverage, based on analyses of 1000 simulated random effects (**SS**) datasets per setting ($n$ = number of subjects, $t$ = number of time points, $\rho$ = autocorrelation). See Table A.1 for coding of statistical methods.

| | | | Statistical Methods | | | | |
|---|---|---|---|---|---|---|---|
| $n$ | $t$ | $\rho$ | $\hat{\beta}^{PA}_{\text{IND}}$ | $\hat{\beta}^{PA}_{\text{EXCH}}$ | $\hat{\beta}^{PA}_{\text{REPL}}$ | $\hat{\beta}^{SS}_{\text{MLa}}$ | $\hat{\beta}^{SS}_{\text{MLb}}$ |
| 100 | 16 | 1 | .286 (.20,.19) .94 | .286 (.20,.19) .93 | .287 (.20,.12) .75 | .343 (.24,.23) .94 | .343 (.24,.23) .94 |
| | | .7 | .287 (.14,.13) .93 | .287 (.14,.13) .92 | .287 (.14,.12) .90 | .298 (.14,.14) .91 | .299 (.14,.14) .91 |
| | | .5 | .291 (.12,.12) .93 | .291 (.12,.12) .93 | .291 (.12,.11) .93 | .295 (.12,.12) .91 | .296 (.12,.12) .90 |
| | | .2 | .291 (.11,.11) .91 | .291 (.11,.11) .91 | .291 (.11,.11) .95 | .292 (.11,.11) .92 | .292 (.11,.11) .91 |
| | 8 | 1 | .300 (.22,.21) .94 | .300 (.22,.21) .94 | .297 (.21,.17) .88 | .361 (.26,.25) .95 | .361 (.26,.25) .95 |
| | | .7 | .292 (.18,.17) .95 | .292 (.18,.17) .95 | .306 (.17,.16) .93 | .313 (.19,.19) .94 | .313 (.19,.19) .94 |
| | | .5 | .293 (.16,.16) .94 | .293 (.16,.16) .94 | .305 (.16,.15) .94 | .304 (.17,.17) .94 | .303 (.17,.16) .94 |
| | | .2 | .296 (.15,.15) .94 | .296 (.15,.15) .94 | .306 (.15,.15) .95 | .300 (.15,.16) .93 | .299 (.15,.15) .93 |
| | 4 | 1 | .306 (.26,.25) .95 | .306 (.26,.25) .95 | .306 (.26,.23) .92 | .369 (.31,.30) .95 | .369 (.31,.30) .95 |
| | | .7 | .302 (.24,.23) .94 | .302 (.24,.23) .94 | .303 (.24,.22) .93 | .335 (.26,.26) .94 | .335 (.26,.26) .94 |
| | | .5 | .296 (.23,.22) .94 | .296 (.23,.22) .94 | .296 (.23,.22) .94 | .315 (.24,.24) .94 | .315 (.24,.23) .94 |
| | | .2 | .299 (.22,.21) .94 | .299 (.22,.21) .94 | .300 (.22,.21) .94 | .307 (.22,.23) .94 | .307 (.22,.22) .94 |
| 20 | 16 | 1 | .303 (.43,.42) .94 | .305 (.43,.42) .94 | .302 (.43,.28) .78 | .358 (.51,.49) .93 | .357 (.51,.49) .93 |
| | | .7 | .292 (.30,.28) .92 | .292 (.30,.28) .92 | .292 (.30,.26) .91 | .301 (.31,.29) .92 | .301 (.31,.29) .92 |
| | | .5 | .284 (.27,.25) .93 | .284 (.27,.25) .93 | .284 (.27,.25) .94 | .289 (.27,.27) .93 | .288 (.27,.26) .93 |
| | | .2 | .278 (.25,.23) .91 | .278 (.25,.23) .92 | .278 (.25,.24) .94 | .280 (.25,.25) .93 | .280 (.25,.25) .93 |
| | 8 | 1 | .282 (.48,.46) .94 | .282 (.48,.46) .94 | .280 (.48,.38) .89 | .333 (.57,.55) .93 | .333 (.57,.55) .93 |
| | | .7 | .291 (.41,.37) .90 | .291 (.41,.37) .90 | .290 (.41,.36) .91 | .308 (.44,.40) .92 | .308 (.44,.40) .92 |
| | | .5 | .292 (.36,.34) .93 | .292 (.36,.34) .93 | .291 (.36,.35) .94 | .304 (.38,.37) .94 | .304 (.38,.37) .94 |
| | | .2 | .291 (.33,.32) .93 | .291 (.33,.32) .93 | .290 (.33,.34) .95 | .296 (.34,.35) .95 | .296 (.34,.34) .95 |
| | 4 | 1 | .290 (.59,.55) .93 | .289 (.59,.55) .93 | .290 (.59,.53) .91 | .345 (.71,.68) .94 | .345 (.71,.67) .94 |
| | | .7 | .314 (.51,.50) .93 | .314 (.51,.50) .93 | .316 (.52,.51) .95 | .348 (.57,.58) .96 | .348 (.57,.57) .96 |
| | | .5 | .320 (.50,.48) .93 | .320 (.50,.48) .93 | .320 (.51,.50) .94 | .346 (.54,.54) .95 | .346 (.54,.54) .95 |
| | | .2 | .319 (.47,.45) .93 | .319 (.47,.45) .93 | .320 (.47,.48) .96 | .334 (.49,.51) .96 | .334 (49,.51) .96 |

Table A.4: Mean estimate of between-subjects (**BS**) treatment effect (true value = 0.35, marginal true value = 0.302), followed in parenthesis by standard deviation among simulations, mean standard error and confidence interval coverage, based on analyses of 1000 simulated random effects (**SS**) datasets per setting ($n$ = number of subjects, $t$ = number of time points, $\rho$ = autocorrelation). See Table A.1 for coding of statistical methods.

|  |  |  | Statistical Methods |  |  |  |  |
|---|---|---|---|---|---|---|---|
| $n$ | $t$ | $\rho$ | $\hat{\beta}^{PA}_{\text{IND}}$ | $\hat{\beta}^{PA}_{\text{EXCH}}$ | $\hat{\beta}^{PA}_{\text{REPL}}$ | $\hat{\beta}^{SS}_{\text{MLa}}$ | $\hat{\beta}^{SS}_{\text{MLb}}$ |
| 100 | 16 | 1 | .294 (.10,.10) .94 | .294 (.10,.10) .94 | .290 (.10,.12) .99 | .351 (.12,.12) .95 | .351 (.12,.12) .95 |
|  |  | .7 | .293 (.12,.12) .94 | .293 (.12,.12) .94 | .292 (.12,.11) .93 | .305 (.13,.11) .88 | .305 (.13,.11) .88 |
|  |  | .5 | .295 (.11,.11) .96 | .295 (.11,.11) .96 | .295 (.11,.11) .95 | .301 (.12,.11) .90 | .301 (.12,.12) .90 |
|  |  | .2 | .294 (.11,.11) .95 | .294 (.11,.11) .95 | .294 (.11,.11) .95 | .296 (.11,.11) .91 | .296 (.11,.11) .91 |
|  | 8 | 1 | .285 (.14,.13) .92 | .285 (.14,.13) .92 | .288 (.14,.16) .97 | .344 (.16,.16) .93 | .344 (.16,.16) .93 |
|  |  | .7 | .284 (.15,.15) .95 | .284 (.15,.15) .95 | .293 (.15,.15) .97 | .305 (.16,.15) .91 | .305 (.16,.15) .91 |
|  |  | .5 | .284 (.15,.15) .94 | .284 (.15,.15) .94 | .294 (.15,.15) .95 | .294 (.15,.16) .91 | .294 (.15,.15) .91 |
|  |  | .2 | .290 (.15,.15) .95 | .290 (.15,.15) .95 | .296 (.15,.15) .95 | .293 (.15,.19) .93 | .293 (.15,.14) .93 |
|  | 4 | 1 | .291 (.18,.18) .95 | .291 (.18,.18) .95 | .291 (.18,.21) .98 | .354 (.22,.22) .96 | .354 (.22,.22) .96 |
|  |  | .7 | .287 (.19,.20) .95 | .287 (.19,.20) .95 | .287 (.19,.21) .96 | .319 (.22,.21) .94 | .319 (.22,.21) .94 |
|  |  | .5 | .280 (.20,.20) .94 | .280 (.20,.20) .95 | .280 (.20,.21) .95 | .300 (.21,.21) .94 | .300 (.21,.21) .94 |
|  |  | .2 | .284 (.20,.20) .94 | .284 (.20,.20) .94 | .283 (.20,.20) .95 | .292 (.21,.21) .93 | .292 (.21,.21) .93 |
| 20 | 16 | 1 | .282 (.23,.21) .91 | .282 (.23,.21) .91 | .283 (.23,.27) .94 | .336 (.27,.26) .94 | .336 (.27,.26) .94 |
|  |  | .7 | .301 (.27,.26) .93 | .301 (.27,.26) .93 | .301 (.27,.26) .94 | .312 (.28,.24) .91 | .312 (.28,.24) .91 |
|  |  | .5 | .297 (.26,.25) .92 | .297 (.26,.25) .92 | .297 (.26,.25) .94 | .303 (.27,.24) .91 | .303 (.27,.24) .91 |
|  |  | .2 | .295 (.25,.23) .92 | .295 (.25,.23) .92 | .294 (.25,.24) .95 | .298 (.26,.24) .93 | .298 (.26,.24) .93 |
|  | 8 | 1 | .288 (.31,.28) .92 | .288 (.31,.28) .93 | .291 (.32,.36) .97 | .344 (.37,.36) .94 | .344 (.37,.36) .94 |
|  |  | .7 | .289 (.35,.32) .92 | .289 (.35,.32) .92 | .288 (.35,.35) .95 | .310 (.37,.34) .92 | .310 (.37,.34) .92 |
|  |  | .5 | .304 (.34,.32) .93 | .304 (.34,.32) .93 | .303 (.34,.34) .95 | .317 (.36,.33) .94 | .317 (.36,.33) .94 |
|  |  | .2 | .308 (.33,.32) .93 | .308 (.33,.32) .93 | .307 (.33,.33) .95 | .314 (.34,.33) .95 | .314 (.34,.33) .95 |
|  | 4 | 1 | .318 (.42,.40) .93 | .318 (.42,.40) .93 | .318 (.42,.47) .97 | .390 (.53,.52) .96 | .390 (.53,.52) .96 |
|  |  | .7 | .329 (.47,.44) .94 | .329 (.47,.44) .94 | .326 (.47,.47) .96 | .372 (.54,.49) .95 | .371 (.54,.49) .95 |
|  |  | .5 | .319 (.49,.45) .92 | .319 (.49,.45) .92 | .319 (.49,.47) .94 | .347 (.53,.48) .94 | .347 (.53,.48) .94 |
|  |  | .2 | .319 (.48,.45) .93 | .319 (.48,.45) .93 | .318 (.48,.47) .95 | .337 (.50,.48) .95 | .337 (.50,.48) .93 |

