Analysis of Repeated Measurements from Medical Research when Observations are Missing

Thesis submitted to the University of London for the degree of Doctor of Philosophy in the Faculty of Science

by

Kate Walker

Department of Statistical Science
University College London

June 2007

Acknowledgements

I am grateful to the Medical Research Council for funding this work.

My thanks go to my supervisors, Rumana Omar and Andrew Copas, for their invaluable support and insight. They were readily available to give advice and encouragement, and dedicated a great deal of time to the project.

It has been a pleasure to work in the Department of Statistical Science, thanks to the welcoming and friendly environment provided by all the staff and students. In particular, I thank Susan, Nadja and Christian for their continuous encouragement and friendship, and Gareth for many inpromptu discussions about my work.

Finally, a huge thank you to all my friends and family. My parents have provided much wisdom and support, particularly throughout my PhD. My housemates, Kate, Danie, Jordi, Tom and Alex, have been a second family to me. Richard, thanks for your unwavering confidence in me and for keeping an eye on my sanity. And Gabby, thank you for everything... tea, chats, enthusiasm, laughs, belief, and more...

Subject dropout is a common problem in repeated measurements health studies. Where dropout is related to the response, the results obtained can be substantially biased. The research in this thesis is motivated by a repeated measurements asthma clinical trial with substantial patient dropout.

In practice the extent to which missing observations affect parameter estimates and their efficiency is not clear. Through extensive simulation studies under various scenarios and missing data mechanisms, the effect on parameter estimates of missing observations is explored and compared. Bias in the model estimates is found to be sensitive to the missing data mechanism, the type of model used, the estimation method, and the type of response variable, amongst other factors.

Findings from the simulation study highlight the importance of considering the likely dropout mechanism in choosing a model for the analysis of incomplete repeated measurements. For example, generalised estimating equations (GEE) require a missing completely at random (MCAR) assumption in general, as does the summary statistics method. Several formal tests of MCAR have been published, and these tests are compared both quantitatively, and in terms of their various merits and limitations.

Other than the sensitivity analysis, there are no widely accepted methods for analysing data with missing observations missing not at random (MNAR), as strong assumptions are required about the missing data mechanism. A method for incorporating cause of dropout into the analysis is proposed for MNAR data. A Bayesian hierarchical model is developed with informative priors for the bias of dropouts compared to completers for each cause of dropout. The feasibility of the proposed prior elicitation is investigated by consultation with clinicians. And the model is assessed through simulation studies, in which the sensitivity of the approach to misspecification of the parameters of the dropout mechanism is examined.

Contents

List of Tables

Chapter 1

Introduction

Medical studies frequently involve the collection of repeated measurements on subjects over the duration of the study. Such longitudinal studies allow the change in response to be measured over time, and provide more reliable information about chronic diseases, which are characterised by fluctuating symptoms. Collecting several measurements on each subject also enables within and between subject effects to be estimated.

Repeated measurements data are a special case of clustered data, in which observations within subject are correlated. Standard generalised linear models do not apply because of the lack of independence between measurements, and models have been developed to handle this dependence. The main models in use are the random effects model and generalised estimating equations (GEE) model. The summary statistics method, in which the observations for a subject are reduced to a single summary measure, is commonly used in medical research, and is a simple and generally valid approach. These methods for the analysis of clustered data are described in chapter 2.

Patient dropout is invariably a problem in repeated measurements studies. In a review of published randomised controlled trials, Wood et. al. [1] found almost 90% of them to contain missing data, and in many of these studies the missing data were handled inadequately. As well as reducing efficiency, missing data can cause substantial bias if, as is likely, the subjects that drop out are not representative of the whole sample. For example, results may be biased if sicker patients drop out of the study.

Missing data are commonly categorised into three levels, according to the severity of the bias they are likely to cause. Missing completely at random (MCAR) implies that the missing observations are are a random sample of all observations. A less strong assumption about the missing data is that they are missing at random (MAR), which means that although the probability of being missing is independent of the values of the missing observations, missingness may depend on the values of other observations in the data. The most difficult category of missing data to deal with is missing not at random (MNAR), where the probability of the observation being missing depends on its value.

If standards are to improve in the analysis of incomplete data, a better understanding of the extent to which missing data affects the parameter estimates is vital. The consensus of opinion is that GEE models require an assumption of missing completely at random (MCAR) data, while random effects models are robust to the more relaxed missing at random (MAR) assumption, except in the case of gaussian response data when both the random effects model and GEE, with correctly specified correlation, are robust to MAR data. The summary statistics method is only valid under the strict assumption of MCAR. In practice, the boundaries are not as clear as this. For example: Park [2] demonstrates that the GEE model is not always valid with MCAR data, even when the data are gaussian; it is unclear whether random effects models are robust to MAR data when covariates are cluster-varying [3]; and it has been suggested that GEE models may be robust to MAR observations for non-gaussian data if the correlation structure is correctly specified [4]. In chapter 3 of this thesis, a thorough investigation is carried out, using simulation studies together with evidence from the literature, in order to clarify these issues.

Armed with a clear framework on the robustness of methods to missing data, the researcher must be able to identify categories of missing data in order to assess the likely impact of missing observations on the parameter estimates. By definition, it is impossible to distinguish missing not at random (MNAR) and MAR mechanisms from the observed data. It is, however, possible to discriminate MCAR and MAR data from the information available. Several tests of MCAR have been published over the last few decades, each approaching the problem from a different angle, some more accessible to the applied statistician than others. Chapter 4 assesses these tests of MCAR, both qualitatively, in terms of their ease of implementation and flexibility to various scenarios, and quantitatively via simulation studies, and recommendations are made on the best approaches to employ.

A motivation for the work in this thesis was a repeated measurements asthma clinical trial, from which over 20% of subjects dropped out. In chapter 5 the data from the trial are analysed using the methods described in chapter 2. The missing data mechanism is explored using the approaches recommended in chapter 4, and the likely bias in the parameter estimates is discussed, based on the findings of chapter 3.

Missing not at random dropout in the asthma clinical trial cannot be ruled out. The issue of MNAR data poses a difficult challenge because of the necessity to make strong and untestable assumptions about the nature of the missing data mechanism. It is therefore a vastly overlooked issue, with the sensitivity analysis as the only standard method in use. In the sensitivity analysis, a model is constructed for the missing data mechanism, and a range of plausible parameter values for this mechanism results in a range of parameter estimates for the model of interest. This is somewhat unsatisfactory if a point estimate is required, and in fact the approach is often used as a test for the robustness of the model to assumptions about the missing data mechanism rather than a method for MNAR data; if the parameter estimates are little affected by the choice of missing data mechanism, the point estimate, ignoring the missing observations, is considered robust to missing data.

Clinicians are advised, or even required by study protocol, to collect information on the cause of patient dropout, but this information is generally used to justify the assumptions about the missing data mechanism rather than incorporated into the analysis. In chapter 6 a model is proposed to handle MNAR dropout in which information on the cause of dropout is fully exploited. Clinicians' knowledge and uncertainty about the missing data, based on the cause of patient dropout, is quantified and incorporated into the model. This information is elicited from clinicians in the form of prior distributions, and a Bayesian analysis is performed.

In contrast to the sensitivity analysis, the Bayesian model provides clinicians with point estimates and associated credible intervals that incorporate their uncertainty due to dropout in a formal manner. In a sensitivity analysis, clinicians are presented with a range of plausible parameter estimates which they must incorporate with their own beliefs about the dropout mechanism in an ad hoc way. Some may prefer this approach, on the grounds that the Bayesian point estimate, averaged over plausible dropout mechanisms, applies to no particular dropout scenario and, in practice, such a combination of dropout mechanisms could be impossible.

The sensitivity of the Bayesian model to mis-specification of the prior distribution, under various missing data mechanisms, is tested using simulation studies, and clinicians are consulted on the practicalities of elicitation of information about dropout bias.

Chapter 7 provides a discussion of the work in the thesis, highlighting the major findings of the research, and suggests directions in which future research could be taken.

Chapter 2

Review of methods for the analysis of clustered data, and introduction to missing data issues

2.1 Introduction to modelling clustered data

Clustered data arise when, for example, repeated measures are taken on each subject (longitudinal data), or when data are collected from different GP practices, hospitals, schools, communities etc.. In the latter case, treatments may be randomly assigned either by cluster, in which case this is termed a cluster randomised design, or by individuals within the cluster. In longitudinal studies the clusters are the individuals, each with several observations at different time-points, and typically there are a large number of clusters, each with only several observations. The observations within subjects tend to be fairly highly correlated because observations on the same subject are likely to be much more similar than observations between subjects [5]. Studies where the subjects are clustered by GP practice etc. are characterised by a relatively small number of clusters compared to the number of observations within clusters, and designs of this type tend to have a smaller within-group correlation than in longitudinal studies.

Classic generalised linear models (GLMs) cannot in general be used to model clustered data because they require an assumption of independence between all observations. The most common statistically valid methods of analysis for clustered data are random effects models, GEE models and the summary statistics approach, described below.

2.2 Random effects models

It is expected that in clustered data, observations on the same cluster will be correlated, and that there will be less variability between observations within the same cluster than between observations across clusters. In random effects models, also called hierarchical models or multilevel models, the heterogeneity between clusters is modelled by allowing certain model parameters to vary randomly between clusters [6]. This imposes correlation between observations within clusters. It also allows the between- and within- group variances to be estimated. In the simplest case, the intercept takes on different values for each cluster, while the other regression coefficients remain fixed. For example a random intercept model for gaussian data is as follows:

$$\begin{aligned} Y_{ij} &= (\alpha + u_i) + \beta \mathbf{x}_{ij} + \epsilon_{ij} \qquad (2.1) \\ u_i &\sim N(0, \sigma_u^2) \\ e_{ij} &\sim N(0, \sigma_e^2) \end{aligned}$$

where y_{ij} and x_{ij} are the observations on the outcome variable and explanatory variable respectively for the jth observation on the ith cluster. α represents the underlying average intercept for all clusters, and u_i gives the deviation of the ith cluster's intercept from the overall average, α. The regression coefficients, β and α are termed fixed effects while u_i and ϵ_{ij} are random effects with variances σ_u^2, the between-cluster variance, and σ_e^2, the within-cluster variance, respectively. The intra-cluster correlation coefficient (**ICC**) is defined as the proportion of the total variance that is between-cluster variance:

$$ICC = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} \tag{2.2}$$

The ICC is a positive coefficient ranging from zero, when the observations within clusters are independent, up to one, for perfect correlation within clusters.