Table A.5: Mean estimate of interaction effect in **(interaction model)** (true value = - 0.15), followed in parenthesis by standard deviation among simulations, mean standard error and confidence interval coverage, based on analyses of 1000 simulated marginal (**PA**) datasets per setting ($n$ = number of subjects, $t$ = number of time points, $\rho$ = autocorrelation). Analysis by procedure B of type $A$ is designated by $\hat{\beta}^{A}_{\mathrm{B}}$, where $A$ = $PA$ (population-averaged) or $SS$ (subject-specific), and B = AR (GEE with autoregressive correlation), ALR (alternating logistic regression), ML (maximum likelihood), MCMC (Bayesian Markov chain Monte Carlo).

| | | | Statistical Methods[T] | | | |
|---|---|---|---|---|---|---|
| $n$ | $t$ | $\rho$ | $\hat{\beta}^{PA}_{\mathrm{AR}}$ | $\hat{\beta}^{PA}_{\mathrm{ALR}}$ | $\hat{\beta}^{SS}_{\mathrm{ML}}$ | $\hat{\beta}^{SS}_{\mathrm{MCMC}}$ |
| 100 | 16 | .7 | −.152 (.04,.04) .95 | −.151 (.04,.04) .96 | −.156 (.04,.02) .67 | −.156 (.04,.02) .65 |
| | | .5 | −.151 (.03,.03) .94 | −.151 (.04,.04) .94 | −.155 (.04,.02) .77 | −.155 (.04,.02) .75 |
| | | .2 | −.151 (.03,.03) .95 | −.151 (.03,.03) .95 | −.152 (.03,.02) .90 | −.153 (.03,.02) .87 |
| | 8 | .7 | −.154 (.08,.08) .94 | −.154 (.09,.08) .94 | −.160 (.09,.05) .73 | −.159 (.09,.05) .70 |
| | | .5 | −.154 (.08,.08) .94 | −.153 (.09,.08) .94 | −.160 (.09,.06) .80 | −.161 (.09,.06) .78 |
| | | .2 | −.152 (.07,.07) .94 | −.152 (.07,.07) .95 | −.154 (.08,.06) .90 | −.162 (.08,.06) .86 |
| | 4 | .7 | −.162 (.16,.16) .95 | −.160 (.16,.16) .94 | −.168 (.17,.13) .86 | −.176 (.18,.13) .85 |
| | | .5 | −.159 (.19,.18) .94 | −.157 (.19,.18) .94 | −.166 (.20,.16) .88 | −.166 (.20,.15) .85 |
| | | .2 | −.158 (.20,.19) .93 | −.157 (.20,.19) .93 | −.161 (.20,.18) .91 | −.179 (.21,.17) .87 |
| 20 | 16 | .7 | −.161 (.09,.09) .92 | −.160 (.10,.10) .92 | −.163 (.10,.05) .69 | −.163 (.10,.05) .65 |
| | | .5 | −.156 (.08,.07) .93 | −.155 (.08,.08) .93 | −.158 (.08,.05) .79 | −.159 (.08,.05) .76 |
| | | .2 | −.152 (.06,.06) .93 | −.152 (.06,.06) .93 | −.153 (.06,.05) .91 | −.156 (.06,.05) .87 |
| | 8 | .7 | −.175 (.18,.18) .95 | −.174 (.19,.19) .93 | −.176 (.20,.12) .77 | −.180 (.20,.12) .72 |
| | | .5 | −.169 (.18,.18) .95 | −.167 (.19,.18) .94 | −.173 (.19,.14) .85 | −.178 (.20,.13) .76 |
| | | .2 | −.166 (.16,.16) .94 | −.165 (.16,.16) .94 | −.167 (.16,.14) .92 | −.174 (.17,.14) .89 |

[T] Note that SS estimates were converted to PA value (see text).

Table A.6: Mean estimate of treatment main effect in **(interaction model)** (true value = 0.35), followed in parenthesis by standard deviation among simulations, mean standard error and confidence interval coverage, based on analyses of 1000 simulated marginal (**PA**) datasets per setting ($n$ = number of subjects, $t$ = number of time points, $\rho$ = autocorrelation). See Table A.5 for coding of statistical methods.

| | | | Statistical Methods[T] | | | |
|---|---|---|---|---|---|---|
| $n$ | $t$ | $\rho$ | $\hat{\beta}^{PA}_{\text{AR}}$ | $\hat{\beta}^{PA}_{\text{ALR}}$ | $\hat{\beta}^{SS}_{\text{ML}}$ | $\hat{\beta}^{SS}_{\text{MCMC}}$ |
| 100 | 16 | .7 | .358 ( .37, .38) .96 | .356 ( .40, .40) .95 | .338 ( .42,.30) .83 | .342 ( .42,.29) .79 |
| | | .5 | .356 ( .32, .32) .95 | .356 ( .33, .33) .95 | .360 ( .34,.26) .86 | .358 ( .40,.25) .82 |
| | | .2 | .349 ( .25, .25) .95 | .350 ( .25, .25) .96 | .352 ( .26,.23) .92 | .357 ( .26,.22) .89 |
| | 8 | .7 | .352 ( .45, .45) .96 | .353 ( .47, .47) .96 | .349 ( .49,.38) .86 | .352 ( .50,.37) .84 |
| | | .5 | .353 ( .42, .42) .95 | .353 ( .43, .43) .95 | .366 ( .45,.36) .88 | .372 ( .46,.35) .85 |
| | | .2 | .348 ( .35, .36) .95 | .348 ( .36, .36) .96 | .353 ( .36,.33) .93 | .391 ( .37,.31) .90 |
| | 4 | .7 | .364 ( .52, .51) .94 | .360 ( .53, .52) .95 | .374 ( .57,.47) .89 | .399 ( .59,.47) .87 |
| | | .5 | .357 ( .55, .53) .94 | .354 ( .56, .54) .94 | .375 ( .59,.50) .90 | .378 ( .58,.48) .87 |
| | | .2 | .355 ( .54, .53) .94 | .354 ( .54, .53) .94 | .364 ( .55,.50) .93 | .412 ( .57,.49) .90 |
| 20 | 16 | .7 | .354 ( .91, .84) .94 | .358 ( .96, .89) .93 | .334 (1.00,.63) .78 | .332 (1.00,.67) .80 |
| | | .5 | .338 ( .74, .70) .93 | .339 ( .76, .73) .94 | .342 ( .78,.57) .85 | .337 ( .80,.56) .84 |
| | | .2 | .338 ( .57, .55) .93 | .337 ( .57, .55) .93 | .338 ( .58,.51) .92 | .348 ( .60,.51) .89 |
| | 8 | .7 | .433 (1.08,1.03) .95 | .437 (1.14,1.07) .90 | .440 (1.15,.80) .84 | .434 (1.21,.87) .81 |
| | | .5 | .413 ( .99, .94) .95 | .414 (1.02, .97) .94 | .424 (1.06,.81) .87 | .443 (1.09,.81) .85 |
| | | .2 | .409 ( .82, .79) .94 | .409 ( .82, .79) .94 | .413 ( .83,.75) .93 | .434 ( .88,.76) .90 |

[T] Note that SS estimates were converted to PA value (see text).

Table A.7: Mean estimate of interaction effect in **(interaction model)** (true value = − 0.15, true marginal value = − 0.129), followed in parenthesis by standard deviation among simulations, mean standard error and confidence interval coverage, based on analyses of 1000 simulated random effects (**SS**) datasets per setting ($n$ = number of subjects, $t$ = number of time points, $\rho$ = autocorrelation). See Table A.5 for coding of statistical methods.

| | | | Statistical Methods | | | |
|---|---|---|---|---|---|---|
| $n$ | $t$ | $\rho$ | $\hat{\beta}^{PA}_{\text{AR}}$ | $\hat{\beta}^{PA}_{\text{ALR}}$ | $\hat{\beta}^{SS}_{\text{ML}}$ | $\hat{\beta}^{SS}_{\text{MCMC}}$ |
| 100 | 16 | 1 | −.126 (.02,.02) .96 | −.126 (.02,.02) .96 | −.151 (.02,.03) .96 | −.152 (.02,.03) .95 |
| | | .7 | −.126 (.03,.03) .95 | −.125 (.03,.03) .96 | −.131 (.03,.02) .80 | −.131 (.03,.02) .78 |
| | | .5 | −.125 (.02,.03) .94 | −.125 (.03,.03) .94 | −.127 (.03,.02) .80 | −.130 (.03,.02) .79 |
| | | .2 | −.125 (.02,.02) .94 | −.125 (.02,.02) .94 | −.125 (.02,.02) .79 | −.131 (.02,.02) .80 |
| | 8 | 1 | −.126 (.06,.06) .93 | −.126 (.06,.06) .94 | −.152 (.07,.07) .94 | −.153 (.07,.07) .92 |
| | | .7 | −.128 (.07,.07) .94 | −.128 (.07,.07) .94 | −.137 (.07,.07) .92 | −.143 (.07,.06) .89 |
| | | .5 | −.128 (.07,.07) .94 | −.128 (.07,.07) .95 | −.133 (.07,.06) .92 | −.144 (.07,.06) .90 |
| | | .2 | −.127 (.07,.06) .94 | −.127 (.07,.06) .94 | −.128 (.07,.06) .93 | −.144 (.07,.06) .91 |
| | 4 | 1 | −.125 (.17,.17) .95 | −.125 (.17,.16) .95 | −.152 (.20,.20) .96 | −.160 (.20,.19) .92 |
| | | .7 | −.126 (.18,.18) .95 | −.126 (.18,.18) .95 | −.140 (.20,.19) .94 | −.159 (.20,.18) .91 |
| | | .5 | −.130 (.18,.18) .95 | −.131 (.18,.18) .95 | −.139 (.19,.19) .94 | −.163 (.20,.18) .91 |
| | | .2 | −.128 (.18,.18) .95 | −.129 (.18,.18) .95 | −.132 (.19,.18) .94 | −.162 (.20,.18) .92 |
| 20 | 16 | 1 | −.129 (.05,.05) .92 | −.129 (.05,.05) .92 | −.152 (.06,.06) .95 | −.154 (.06,.06) .93 |
| | | .7 | −.126 (.06,.06) .92 | −.126 (.06,.06) .92 | −.130 (.06,.05) .88 | −.127 (.06,.06) .92 |
| | | .5 | −.127 (.06,.05) .92 | −.126 (.06,.05) .92 | −.128 (.06,.05) .90 | −.126 (.06,.06) .94 |
| | | .2 | −.126 (.05,.05) .91 | −.126 (.05,.05) .91 | −.127 (.06,.05) .90 | −.124 (.06,.06) .96 |
| | 8 | 1 | −.131 (.14,.13) .93 | −.131 (.14,.13) .93 | −.154 (.16,.16) .95 | −.154 (.17,.16) .91 |
| | | .7 | −.127 (.15,.15) .94 | −.126 (.15,.15) .94 | −.134 (.16,.15) .93 | −.123 (.17,.15) .91 |
| | | .5 | −.130 (.15,.15) .94 | −.131 (.15,.15) .93 | −.135 (.16,.15) .95 | −.123 (.16,.15) .92 |
| | | .2 | −.131 (.15,.14) .92 | −.131 (.15,.14) .92 | −.133 (.15,.14) .94 | −.120 (.16,.15) .92 |

Table A.8: Mean estimate of treatment main effect in **(interaction model)** (true value = – 0.15, true marginal value = – 0.129), followed in parenthesis by standard deviation among simulations, mean standard error and confidence interval coverage, based on analyses of 1000 simulated random effects (**SS**) datasets per setting ($n$ = number of subjects, $t$ = number of time points, $\rho$ = autocorrelation). See Table A.5 for coding of statistical methods.

| | | | Statistical Methods | | | |
|---|---|---|---|---|---|---|
| $n$ | $t$ | $\rho$ | $\hat{\beta}^{PA}_{\text{AR}}$ | $\hat{\beta}^{PA}_{\text{ALR}}$ | $\hat{\beta}^{SS}_{\text{ML}}$ | $\hat{\beta}^{SS}_{\text{MCMC}}$ |
| 100 | 16 | 1 | .289 (.25,.26) .96 | .289 (.25,.26) .96 | .345 (.30,.31) .96 | .349 ( .31,.30) .93 |
| | | .7 | .293 (.26,.25) .94 | .292 (.26,.25) .94 | .304 (.27,.23) .92 | .311 ( .27,.23) .89 |
| | | .5 | .292 (.24,.23) .95 | .292 (.24,.23) .95 | .297 (.24,.22) .93 | .322 ( .25,.21) .90 |
| | | .2 | .290 (.22,.22) .95 | .290 (.22,.22) .95 | .291 (.22,.22) .95 | .340 ( .23,.20) .91 |
| | 8 | 1 | .300 (.34,.33) .94 | .301 (.34,.33) .94 | .362 (.41,.40) .95 | .373 ( .42,.39) .92 |
| | | .7 | .305 (.36,.34) .93 | .305 (.36,.34) .93 | .327 (.38,.35) .92 | .360 ( .39,.33) .89 |
| | | .5 | .308 (.34,.33) .94 | .309 (.34,.34) .94 | .319 (.35,.33) .94 | .375 ( .36,.31) .90 |
| | | .2 | .302 (.33,.32) .94 | .302 (.33,.32) .94 | .305 (.33,.32) .94 | .383 ( .34,.30) .91 |
| | 4 | 1 | .307 (.50,.48) .94 | .307 (.50,.48) .95 | .372 (.60,.59) .95 | .398 ( .60,.56) .92 |
| | | .7 | .303 (.51,.50) .95 | .303 (.51,.50) .95 | .336 (.57,.54) .95 | .393 ( .57,.51) .91 |
| | | .5 | .309 (.51,.50) .94 | .310 (.51,.50) .95 | .331 (.54,.52) .94 | .398 ( .57,.50) .90 |
| | | .2 | .309 (.51,.50) .95 | .309 (.51,.50) .96 | .318 (.52,.51) .96 | .402 ( .56,.49) .91 |
| 20 | 16 | 1 | .317 (.60,.56) .92 | .318 (.60,.56) .93 | .373 (.71,.68) .94 | .380 ( .74,.70) .92 |
| | | .7 | .289 (.59,.54) .91 | .288 (.59,.54) .91 | .297 (.61,.52) .91 | .265 ( .62,.56) .94 |
| | | .5 | .296 (.55,.50) .92 | .295 (.54,.50) .92 | .300 (.55,.50) .93 | .267 ( .56,.55) .95 |
| | | .2 | .295 (.51,.46) .91 | .295 (.51,.46) .92 | .297 (.52,.49) .94 | .264 ( .52,.54) .96 |
| | 8 | 1 | .300 (.81,.73) .93 | .297 (.80,.73) .93 | .350 (.96,.90) .94 | .337 (1.02,.93) .92 |
| | | .7 | .301 (.80,.76) .94 | .300 (.79,.76) .94 | .317 (.85,.78) .94 | .262 ( .87,.80).94 |
| | | .5 | .315 (.78,.74) .94 | .317 (.78,.74) .94 | .329 (.81,.75) .94 | .272 ( .83,.78) .94 |
| | | .2 | .317 (.76,.71) .92 | .317 (.76,.71) .92 | .323 (.77,.73) .94 | .260 ( .79,.77) .94 |