More complex models allow several regression coefficients to vary between clusters. A more general linear random effects model is:

$$Y_{ij} = (\alpha + u_i) + \sum_k (\beta_k z_{ijk} + \delta_{ik} z_{ijk}) + \sum_l \gamma_l x_{ijl} + \epsilon_{ij}$$

where α, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are fixed effects and u_i, $\boldsymbol{\delta_i}$ and ϵ_{ij} are random effects. In this example, z_{ijk} denotes the kth variable with both fixed and random effects, and x_{ijl} the lth variable with only fixed effects. The fixed effects measure the average effect of each covariate across the clusters, while the random effects give estimates of the deviation from the average effect, in the ith cluster.

Data with any distribution from the exponential family can be modelled using a random effects model, by applying the same link function as in the GLM. As in the GLM framework, further complications to the structure of the data can be incorporated into the model, such as interaction terms, quadratic terms etc..

Extension to multiple levels of clustering, such as longitudinal data clustered by GP practice, is straightforward. Hence the terms multilevel models and hierarchical models.

Random effects models are sometimes described as cluster-specific models because the random effects allow each cluster to have its own regression equation. The fixed regression coefficients are interpreted as the expected change in the outcome within a specific cluster for one unit change in the covariate.

Estimation of parameters is by maximum likelihood methods or, where the number of clusters is small, by restricted maximum likelihood methods (REML). Where estimation is by maximum likelihood, significance of the parameters can be tested by either Likelihood ratio or Wald tests. Only Wald tests are valid if REML methods are used [7].

The random effects model is robust to missing data that are missing at random (MAR). This is because the inference is likelihood-based, and the likelihood of the complete data factorises into terms dependent on missing-data mechanisms and terms that depend on the model parameters [8]. This is explained further in section 2.6.2.

2.3 GEE models

Generalised estimating equations (GEE), or *marginal models*, as implied by their name, model the marginal expectation of the response across all clusters. Whereas random effects models explicitly model the correlation within clusters by allowing parameters of the model to vary randomly between clusters, generalized estimating equations treat this correlation as a nuisance and adjust for it. The correlation is specified separately to the regression model, in an ad hoc way rather than parametrically. A structure for the "working" correlation matrix, $\mathbf{R}_i(\alpha)$, which is assumed to be the same for each cluster, is chosen by the user. The model is relatively robust to mis-specification of this working correlation matrix.

The score-like equations solved in GEE models are: [4]

$$\mathbf{U_i} = \mathbf{D_i^T V_i^{-1} S_i} = \mathbf{0} \tag{2.3}$$

where

$$\mathbf{V_i} = \frac{\mathbf{A}_i^{1/2}\mathbf{R}_i(\alpha)\mathbf{A}_i^{1/2}}{\phi}$$

In this representation of the score equations $\mathbf{A}_i$ is a diagonal variance matrix of $\mathbf{Y}_i$ and $\mathbf{V}_i$ is the estimated covariance matrix of $\mathbf{Y}_i$. $\mathbf{S}_i$ is a vector of residuals for the ith cluster, $(\mathbf{Y}_i - \mu_i)$, and $\mathbf{D}_i = \frac{\partial \mu_i}{\partial \beta}$. The generalised estimating equations reduce to the maximum likelihood score equations when the response is gaussian, provided the relevant correlation structure is specified.

The generalised estimating equations are solved iteratively. Initially, a standard generalised linear model is fitted, assuming all observations are independent, to produce initial estimates of the regression parameters. Next the residuals, $\mathbf{S}_i$, are used to estimate $\boldsymbol{\alpha}$, the parameters of the working correlation matrix. The model is then re-fitted by solving equation 2.3 using this new working correlation matrix, and the parameter estimates are updated. This process of updating the correlation parameters and regression parameters is repeated iteratively until convergence is reached.

Liang and Zeger demonstrated that as long as the correct regression model is applied, the regression parameters are consistent even if the wrong correlation structure is chosen, although incorrect correlation structure leads to a loss in efficiency [4]. They also proposed a robust estimator of the standard error of the regression coefficients, the sandwich estimator, that is consistent under mis-specification of the correlation matrix:

$$(\sum_{i=1}^{N} \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_i)^{-1} [\sum_{i=1}^{N} \mathbf{D}_i^T \mathbf{V}_i^{-1} cov(\mathbf{y}_i) \mathbf{V}_i^{-1} \mathbf{D}_i] (\sum_{i=1}^{N} \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_i)^{-1} \tag{2.4}$$

When clusters are of equal size, the parameter estimate in the GEE model is the same as that specified in a GLM, treating all observations as independent, but the same is not true for the standard error.

There are several standard choices for the structure of the working correlation matrix, for example [6] [9]:

- Independence: The correlation between different observations in the same cluster is zero. The independence GEE model is equivalent to the generalized linear model. If ρ_{jk} is the correlation between the jth and kth observations in the same cluster, then $\rho_{jj} = 1$ and $\rho_{jk} = 0$ where j$\neq k$.

- Exchangeable: Each observation within a cluster is equally correlated with every other observation within that cluster. Here $\rho_{jj} = 1$ and $\rho_{jk} = \alpha$ if j$\neq k$, where α is an estimate of the intra-cluster correlation coefficient. A model with this correlation matrix is equivalent to a random effects model with a random intercept.

- Unstructured: The correlation between each pair of observations is allowed to take any value between -1 and +1, and is estimated from the data. $\rho_{jj} = 1$ and $\rho_{jk} = r_{jk}$ where j$\neq k$.

- Autoregressive: This applies only to repeated measurements data. The correlation between pairs of observations in the same cluster decreases as the time between the measurements increases. For first order autoregressive correlation, $\rho_{jj} = 1$ and $\rho_{jk} = \alpha^{|j-k|}$ for observations $|j-k|$ time-points apart.

- User fixed: The user sets the value of the correlation between all pairs of observations prior to the analysis.

Likelihood-ratio tests are not appropriate for the GEE model because the approach is not likelihood-based, and Wald tests are used instead [6].

The approach described above is called GEE-1 methodology because only first order moments are estimated, and parameters of the correlation are treated as nuisance parameters. Extensions to this approach have been developed which estimate the first and second order moments, and these are termed GEE-2 methods [10]. These methods can add efficiency to the analysis, but have the disadvantage that the correlation structure must be correctly specified in order to obtain unbiased estimates of the model parameters.

2.4 Summary statistics method

The simplest approach to dealing with clustered data is to reduce the observations in each cluster to a single measurement, and analyse this set of summary statistics as if they were the raw data. This removes from the data any correlation within clusters and offers a statistically valid approach if a sensible choice of summary measure is made and any missing observations are missing completely at random (MCAR).

The choice of summary statistic depends on the shape of the data and the particular research question. For example, if a researcher is interested in which of several treatments acts the fastest, the time to reach a maximum or minimum value may be a suitable summary statistic. The mean response is a common choice of summary measure, but where measurements are made at unequal points in time, the area under the curve is a better statistic to use. Matthews et. al. [11] offer a discussion of various choices of summary measure. Whatever the choice of summary statistic, this decision should be made in the design stage, before the results are obtained.

The simple summary statistics method, with the mean as the summary measure, weights all clusters equally, regardless of the number of observations in the cluster. This ignores the fact that clusters of different size estimate the summary measure with different accuracy, and may result in a loss of efficiency. An alternative approach is to weight clusters according to their size, but this ignores the correlation within clusters, and provides the correct inference only if observations within clusters are independent. A correct weighting would take into account the correlation structure within clusters, and weighted summary statistics approaches have been proposed that do just this [12]. Such methods do not have the appeal of simplicity that the unweighted summary statistics approach offers.

2.5 Comparison of random effects and GEE models

2.5.1 Gaussian data

For gaussian data the generalized estimating equations, with a linear regression model, reduce to the score equations used in random effects models [4], so that although the two types of models are specified differently, the same model is applied. The choice of correlation matrix in the marginal model determines the structure of the corresponding random effects model. For example, a marginal model with independent correlation is equivalent to a generalized linear model, and with exchangeable correlation corresponds to a random-intercept model.

In theory, equivalent marginal and random effects models produce identical regression parameters, but where there is missing data this is not necessarily the case. The issue of missing data is discussed in detail in section 2.6. Aside from missing data considerations, the choice between types of model for gaussian data comes down to whether there is direct interest in the variance and correlation parameters, and perhaps whether a likelihood-based method is preferred.

2.5.2 Likelihood based inference

An important distinction between random effects and GEE models is that the former are likelihood-based and the latter are not. There are various merits and limitations associated with each model for this reason. Likelihood-based inference is well researched and therefore methods of estimation and significance testing are well founded; for example, maximum likelihood methods can be used for parameter estimation and model adequacy can be assessed using likelihood ratio tests.

Random effects models are only valid if certain assumptions about the data hold. In particular, the random effects must be normally distributed, within- and between- cluster residuals must be independent, and random effects must be homogeneous [7]. GEE models are more robust than random effects models to deviation from these assumptions. But this robustness of GEE methodology is at the cost of efficiency: Because marginal models avoid specifying the full distribution, by specifying only the mean and covariance structure, they do not make full use of all the available information and in general are less efficient than random effects models [8] [13].

As discussed in section 2.6.2, likelihood-based methods such as random effects models, are robust to missing data that are missing at random (MAR), whereas GEE models, in general, are not.