# Appendix B

# Additional tables of Chapter 4

Table B.1: Relative bias of estimates and standard errors to the true values (RBT) with a significance indication, based on analyses of 1000 simulated datasets generated by random intercept model ($\rho$ = **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: $\beta_0$ (intercept), $\beta_1$ (time coefficient), $\beta_2$ (subject level factor), $\beta_3$ (cluster level factor), $\sigma_2^2$ (variance at subject level), $\sigma_3^2$ (variance at cluster level), $\phi$ (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).

| | | Statistical Methods | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Scen- | parm- | PQL | | PQLx | | ML | | MCMC | |
| ario | eter. | Est. | SE | Est. | SE | Est. | SE | Est. | SE |
| scc40 | $\beta_0$ | $-1.3$ | $-4.9^{*}$ | $2.7^{\ddagger}$ | $-5.9^{*}$ | $0.1$ | $-4.3^{*}$ | $0.0$ | $-3.1$ |
| | $\beta_1$ | $-0.7^{\ddagger}$ | $-8.7^{*}$ | $4.0^{\ddagger}$ | $-16.7^{*}$ | $0.0$ | $-4.3$ | $0.0$ | $-4.3$ |
| | $\beta_2$ | $-4.2^{\dagger}$ | $-0.6$ | $-4.2$ | $-0.6$ | $-0.5$ | $0.0$ | $-0.1$ | $0.0$ |
| | $\beta_3$ | $-2.0$ | $-4.3^{*}$ | $1.5$ | $-4.1$ | $1.5$ | $-4.1$ | $1.7$ | $-1.3$ |
| | $\sigma_2^2$ | $-10.8^{\ddagger}$ | $-15.1^{*}$ | $7.5^{\ddagger}$ | $-23.2^{*}$ | $0.8^{\dagger}$ | $-2.8$ | $2.5^{\ddagger}$ | $-2.0$ |
| | $\sigma_3^2$ | $-15.4^{\ddagger}$ | $-5.1^{*}$ | $-9.2^{\ddagger}$ | $-5.2^{*}$ | $-9.2^{\ddagger}$ | $-3.1$ | $0.1$ | $8.8^{*}$ |
| | $\phi$ | | | $-17.8^{\ddagger}$ | $37.5^{*}$ | | | | |
| MARL | $\beta_0$ | $-1.0$ | $-2.4$ | $4.6^{\ddagger}$ | $-3.4$ | $1.1$ | $-1.1$ | $3.2^{\dagger}$ | $-5.2^{*}$ |
| | $\beta_1$ | $-2.1^{\ddagger}$ | $-5.0^{*}$ | $13.8^{\ddagger}$ | $-17.0^{*}$ | $0.8$ | $0.1$ | $2.1^{\ddagger}$ | $-1.7$ |
| | $\beta_2$ | $-4.5^{\ddagger}$ | $-2.0$ | $0.0$ | $-1.2$ | $-1.0$ | $0.1$ | $-0.1$ | $-1.0$ |
| | $\beta_3$ | $-3.2^{\ddagger}$ | $-5.2^{*}$ | $1.4$ | $-5.3^{*}$ | $0.2$ | $-4.4^{*}$ | $2.2$ | $-3.4$ |
| | $\sigma_2^2$ | $-12.1^{\ddagger}$ | $-12.4^{*}$ | $10.7^{\ddagger}$ | $-23.2^{*}$ | $0.7$ | $2.4$ | $3.5^{\ddagger}$ | $-0.7$ |
| | $\sigma_3^2$ | $-14.0^{\ddagger}$ | $-6.4^{*}$ | $-5.5^{\ddagger}$ | $-6.2^{*}$ | $-7.6^{\ddagger}$ | $-3.8^{*}$ | $2.0$ | $11.0^{*}$ |
| | $\phi$ | | | $-19.0^{\ddagger}$ | $25.6^{*}$ | | | | |
| MARH | $\beta_0$ | $-3.2^{\ddagger}$ | $-5.3^{*}$ | $13.6^{\ddagger}$ | $-6.3^{*}$ | $1.3$ | $-2.1$ | $1.6$ | $-0.4$ |
| | $\beta_1$ | $-21.7^{\ddagger}$ | $-15.7^{*}$ | $92.9^{\ddagger}$ | $-45.6^{*}$ | $1.2$ | $2.0$ | $4.4^{\ddagger}$ | $-0.3$ |
| | $\beta_2$ | $-7.1^{\ddagger}$ | $-7.7^{*}$ | $11.0^{\ddagger}$ | $-3.1$ | $-1.0$ | $-0.6$ | $-0.2$ | $-0.7$ |
| | $\beta_3$ | $-5.5^{\ddagger}$ | $-8.0^{*}$ | $12.9^{\ddagger}$ | $-7.3^{*}$ | $0.5$ | $-5.4^{*}$ | $1.3$ | $-3.5$ |
| | $\sigma_2^2$ | $-26.0^{\ddagger}$ | $-26.7^{*}$ | $66.7^{\ddagger}$ | $-49.3^{*}$ | $1.4^{\dagger}$ | $3.6$ | $5.0^{\ddagger}$ | $1.3$ |
| | $\sigma_3^2$ | $-19.0^{\ddagger}$ | $-9.5^{*}$ | $15.8^{\ddagger}$ | $-6.0^{*}$ | $-8.1^{\ddagger}$ | $-3.4$ | $1.7$ | $8.4^{*}$ |
| | $\phi$ | | | $-27.6^{\ddagger}$ | $-26.7^{*}$ | | | | |
| NMARL | $\beta_0$ | $-2.6^{\ddagger}$ | $-1.6$ | $2.7^{\ddagger}$ | $-2.5$ | $0.2$ | $-0.7$ | $0.6$ | $-1.8$ |
| | $\beta_1$ | $-78.8^{\ddagger}$ | $-3.8$ | $-71.3^{\ddagger}$ | $-14.7^{*}$ | $-77.7^{\ddagger}$ | $-1.8$ | $-77.4^{\ddagger}$ | $-2.4$ |
| | $\beta_2$ | $-4.8^{\ddagger}$ | $-0.4$ | $-0.6$ | $0.4$ | $-1.5^{\ddagger}$ | $0.7$ | $-1.1^{\dagger}$ | $-0.4$ |
| | $\beta_3$ | $-3.3^{\ddagger}$ | $-4.6^{*}$ | $1.1$ | $-4.6^{*}$ | $-0.2$ | $-4.1$ | $0.4$ | $-3.2$ |
| | $\sigma_2^2$ | $-11.2^{\ddagger}$ | $-11.9^{*}$ | $11.2^{\ddagger}$ | $-22.0^{*}$ | $0.2$ | $1.3$ | $1.9^{\ddagger}$ | $1.1$ |
| | $\sigma_3^2$ | $-13.9^{\ddagger}$ | $-6.5^{*}$ | $-6.0^{\ddagger}$ | $-6.5^{*}$ | $-8.4^{\ddagger}$ | $-4.8^{*}$ | $0.8$ | $7.8^{*}$ |
| | $\phi$ | | | $-19.8^{\ddagger}$ | $16.4^{*}$ | | | | |
| NMARH | $\beta_0$ | $11.5^{\ddagger}$ | $-6.3^{*}$ | $26.2^{\ddagger}$ | $-5.7^{*}$ | $14.7^{\ddagger}$ | $-3.1$ | $15.1^{\ddagger}$ | $-1.0$ |
| | $\beta_1$ | $-317.4^{\ddagger}$ | $-5.7^{*}$ | $-296.5^{\ddagger}$ | $-22.8^{*}$ | $-318.0^{\ddagger}$ | $-10.3^{*}$ | $-318.1^{\ddagger}$ | $-10.9^{*}$ |
| | $\beta_2$ | $-8.7^{\ddagger}$ | $-4.0$ | $4.4^{\ddagger}$ | $1.4$ | $-6.2^{\ddagger}$ | $-2.2$ | $-5.5^{\ddagger}$ | $-2.2$ |
| | $\beta_3$ | $-7.3^{\ddagger}$ | $-7.7^{*}$ | $6.3^{\ddagger}$ | $-6.2^{*}$ | $-5.3^{\ddagger}$ | $-6.5^{*}$ | $-4.7^{\ddagger}$ | $-4.3^{*}$ |
| | $\sigma_2^2$ | $-23.0^{\ddagger}$ | $-24.8^{*}$ | $53.1^{\ddagger}$ | $-40.9^{*}$ | $-11.3^{\ddagger}$ | $0.8$ | $-8.7^{\ddagger}$ | $0.3$ |
| | $\sigma_3^2$ | $-19.8^{\ddagger}$ | $-7.9^{*}$ | $5.5^{\ddagger}$ | $-3.6$ | $-16.3^{\ddagger}$ | $-4.3^{*}$ | $-7.8^{\ddagger}$ | $7.0^{*}$ |
| | $\phi$ | | | $-28.3^{\ddagger}$ | $-47.3^{*}$ | | | | |