2.5.3 Model fitting in random effects models

Parameter estimation is generally simpler under GEE methodology than in random effects modelling because it is often impossible to directly evaluate the likelihood in random effects models, particularly for non-gaussian data. This is because the likelihood is determined by integrating the joint distribution of the outcome variable, $\mathbf{Y}_{ij}$, and the random effects, $\mathbf{U}_i$, with respect to $\mathbf{U}_i$, and this integral is no longer of closed form for link functions other than the identity function.

Numerical integration methods, linear approximations to the link function, or computer intensive methods based on the bootstrap or the Gibbs sampler have been developed to fit random effects models for non-gaussian data [14].

Numerical integration for random effects generalised linear models, such as Monte Carlo integration or quadrature methods, is only possible where the dimension of the random effects matrix, $\mathbf{U_i}$, is small [9]. As the dimension of the random effects increases above two dimensions, a huge number of quadrature points are needed for quadrature estimation and the process becomes extremely computationally intensive. Stata uses quadrature methods to estimate parameters in random effects logistic models, and SAS also uses quadrature methods in some of its algorithms. Bayesian methods are making numerical integration of higher dimensions much more feasible, and MCMC methods can be carried out in statistical packages such as WinBUGS or MLwiN.

Many statistical packages, including MLwiN, VARCL, SAS and HLM, use less computationally intensive linear approximation methods such as marginal quasi-likelihood (MQL) or penalised quasi-likelihood (PQL) for random effects generalised linear modelling [15]. These methods approximate the link function to a linear function using Taylor expansion. The difference between MQL and PQL is that MQL iterates on the fixed effects only while PQL makes use of the current estimators of both the fixed and random effects in its iterations. Second-order MQL and PQL include first- and second-order derivatives whereas first-order quasi-likelihood methods approximate only as far as the first term in the Taylor expansion.

2.5.4 Interpretation of parameters in random effects and GEE models

Random effects models estimate conditional means, where the conditioning is on the random effect, for example:

$$g[E(Y_{ij} \mid u_i)] = X_{ij}\beta_C + u_i \tag{2.5}$$

In contrast, GEE models produce marginal estimates, averaged across all clusters. The marginal equivalent to equation 2.5 is:

$$g[E(Y_{ij})] = X_{ij}\beta_M \tag{2.6}$$

When the model is linear, parameters of the conditional model, β_C, and the marginal model, β_M, are in general the same [4]. For non-linear models, such as the logistic regression model, the parameters are different and have a different interpretation. The conditional estimate, β_C, is the expected influence of the covariate on the response of a cluster with a particular value of random effect. The marginal estimate, β_M, is the expected influence of the covariate on the response averaged over all clusters. Because of this, random effects models are commonly referred to as cluster-specific, and GEE models as population-averaged.

To understand why random effects models and GEE models provide different parameter estimates for non-linear models, consider the logistic regression model. The random effects model fits a separate logistic regression model to each cluster with a unique random effect, estimating a separate odds ratio for each covariate for each cluster. The overall odds ratio from the random effects model is the expected value of these individual odds ratios. In contrast, the GEE model estimates one marginal odds ratio for each covariate, as the ratio of the mean odds across all clusters. The random effects model estimates the mean of many odds ratios while the GEE model estimates the ratio of the mean odds, resulting in two different estimates.

Zeger et. al. [16] introduce a simple equation that can be used to approximately convert a cluster-specific coefficient to a population-averaged coefficient, as follows:

$$\beta_{PA} = \frac{\beta_{CS}}{\sqrt{1 + c\sigma_u^2}} \tag{2.7}$$

where

$$c = \left(\frac{16\sqrt{3}}{15\pi}\right)^2$$

β_{CS} is the cluster-specific coefficient and β_{PA} is the coefficient with a population-averaged interpretation. σ_u^2 represents the between-cluster variance. This approximation relies on the assumption that the logistic function can be approximated to the cumulative gaussian function. The smaller the between-cluster variance, the better the approximation.

It is widely stated in the literature that the choice between marginal and conditional models should be based on whether the question of interest requires a population-averaged or cluster-specific estimate. For example, in an epidemiological study, a marginal comparison between two populations of subjects, such as an exposed and an unexposed group, may be required. In contrast, the random effects model allows inference to be made about the effects of explanatory variables on an individual.

Lindsey and Lambert [17] dispute the argument that if the required prediction is for population-averaged inference, then a marginal model should be fitted. They argue that the marginal model is only population-averaged if the sample in the study is representative of the population of interest. Clinical trials are given as an example of when the participants are not a representative sample, as participation in the study is voluntary. They are highly critical of GEE models, giving several examples of where they may fail. The basis of most of their arguments is that GEE models treat the data from longitudinal studies as cross-sectional and do not adequately model the change in individuals over time. For example, small mean differences could mask heterogeneous differences amongst subjects. Lee and Nelder [18] continue the arguments of Lindsey and Lambert, going as far as to recommend random effects models as the analysis of choice over GEE models, because of the issues raised by Lindsey and Lambert, and also because of the difficulty in checking assumptions of models that are not likelihood-based. They also argue that model selection is a separate stage in the analysis to the prediction stage, and that each stage does not necessarily need to use the same model. Whether the inference required is population-averaged or cluster-specific is something that should govern the model used in the prediction stage.

Neuhaus et. al. [19] discuss the appropriateness of GEE models to different applications. Like Lindsey and Lambert, they argue that if an important predictor changes within clusters, for example the age of subjects within GP practices, a random effects model is more informative because it provides information about the variation of the effect of age within GP practice, while a population-averaged estimate will miss any heterogeneity of the effect within clusters. But they are not without criticism of the random effects model, using the cluster-randomised trial as a scenario in which parameters from random effects models may be difficult to interpret. A random effects model assumes that there are underlying latent groups on which the model conditions, and parameter estimation is with reference to these groups. Interpretation of cluster-specific parameters may be difficult if these latent groups cannot be identified.

Further to the debate about conditional versus marginal models, Begg and Parides [20] raise an interesting point. They argue that between-cluster and within-cluster effects can be estimated using either random effects models or GEE models as long as an appropriate model is constructed. For example, a marginal model could provide estimates of the within-cluster effects if, for each cluster, the covariate centred by the cluster-level mean, $(X_{ij} - \overline{X}_i)$, is included as a covariate in the model:

$$E(Y_{ij}) = \overline{X_{ij}}\beta + (X_{ij} - \overline{X}_i)\gamma \tag{2.8}$$

2.5.5 Versatility of the models

The hierarchical structure of random effects models extends to more than two levels. This allows data with a more complex structure to be analysed, such as clusters nested within clusters, for example, repeated measures on subjects clustered by general practice. It also allows multivariate data to be analysed by introducing extra outcome variables at the appropriate level in the hierarchy [21].

GEE models have the advantage of being robust to the mis-specification of the correlation structure. If this structure is not known, or is complicated in a way that makes modelling it via a random effects model difficult, a GEE model may be preferable [21] [22].

2.6 Missing observations in clustered data

2.6.1 Classifying missing data

Little and Rubin [23] classified missing observations into three types, missing completely at random (MCAR), missing at random (MAR) and informative or missing not at random (MNAR). The complete set of observations that would have been observed had there been no missing data, $\mathbf{Y}^*$, is partitioned into observed and missing data, $\mathbf{Y}^* = [\mathbf{Y}^O, \mathbf{Y}^M]$. $\mathbf{M}$ is a random variable that indicates whether an observation is missing ($M_{ij} = 1$) or observed ($M_{ij} = 0$) and the missing data mechanism is denoted $f(\mathbf{M} \backslash \mathbf{Y}^*, \phi)$, where ϕ represents the parameters of the missing data mechanism. The three categories of missing data mechanism are then defined as follows:

- **Missing completely at random (MCAR)**: The missing data mechanism is independent of both observed and unobserved data, $\mathbf{Y}^O$ and $\mathbf{Y}^M$:

  $$f(\mathbf{M} \mid \mathbf{Y}^*, \phi) = f(\mathbf{M} \mid \phi) \qquad \textit{for all } \mathbf{Y}^*,\ \phi \tag{2.9}$$

  Under MCAR, analysis of the available data provides valid inference, i.e. estimates are asymptotically unbiased, if less efficient.

- **MCAR conditional on the covariates**: A more general case is where the missing data mechanism depends on one or more of the observed covariates, $\mathbf{X}_1, ...\mathbf{X}_p$, and is independent of the observed and unobserved outcomes, $\mathbf{Y}^*$, conditional on the observed covariates. Analysis of the available data is then valid if these covariates are included in the model of interest [24]. Little calls this missing data mechanism

covariate-dependent missing, distinguishing it from true MCAR. Formally:

$$f(\mathbf{M} \mid \mathbf{Y}^{*}, \mathbf{X}_1, ...\mathbf{X}_p; \phi) = f(\mathbf{M} \mid \mathbf{X}_1, ...\mathbf{X}_p; \phi) \qquad (2.10)$$

- **Missing at random (MAR)**: The missing data mechanism is independent of the unobserved data but depends on the observed data, $\mathbf{Y}^O$:

$$f(\mathbf{M} \mid \mathbf{Y}^{*}, \phi) = f(\mathbf{M} \mid \mathbf{Y}^O, \phi) \qquad \textit{for all } \mathbf{Y}^M,\ \phi \qquad (2.11)$$

- **Missing not at random (MNAR)**: The missing data mechanism is dependent on the unobserved data, $\mathbf{Y}^M$.

In a hypothetical study of blood pressure in patients with hypertension, data would be classified as MNAR if, for example, a subject failed to attend clinic for follow-up because they had an acute exacerbation of hypertension symptoms, which could not be predicted from their preceding blood pressure measurements. Alternatively, if repeatedly low blood pressure readings predicted that a subject would go on to drop out of the study, their observations would be MAR. In an ideal situation, patients' failure to attend their follow-up appointments is independent of both their observed and unobserved readings, in which case the data are classified as MCAR. The assumption of MCAR dropout would be a reasonable in cases where a patient's dropout is unrelated to the study. For example, the patient moved away from the area in which the study was carried out.