$^{\dagger}$ significant bias in estimate at $P < 0.05$; $^{\ddagger}$ significant bias in estimate at $P < 0.01$; $^{*}$ significant bias in standard error at $P < 0.05$

Table B.2: Relative bias of estimates and standard errors to the true values (RBT) 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: $\beta_0$ (intercept), $\beta_1$ (time coefficient), $\beta_2$ (subject level factor), $\beta_3$ (cluster level factor), $\sigma_2^2$ (variance at subject level), $\sigma_3^2$ (variance at cluster level), $\phi$ (extra-binomial dispersion). see Table B.1 for coding of estimation procedures.

|  |  | Statistical Methods |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
| Scen- | Parm- | PQL |  | PQLx |  | ML |  | MCMC |  |
| ario | eter | Est. | SE | Est. | SE | Est. | SE | Est. | SE |
| scc40 | $\beta_0$ | $-6.1^{\ddagger}$ | $-1.3$ | $-2.7^{\ddagger}$ | $-2.1$ | $-5.3^{\ddagger}$ | $-0.9$ | $-5.1^{\ddagger}$ | $-1.3$ |
|  | $\beta_1$ | $-6.5^{\ddagger}$ | $-9.1^{*}$ | $-1.3^{\ddagger}$ | $-13.6^{*}$ | $-4.7^{\ddagger}$ | $-4.5^{*}$ | $-4.7^{\ddagger}$ | $-4.5^{*}$ |
|  | $\beta_2$ | $-6.4^{\ddagger}$ | $-1.4$ | $-3.4^{\ddagger}$ | $-0.7$ | $-3.9^{\ddagger}$ | $0.7$ | $-3.7^{\ddagger}$ | $0.0$ |
|  | $\beta_3$ | $-6.5^{\ddagger}$ | $-3.8$ | $-3.5^{\dagger}$ | $-3.7$ | $-4.2^{\dagger}$ | $-3.4$ | $-3.9^{\ddagger}$ | $-2.0$ |
|  | $\sigma_2^2$ | $-25.2^{\ddagger}$ | $-16.6^{*}$ | $-10.3^{\ddagger}$ | $-24.7^{*}$ | $-16.8^{\ddagger}$ | $-2.9$ | $-15.6^{\ddagger}$ | $-2.8$ |
|  | $\sigma_3^2$ | $-20.7^{\ddagger}$ | $-12.5^{*}$ | $-15.4^{\ddagger}$ | $-12.2^{*}$ | $-16.6^{\ddagger}$ | $-9.7^{*}$ | $-8.3^{\ddagger}$ | $2.4$ |
|  | $\phi$ |  |  | $-16.0^{\ddagger}$ | $46.7^{*}$ |  |  |  |  |
| MARL | $\beta_0$ | $-6.5^{\ddagger}$ | $-5.7^{*}$ | $-1.7$ | $-6.5^{*}$ | $-5.2^{\ddagger}$ | $-4.1$ | $-5.2^{\ddagger}$ | $-3.7$ |
|  | $\beta_1$ | $-6.1^{\ddagger}$ | $-11.1^{*}$ | $7.7^{\ddagger}$ | $-21.6^{*}$ | $-3.8^{\ddagger}$ | $-5.9^{*}$ | $-3.2^{\ddagger}$ | $-6.2^{*}$ |
|  | $\beta_2$ | $-6.5^{\ddagger}$ | $-0.4$ | $-2.5^{\ddagger}$ | $-0.5$ | $-4.1^{\ddagger}$ | $2.3$ | $-3.8^{\ddagger}$ | $2.2$ |
|  | $\beta_3$ | $-7.4^{\ddagger}$ | $-5.4^{*}$ | $-3.4^{\ddagger}$ | $-5.4^{*}$ | $-5.1^{\ddagger}$ | $-4.1$ | $-5.0^{\ddagger}$ | $-2.6$ |
|  | $\sigma_2^2$ | $-26.7^{\ddagger}$ | $-20.2^{*}$ | $-8.0^{\ddagger}$ | $-29.8^{*}$ | $-17.2^{\ddagger}$ | $-4.7^{*}$ | $-15.9^{\ddagger}$ | $-4.6^{*}$ |
|  | $\sigma_3^2$ | $-20.0^{\ddagger}$ | $-5.7^{*}$ | $-12.8^{\ddagger}$ | $-6.0^{*}$ | $-15.9^{\ddagger}$ | $-2.7$ | $-7.4^{\ddagger}$ | $10.8^{*}$ |
|  | $\phi$ |  |  | $-17.1^{\ddagger}$ | $40.7^{*}$ |  |  |  |  |
| MARH | $\beta_0$ | $-8.4^{\ddagger}$ | $-8.4^{*}$ | $4.0^{\ddagger}$ | $-9.7^{*}$ | $-5.7^{\ddagger}$ | $-4.9^{*}$ | $-5.4^{\ddagger}$ | $-3.5$ |
|  | $\beta_1$ | $-15.0^{\ddagger}$ | $-25.8^{*}$ | $78.1^{\ddagger}$ | $-50.7^{*}$ | $4.9^{\ddagger}$ | $-9.0^{*}$ | $7.7^{\ddagger}$ | $-10.6^{*}$ |
|  | $\beta_2$ | $-10.4^{\ddagger}$ | $-4.1$ | $5.6^{\ddagger}$ | $-3.2$ | $-6.1^{\ddagger}$ | $2.8$ | $-5.5^{\ddagger}$ | $2.8$ |
|  | $\beta_3$ | $-11.2^{\ddagger}$ | $-6.0^{*}$ | $4.7^{\ddagger}$ | $-6.2^{*}$ | $-7.1^{\ddagger}$ | $-3.3$ | $-6.6^{\ddagger}$ | $-1.6$ |
|  | $\sigma_2^2$ | $-46.5^{\ddagger}$ | $-35.0^{*}$ | $17.8^{\ddagger}$ | $-57.7^{*}$ | $-27.9^{\ddagger}$ | $-8.7^{*}$ | $-25.6^{\ddagger}$ | $-10.0^{*}$ |
|  | $\sigma_3^2$ | $-27.6^{\ddagger}$ | $-9.3^{*}$ | $1.0$ | $-9.2^{*}$ | $-20.4^{\ddagger}$ | $-2.5$ | $-12.0^{\ddagger}$ | $10.1^{*}$ |
|  | $\phi$ |  |  | $-22.3^{\ddagger}$ | $-8.6^{*}$ |  |  |  |  |
| NMARL | $\beta_0$ | $-11.5^{\ddagger}$ | $-4.9^{*}$ | $-7.8^{\ddagger}$ | $-5.8^{*}$ | $-10.7^{\ddagger}$ | $-3.5$ | $-10.5^{\ddagger}$ | $-2.4$ |
|  | $\beta_1$ | $-83.8^{\ddagger}$ | $-12.2^{*}$ | $-79.1^{\ddagger}$ | $-20.3^{*}$ | $-83.2^{\ddagger}$ | $-9.8^{*}$ | $-82.9^{\ddagger}$ | $-10.3^{*}$ |
|  | $\beta_2$ | $-10.0^{\ddagger}$ | $0.0$ | $-6.7^{\ddagger}$ | $0.5$ | $-8.7^{\ddagger}$ | $3.0$ | $-8.4^{\ddagger}$ | $2.0$ |
|  | $\beta_3$ | $-10.8^{\ddagger}$ | $-3.5$ | $-7.5^{\ddagger}$ | $-3.6$ | $-9.7^{\ddagger}$ | $-2.9$ | $-9.5^{\ddagger}$ | $-1.0$ |
|  | $\sigma_2^2$ | $-40.9^{\ddagger}$ | $-20.4^{*}$ | $-26.3^{\ddagger}$ | $-29.3^{*}$ | $-35.0^{\ddagger}$ | $-5.0$ | $-34.1^{\ddagger}$ | $-5.0^{*}$ |
|  | $\sigma_3^2$ | $-25.7^{\ddagger}$ | $-6.6^{*}$ | $-20.0^{\ddagger}$ | $-7.0^{*}$ | $-23.8^{\ddagger}$ | $-3.3$ | $-15.9^{\ddagger}$ | $10.0^{*}$ |
|  | $\phi$ |  |  | $-15.6^{\ddagger}$ | $49.5^{*}$ |  |  |  |  |
| NMARH | $\beta_0$ | $5.5^{\ddagger}$ | $-6.8^{*}$ | $18.1^{\ddagger}$ | $-7.9^{*}$ | $6.8^{\ddagger}$ | $-3.5$ | $7.2^{\ddagger}$ | $-0.8$ |
|  | $\beta_1$ | $-309.6^{\ddagger}$ | $-6.9^{*}$ | $-297.9^{\ddagger}$ | $-20.9^{*}$ | $-309.8^{\ddagger}$ | $-10.6^{*}$ | $-309.9^{\ddagger}$ | $-10.9$ |
|  | $\beta_2$ | $-12.9^{\ddagger}$ | $-2.4$ | $-1.6^{\ddagger}$ | $-1.0$ | $-11.5^{\ddagger}$ | $0.4$ | $-11.0^{\ddagger}$ | $0.2$ |
|  | $\beta_3$ | $-12.9^{\ddagger}$ | $-3.5$ | $-1.4$ | $-3.1^{*}$ | $-12.0^{\ddagger}$ | $-2.3$ | $-11.6^{\ddagger}$ | $0.8$ |
|  | $\sigma_2^2$ | $-44.5^{\ddagger}$ | $-26.8^{*}$ | $10.0^{\ddagger}$ | $-48.0^{*}$ | $-36.5^{\ddagger}$ | $3.7$ | $-35.1^{\ddagger}$ | $-5.4^{*}$ |
|  | $\sigma_3^2$ | $-28.0^{\ddagger}$ | $-6.4^{*}$ | $-7.5^{\ddagger}$ | $-5.4^{*}$ | $-26.3^{\ddagger}$ | $-1.8$ | $-18.8^{\ddagger}$ | $10.4^{*}$ |
|  | $\phi$ |  |  | $-23.1^{\ddagger}$ | $-39.6^{*}$ |  |  |  |  |

$^{\dagger}$ significant bias in estimate at $P < 0.05$; $^{\ddagger}$ significant bias in estimate at $P < 0.01$; $^{*}$ significant bias in standard error at $P < 0.05$

Table B.3: Relative bias of estimates and standard errors to the true values (RBT) 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: $\beta_0$ (intercept), $\beta_1$ (time coefficient), $\beta_2$ (subject level factor), $\beta_3$ (cluster level factor), $\sigma_2^2$ (variance at subject level), $\sigma_3^2$ (variance at cluster level), $\phi$ (extra-binomial dispersion). see Table B.1 for coding of estimation procedures.

| | | Statistical Methods | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Scen- | Parm- | PQL | | PQLx | | ML | | MCMC | |
| ario | eter | Est. | SE | Est. | SE | Est. | SE | Est. | SE |
| scc40 | $\beta_0$ | $-16.7^{\ddagger}$ | $-2.5$ | $-14.7^{\ddagger}$ | $-3.5$ | $-16.8^{\ddagger}$ | $-1.0$ | $-17.1^{\ddagger}$ | 0.0 |
| | $\beta_1$ | $-16.7^{\ddagger}$ | 0.0 | $-14.7^{\ddagger}$ | $-5.3^{*}$ | $-16.7^{\ddagger}$ | 0.0 | $-16.7^{\ddagger}$ | 0.0 |
| | $\beta_2$ | $-16.7^{\ddagger}$ | $-2.7$ | $-14.9^{\ddagger}$ | $-3.5$ | $-16.5^{\ddagger}$ | 0.9 | $-16.3^{\ddagger}$ | 0.0 |
| | $\beta_3$ | $-16.8^{\ddagger}$ | $-2.4$ | $-15.0^{\ddagger}$ | $-2.4$ | $-16.7^{\ddagger}$ | $-1.6$ | $-16.7^{\ddagger}$ | 1.6 |
| | $\sigma_2^2$ | $-68.4^{\ddagger}$ | $-14.2^{*}$ | $-61.9^{\ddagger}$ | $-21.6^{*}$ | $-65.5^{\ddagger}$ | $-1.6$ | $-65.2^{\ddagger}$ | 0.0 |
| | $\sigma_3^2$ | $-36.4^{\ddagger}$ | $-10.2^{*}$ | $-33.5^{\ddagger}$ | $-11.1^{*}$ | $-36.2^{\ddagger}$ | $-7.0^{*}$ | $-29.6^{\ddagger}$ | $6.3^{*}$ |
| | $\phi$ | | | $-10.2^{\ddagger}$ | $140.0^{*}$ | | | | |
| MARL | $\beta_0$ | $-16.9^{\ddagger}$ | $-5.5^{*}$ | $-14.2^{\ddagger}$ | $-6.4^{*}$ | $-16.8^{\ddagger}$ | $-3.9$ | $-16.9^{\ddagger}$ | $-0.9$ |
| | $\beta_1$ | $-18.5^{\ddagger}$ | $-8.9^{*}$ | $-10.8^{\ddagger}$ | $-15.8^{*}$ | $-17.3^{\ddagger}$ | $-2.7$ | $-17.0^{\ddagger}$ | $-3.6$ |
| | $\beta_2$ | $-16.9^{\ddagger}$ | $-4.9^{*}$ | $-14.5^{\ddagger}$ | $-4.8^{*}$ | $-16.6^{\ddagger}$ | 0.9 | $-16.4^{\ddagger}$ | $-1.0$ |
| | $\beta_3$ | $-16.0^{\ddagger}$ | $-7.2^{*}$ | $-13.6^{\ddagger}$ | $-7.5^{*}$ | $-15.7^{\ddagger}$ | $-6.3^{*}$ | $-15.7^{\ddagger}$ | $-2.7$ |
| | $\sigma_2^2$ | $-68.7^{\ddagger}$ | $-16.1^{*}$ | $-60.8^{\ddagger}$ | $-26.0^{*}$ | $-65.3^{\ddagger}$ | 1.2 | $-65.0^{\ddagger}$ | $-0.2$ |
| | $\sigma_3^2$ | $-35.4^{\ddagger}$ | $-7.6^{*}$ | $-31.6^{\ddagger}$ | $-8.1^{*}$ | $-35.0^{\ddagger}$ | $-3.4$ | $-28.3^{\ddagger}$ | $10.4^{*}$ |
| | $\phi$ | | | $-10.9^{\ddagger}$ | $111.7^{*}$ | | | | |
| MARH | $\beta_0$ | $-18.4^{\ddagger}$ | $-5.6^{*}$ | $-12.6^{\ddagger}$ | $-8.2^{*}$ | $-17.5^{\ddagger}$ | $-3.4$ | $-17.7^{\ddagger}$ | $-0.1$ |
| | $\beta_1$ | $-18.5^{\ddagger}$ | $-24.5^{*}$ | $20.9^{\ddagger}$ | $-47.8^{*}$ | $-8.9^{\ddagger}$ | $-7.0^{*}$ | $-10.1^{\ddagger}$ | $-12.4^{*}$ |
| | $\beta_2$ | $-22.0^{\ddagger}$ | $-8.0^{*}$ | $-14.9^{\ddagger}$ | $-13.1^{*}$ | $-20.6^{\ddagger}$ | $-1.9$ | $-20.8^{\ddagger}$ | $-3.1$ |
| | $\beta_3$ | $-21.2^{\ddagger}$ | $-7.6^{*}$ | $-13.9^{\ddagger}$ | $-10.2^{*}$ | $-19.8^{\ddagger}$ | $-6.0^{*}$ | $-19.9^{\ddagger}$ | $-1.3$ |
| | $\sigma_2^2$ | $-85.5^{\ddagger}$ | $-28.1^{*}$ | $-69.1^{\ddagger}$ | $-57.3^{*}$ | $-80.6^{\ddagger}$ | $-7.0^{*}$ | $-81.3^{\ddagger}$ | $-19.8^{*}$ |
| | $\sigma_3^2$ | $-43.3^{\ddagger}$ | $-11.5^{*}$ | $-32.2^{\ddagger}$ | $-16.1^{*}$ | $-41.3^{\ddagger}$ | $-5.3^{*}$ | $-35.5^{\ddagger}$ | $8.0^{*}$ |
| | $\phi$ | | | $-11.2^{\ddagger}$ | $9.9^{*}$ | | | | |
| NMARL | $\beta_0$ | $-22.7^{\ddagger}$ | $-5.7^{*}$ | $-21.2^{\ddagger}$ | $-6.5^{*}$ | $-22.8^{\ddagger}$ | $-4.5^{*}$ | $-23.0^{\ddagger}$ | $-1.7$ |
| | $\beta_1$ | $-85.2^{\ddagger}$ | $-12.8^{*}$ | $-84.1^{\ddagger}$ | $-16.6^{*}$ | $-85.0^{\ddagger}$ | $-10.9^{*}$ | $-85.0^{\ddagger}$ | $-11.1^{*}$ |
| | $\beta_2$ | $-22.4^{\ddagger}$ | $-3.6$ | $-21.1^{\ddagger}$ | $-4.0$ | $-22.5^{\ddagger}$ | $-0.1$ | $-22.5^{\ddagger}$ | $-0.4$ |
| | $\beta_3$ | $-21.5^{\ddagger}$ | $-7.1^{*}$ | $-20.1^{\ddagger}$ | $-7.3^{*}$ | $-21.6^{\ddagger}$ | $-6.3^{*}$ | $-21.7^{\ddagger}$ | $-2.4$ |
| | $\sigma_2^2$ | $-86.1^{\ddagger}$ | $-13.3^{*}$ | $-82.0^{\ddagger}$ | $-24.6^{*}$ | $-84.5^{\ddagger}$ | $-1.2$ | $-84.7^{\ddagger}$ | $-6.3^{*}$ |
| | $\sigma_3^2$ | $-43.8^{\ddagger}$ | $-7.0^{*}$ | $-41.8^{\ddagger}$ | $-7.6^{*}$ | $-44.0^{\ddagger}$ | $-3.2$ | $-38.2^{\ddagger}$ | $10.8^{*}$ |
| | $\phi$ | | | $-7.3^{\ddagger}$ | $107.1^{*}$ | | | | |
| NMARH | $\beta_0$ | $-4.8^{\ddagger}$ | $-6.7^{*}$ | $1.4^{\ddagger}$ | $-9.9^{*}$ | $-4.9^{\ddagger}$ | $-4.3^{*}$ | $-5.1^{\ddagger}$ | $-1.4$ |
| | $\beta_1$ | $-274.6^{\ddagger}$ | $-12.5^{*}$ | $-275.7^{\ddagger}$ | $-18.2^{*}$ | $-274.3^{\ddagger}$ | $-13.4^{*}$ | $-274.4^{\ddagger}$ | $-13.6^{*}$ |
| | $\beta_2$ | $-23.5^{\ddagger}$ | 0.4 | $-18.8^{\ddagger}$ | $-2.2$ | $-23.2^{\ddagger}$ | 3.5 | $-23.3^{\ddagger}$ | 2.6 |
| | $\beta_3$ | $-22.5^{\ddagger}$ | $-7.2^{*}$ | $-17.5^{\ddagger}$ | $-8.2^{*}$ | $-22.4^{\ddagger}$ | $-6.1^{*}$ | $-22.4^{\ddagger}$ | $-2.4$ |
| | $\sigma_2^2$ | $-84.1^{\ddagger}$ | $-14.8^{*}$ | $-68.6^{\ddagger}$ | $-46.9^{*}$ | $-81.8^{\ddagger}$ | $-1.5$ | $-83.2^{\ddagger}$ | $-16.8^{*}$ |
| | $\sigma_3^2$ | $-44.6^{\ddagger}$ | $-10.0^{*}$ | $-37.2^{\ddagger}$ | $-12.4^{*}$ | $-44.3^{\ddagger}$ | $-5.1^{*}$ | $-38.7^{\ddagger}$ | $7.4^{*}$ |
| | $\phi$ | | | $-11.8^{\ddagger}$ | $-21.4^{*}$ | | | | |

$^{\dagger}$ significant bias in estimate at $P < 0.05$; $^{\ddagger}$ significant bias in estimate at $P < 0.01$; $^{*}$ significant bias in standard error at $P < 0.05$