Incomplete repeated measures, or longitudinal data, can be further categorised into data that are missing intermittently, and dropout. Dropout, also termed attrition or monotone missing, refers to a situation where once a subject has one missing observation, all following observations are missing. In contrast to intermittent missingness, it is often impossible to find out why subjects drop out of a study, because they do not attend clinic again. Diggle et. al argue that subjects who miss follow-up appointments intermittently are more likely to be missing at random than dropouts [9]. Clearly, this assertion will not always hold. More importantly, it should be possible to record reasons for their missing observations, which can be used to judge the type of missing data mechanism present, or used directly in the analysis, as demonstrated in chapter 6.

2.6.2 Likelihood-based inference and missing data

Likelihood-based methods, such as the random effects model for clustered data, produce an unbiased analysis for data that are MAR, and the stricter assumption of MCAR is not necessary [23]. This follows from noting that, when data are MAR, the full likelihood of the observed data and missing data indicator, $L_{full}^{obs}(\theta, \phi \mid \mathbf{Y}^O, \mathbf{M})$, factorises into a function of the parameters of the missing data mechanism, ϕ, and a function of the parameters of the distribution of the observations, θ:

$$L_{full}^{obs}(\theta, \phi) = L_1(\phi) \times L_2(\theta) \tag{2.12}$$

Likelihood-based inferences for θ, from the full likelihood of the observed data and the missing data indicator, $L_{full}^{obs}(\theta, \phi)$, are then the same as inferences for θ from the likelihood of θ based on the observed data only, ignoring the the missing data mechanism, $L_{ign}^{obs}(\theta)$. Unbiased estimates of θ can then be obtained from the observed data. The proof of equation 2.12 is as follows.

The full likelihood function of the complete data, L_{full}^{comp}, is proportional to the joint density function of the response and the missing data mechanism, given by:

$$L_{full}^{comp}(\theta,\phi) \propto f(\mathbf{Y}^O,\mathbf{Y}^M,\mathbf{M} \mid \theta,\phi) = f(\mathbf{Y}^O,\mathbf{Y}^M \mid \theta) f(\mathbf{M} \mid \mathbf{Y}^O,\mathbf{Y}^M,\phi) \tag{2.13}$$

The full likelihood function for the observed data, $L_{full}^{obs}(\theta,\phi)$, is obtained by integrating the full likelihood function for the complete data in equation 2.13 over $\mathbf{Y^M}$, to give

$$\begin{aligned} L_{full}^{obs} \propto f(\mathbf{Y}^O,\mathbf{M} \mid \theta,\phi) &= \int f(\mathbf{Y}^O,\mathbf{Y}^M,\mathbf{M} \mid \theta,\phi) d\mathbf{Y}^M \qquad (2.14) \\ &= \int f(\mathbf{Y}^O,\mathbf{Y}^M \mid \theta) f(\mathbf{M} \mid \mathbf{Y}^O,\mathbf{Y}^M,\phi) d\mathbf{Y}^M \end{aligned}$$

Now if the data are MAR then from the definition given in equation 2.11, the likelihood for the observed data becomes:

$$\begin{aligned} L_{full}^{obs}(\theta,\phi) &\propto \int f(\mathbf{Y}^O,\mathbf{Y}^M \mid \theta) f(\mathbf{M} \mid \mathbf{Y}^O,\phi) d\mathbf{Y}^M \\ &= f(\mathbf{M} \mid \mathbf{Y}^O,\phi) \int f(\mathbf{Y}^O,\mathbf{Y}^M \mid \theta) d\mathbf{Y}^M \\ &= f(\mathbf{M} \mid \mathbf{Y}^O,\phi) f(\mathbf{Y}^O \mid \theta) \\ &= L_1(\phi) \times L_2(\theta) \qquad (2.15) \end{aligned}$$

$L_{full}^{obs}(\theta,\phi)$ and $L_{ign}^{obs}(\theta)$ are then the same, for inference about θ, as long as θ and ϕ are distinct. The term distinct means that the joint parameter space of θ and ϕ factorises into the product of the parameter space for θ and the parameter space for ϕ. If θ and ϕ are not distinct, inference for θ, from the observed data only, will be unbiased but less efficient.

If the response is incomplete and the missing data mechanism is believed to be MAR, the recommendation is to carry out a likelihood-based analysis on the observed data, ignoring the missing data mechanism. This is called an "available case analysis". The analysis becomes more problematic if there are missing observations in the covariates, because an available case analysis then excludes the whole case for subjects with missing observations on as few as one covariate. The issue of missing data in the covariates is beyond the scope of this work.