Table B.4: Relative bias of estimates and standard errors to the marginal true values (RBT) with a significance indication, 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), MNARL, MNARH (low (31%) and high (52%) proportion of missing values not at random). Parameters: $\beta_0$ (intercept), $\beta_1$ (time coefficient), $\beta_2$ (subject level factor), $\beta_3$ (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 | $\beta_0$ | | $-5.3^{\ddagger}$ | $-50.8^{*}$ | $-4.5^{\ddagger}$ | $-4.7^{*}$ | $-6.4^{\ddagger}$ | $-48.2^{*}$ | $-5.8^{\ddagger}$ | $-1.2$ | $-5.0^{\ddagger}$ | $-47.0^{*}$ | $-4.5^{\ddagger}$ | $-3.7$ |
| | $\beta_1$ | | $-8.2^{\ddagger}$ | 0.0 | $-4.5^{\ddagger}$ | $12.5^{*}$ | $-8.2^{\ddagger}$ | 0.0 | $-5.4^{\ddagger}$ | 0.0 | $-6.4^{\ddagger}$ | $6.3^{*}$ | $-4.5^{\ddagger}$ | $-6.3$ |
| | $\beta_2$ | | $-5.3^{\ddagger}$ | $-36.3^{*}$ | $-5.3^{\ddagger}$ | $-1.9$ | $-4.5^{\ddagger}$ | $-34.5^{*}$ | $-4.6^{\ddagger}$ | $-2.9$ | $-4.4^{\ddagger}$ | $-22.6^{*}$ | $-4.4^{\ddagger}$ | $-2.2$ |
| | $\beta_3$ | | $-3.5^{\ddagger}$ | $-66.7^{*}$ | $-3.4^{\ddagger}$ | $-3.3$ | $-4.6^{\ddagger}$ | $-66.7^{*}$ | $-4.8^{\ddagger}$ | $-2.9$ | $-4.6^{\ddagger}$ | $-64.7^{*}$ | $-4.5^{\ddagger}$ | $-2.0$ |
| MARL | $\beta_0$ | | $-11.5^{\ddagger}$ | $-51.0^{*}$ | $-3.8^{\ddagger}$ | $-1.8$ | $-12.8^{\ddagger}$ | $-52.7^{*}$ | $-5.7^{\ddagger}$ | $-5.7^{*}$ | $-9.7^{\ddagger}$ | $-50.8^{*}$ | $-4.7^{\ddagger}$ | $-5.1^{*}$ |
| | $\beta_1$ | | $-77.0^{\ddagger}$ | $-6.0$ | $-3.7^{\ddagger}$ | $-3.3$ | $-70.6^{\ddagger}$ | $-9.7$ | $-3.8^{\ddagger}$ | $-8.2^{*}$ | $-48.1^{\ddagger}$ | $-6.9^{*}$ | $-4.4^{\ddagger}$ | $-5.7^{*}$ |
| | $\beta_2$ | | $-8.2^{\ddagger}$ | $-34.4^{*}$ | $-5.5^{\ddagger}$ | $-2.9$ | $-6.7^{\ddagger}$ | $-31.2^{*}$ | $-4.5^{\ddagger}$ | $-0.3$ | $-5.9^{\ddagger}$ | $-21.6^{*}$ | $-4.4^{\ddagger}$ | $-3.2$ |
| | $\beta_3$ | | $-7.2^{\ddagger}$ | $-65.3^{*}$ | $-4.6^{\ddagger}$ | $-3.7$ | $-7.8^{\ddagger}$ | $-65.3^{*}$ | $-5.6^{\ddagger}$ | $-4.2^{*}$ | $-5.1^{\ddagger}$ | $-65.1^{*}$ | $-3.5^{\ddagger}$ | $-6.0^{*}$ |
| MARH | $\beta_0$ | | $-5.0^{\ddagger}$ | $-42.4^{*}$ | $-8.8^{\ddagger}$ | $-1.1$ | $-4.6^{\ddagger}$ | $-44.2^{*}$ | $-7.5^{\ddagger}$ | $-5.8^{*}$ | $-2.2^{\ddagger}$ | $-42.4^{*}$ | $-1.5^{\ddagger}$ | $-3.6$ |
| | $\beta_1$ | | $-200.7^{\ddagger}$ | $-10.0^{*}$ | $28.4^{\ddagger}$ | $-4.9^{*}$ | $-162.7^{\ddagger}$ | $-13.1^{*}$ | $42.2^{\ddagger}$ | $-9.6^{*}$ | $-83.7^{\ddagger}$ | $-10.7^{*}$ | $52.6^{\ddagger}$ | $-16.8^{*}$ |
| | $\beta_2$ | | $-16.3^{\ddagger}$ | $-19.3^{*}$ | $-4.7^{\ddagger}$ | $-1.6$ | $-13.3^{\ddagger}$ | $11.8^{*}$ | $-2.8^{\ddagger}$ | 2.9 | $-9.4^{\ddagger}$ | $-6.8^{*}$ | $-2.1^{\ddagger}$ | $-1.7$ |
| | $\beta_3$ | | $-15.0^{\ddagger}$ | $-54.2^{*}$ | $-3.6^{\ddagger}$ | $-1.8$ | $-14.1^{\ddagger}$ | $-53.9^{*}$ | $-3.9^{\ddagger}$ | $-0.7$ | $-8.4^{\ddagger}$ | $-55.2^{*}$ | $-1.2^{\ddagger}$ | $-4.0$ |
| NMARL | $\beta_0$ | | $-11.2^{\ddagger}$ | $-50.8^{*}$ | $-4.1^{\ddagger}$ | $-1.9$ | $-11.2^{\ddagger}$ | $-51.7^{*}$ | $-6.4^{\ddagger}$ | $-4.8^{*}$ | $-6.5^{\ddagger}$ | $-49.6^{*}$ | $-4.7^{\ddagger}$ | $-4.7^{*}$ |
| | $\beta_1$ | | $-127.7^{\ddagger}$ | $-5.9^{*}$ | $-78.3^{\ddagger}$ | $-6.5^{*}$ | $-117.1^{\ddagger}$ | $-11.4^{*}$ | $-82.5^{\ddagger}$ | $-6.2^{*}$ | $-96.0^{\ddagger}$ | $-11.7^{*}$ | $-81.6^{\ddagger}$ | $-6.5^{*}$ |
| | $\beta_2$ | | $-6.7^{\ddagger}$ | $-34.2^{*}$ | $-5.2^{\ddagger}$ | $-4.0$ | $-5.1^{\ddagger}$ | $-27.6^{*}$ | $-4.0^{\ddagger}$ | $-1.5$ | $-4.5^{\ddagger}$ | $-12.3^{*}$ | $-4.1^{\ddagger}$ | $-3.7$ |
| | $\beta_3$ | | $-5.7^{\ddagger}$ | $-65.2^{*}$ | $-4.1^{\ddagger}$ | $-4.7^{*}$ | $-6.2^{\ddagger}$ | $-64.1^{*}$ | $-5.1^{\ddagger}$ | $-3.3$ | $-3.5^{\ddagger}$ | $-64.1^{*}$ | $-3.1^{\ddagger}$ | $-6.3^{*}$ |
| NMARH | $\beta_0$ | | $7.5^{\ddagger}$ | $-41.2^{*}$ | $-12.6^{\ddagger}$ | $-5.9^{*}$ | $8.8^{\ddagger}$ | $-41.7^{*}$ | $11.8^{\ddagger}$ | $-6.9^{*}$ | $-15.0^{\ddagger}$ | $-40.0^{*}$ | $16.1^{\ddagger}$ | $-5.8^{*}$ |
| | $\beta_1$ | | $-450.2^{\ddagger}$ | $-13.7^{*}$ | $-317.7^{\ddagger}$ | $-38.3^{*}$ | $-424.6^{\ddagger}$ | $-12.7^{*}$ | $-329.7^{\ddagger}$ | $-27.9^{*}$ | $-360.7^{\ddagger}$ | $-13.9^{*}$ | $-320.4^{\ddagger}$ | $-15.9^{*}$ |
| | $\beta_2$ | | $-9.1^{\ddagger}$ | $-20.7^{*}$ | $-3.9^{\ddagger}$ | $-8.7^{*}$ | $-7.8^{\ddagger}$ | $-14.1^{*}$ | $-3.5^{\ddagger}$ | $-5.8^{*}$ | $-6.2^{\ddagger}$ | $-3.4$ | $-4.2^{\ddagger}$ | $-1.7$ |
| | $\beta_3$ | | $-8.5^{\ddagger}$ | $-53.1^{*}$ | $-3.2^{\ddagger}$ | $-9.6^{*}$ | $-8.3^{\ddagger}$ | $-51.0^{*}$ | $-4.1^{\ddagger}$ | $-5.7^{*}$ | $-5.5^{\ddagger}$ | $51.9^{*}$ | $-3.1^{\ddagger}$ | $-8.0^{*}$ |

$^{\dagger}$ significant bias in estimate at $P < 0.05$; $^{\ddagger}$ significant bias in estimate at $P < 0.01$; $^{*}$ significant bias in standard error at $P < 0.05$

Table B.5: Relative bias of estimates and standard errors to the marginal true values (RBT) with a significance indication, 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), MNARL, MNARH (low (31%) and high (52%) proportion of missing values not at random). Parameters: $\beta_0$ (intercept), $\beta_1$ (time coefficient), $\beta_2$ (subject level factor), $\beta_3$ (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 | $\beta_0$ | | $-3.4^{\ddagger}$ | 1.9 | $3.2^{\ddagger}$ | 1.3 | $-5.7^{\ddagger}$ | −2.8 | 1.9 | −2.8 | $-4.5^{\ddagger}$ | $-4.2^{*}$ | $9.4^{\ddagger}$ | $-6.1^{*}$ |
| | $\beta_1$ | | $-1.5^{\dagger}$ | 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 |
| | $\beta_2$ | | $-4.9^{\ddagger}$ | 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 |
| | $\beta_3$ | | $-4.2^{\ddagger}$ | −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 | $\beta_0$ | | $10.0^{\ddagger}$ | $-32.9^{*}$ | $-10.8^{\ddagger}$ | $-38.2^{*}$ | $3.6^{\ddagger}$ | $-38.2^{*}$ | $-7.9^{\dagger}$ | $-39.2^{*}$ | $4.0^{\ddagger}$ | $-31.5^{*}$ | 3.5 | $-34.7^{*}$ |
| | $\beta_1$ | | $-37.6^{\ddagger}$ | $-39.1^{*}$ | $-32.4^{\ddagger}$ | $-34.6^{*}$ | $-36.0^{\ddagger}$ | $-36.6^{*}$ | $-29.4^{\ddagger}$ | $-33.4^{*}$ | $-22.9^{\ddagger}$ | $-32.3^{*}$ | $-14.7^{\ddagger}$ | $-29.3^{*}$ |
| | $\beta_2$ | | $-5.0^{\dagger}$ | $-39.3^{*}$ | $-8.7^{\ddagger}$ | $-26.0^{*}$ | $-2.8^{\ddagger}$ | $-40.4^{*}$ | $-7.6^{\ddagger}$ | $-25.7^{*}$ | $-2.1^{\ddagger}$ | $-35.6^{*}$ | $-3.7^{\ddagger}$ | $-27.2^{*}$ |
| | $\beta_3$ | | $-5.2^{\dagger}$ | $-35.5^{*}$ | $-8.4^{\dagger}$ | $-43.1^{*}$ | $-5.4^{\ddagger}$ | $-34.6^{*}$ | −5.6 | $-41.1^{*}$ | $-1.6^{\ddagger}$ | $-32.9^{*}$ | −1.8 | $-41.2^{*}$ |

$^{\dagger}$ significant bias in estimate at $P < 0.05$; $^{\ddagger}$ significant bias in estimate at $P < 0.01$; $^{*}$ significant bias in standard error at $P < 0.05$