2.6.3 Generalized estimating equations and missing data

Because GEE models are not likelihood-based, in general they require data to be MCAR to ensure estimation is robust to mis-specification of the correlation matrix. If the correlation matrix is correctly specified, Liang and Zeger state that a MCAR assumption "may not be necessary" [4]. With MCAR dropout, the data are simply unbalanced in their cluster size, which GEE models, in theory, can handle [6]. The only issue with unbalanced data in GEE is that estimation of the correlation parameters may be biased, which Park demonstrates using simulations [2].

When the data are gaussian, generalised estimating equations reduce to the score equations in likelihood-based models and, provided the correct correlation structure is applied to the GEE model, the two types of model become equivalent [4]. Because likelihood-based models are robust to MAR data, this implies that GEE models can handle MAR data when the response is gaussian. Again, this relies on unbiased estimation of the correlation parameters, and with unbalanced data, GEE do not always provide this. This issue is discussed in greater detail, alongside the results of the simulation studies, in chapter 3, which demonstrate that GEE are fairly robust to MAR dropout if the correct correlation matrix is applied, but that mis-specification of the correlation structure can cause substantial bias.

A new class of weighted estimating equations has been developed, that are robust to MAR dropout in general, provided the probability of dropout, given the observed response, is modelled correctly [25]. The missing observations are effectively imputed by weighting the responses that are observed, analogous to the weighting in survey analysis. The weights applied to the observed elements of the response are the inverse of the probability of remaining in the study, conditional on past observations, estimated by logistic regression. For this reason the method is often referred to as inverse probability weighting (IPW). Heyting et. al. [26] provide a clear introduction to the concept of IPW.

In summary, standard GEE models require a MCAR assumption in general, but in certain circumstances may be biased in the presence MCAR data. For gaussian data, GEE models may be robust to missing observations with a MAR mechanism. Weighted GEE have been developed which, when a model for the probability of dropout is correctly specified, handle MAR missing data in general.

2.6.4 Testing for MCAR

The missing data mechanism should be taken into account when choosing the method of analysis for incomplete clustered data. GEE models require any missing observations to be MCAR in general, as does the summary statistics method.

The clinical causes of missing data should be considered when assessing the missing data mechanism. It is also possible to distinguish between MCAR and MAR data by examining the data. By definition, the observed data do not provide sufficient information to determine if the missing data are MNAR; clearly it is not possible to test if the missingness is dependent on the unobserved data. Inspection of the observed data helps to distinguish between MCAR and MAR data, and Carpenter et. al. [27] suggest plotting the means (± 2 standard errors) for the subjects who drop out at the next time-point compared to those that do not. Several formal tests of MCAR have also been published, and these tests are compared in detail in chapter 4.

2.6.5 Methods for MNAR data

Likelihood-based models are robust to MAR data if the missingness is in the response, but no models that ignore the missing data mechanism can handle MNAR data. It is important, therefore, to check that any findings based on MAR assumptions hold by carrying out a sensitivity analysis. In a sensitivity analysis, the missing data mechanism is incorporated into the model of interest [27] [28]. The missing data mechanism may be modelled using a pattern mixture model or selection model, described in section 2.6.6. The sensitivity of model estimates to changes in the missing data parameters is then examined. The model is fitted with a range of plausible missing data parameters, resulting in a range of plausible estimates for the parameters of interest. If the analysis is sensitive to the missing data parameters, the range of estimates for the parameters of interest should be reported.

If data are MNAR, the common recommendation is to carry out a sensitivity analysis and report a range of estimates for the parameters of interest. Fairly recently, Bayesian models have been proposed as alternatives to the sensitivity analysis [29] [30]. A pattern mixture model or selection model is fitted with informative priors on the missing data parameters. This incorporates the uncertainty due to the missing data into the posterior distribution of the parameters of interest. An extension of these models for repeated measurements is proposed in which the cause of dropout is included in the analysis, in chapter 6.

2.6.6 Modelling the missing data mechanism

Joint modelling of the missing data mechanism and the model of interest can be considered in a pattern-mixture model (PMM) or selection model (SM) framework, defined as follows:

$$\text{SM} \quad f(\mathbf{Y}, \mathbf{M} \backslash \mathbf{X}, \theta, \phi) = f(\mathbf{M} \backslash \mathbf{Y}, \mathbf{X}, \phi) f(\mathbf{Y} \backslash \mathbf{X}, \theta) \tag{2.16}$$

$$\text{PMM} \quad f(\mathbf{Y}, \mathbf{M} \backslash \mathbf{X}, \theta, \phi) = f(\mathbf{Y} \backslash \mathbf{X}, \mathbf{M}, \theta) f(\mathbf{M} \backslash \mathbf{X}, \phi) \tag{2.17}$$

where $\mathbf{M}$ is a dichotomous variable to indicate whether an observation is missing or observed. $\mathbf{Y}$ is a vector of the response, $\mathbf{X}$ represents the covariates, and θ and ϕ are parameters of the model of interest and parameters of the missing data mechanism respectively.

In the pattern-mixture model framework, the data are stratified according to which observations are missing in the response. The distribution of the full data is treated as a mixture of of distributions over these missing data "patterns". For example, subjects that have complete observations have a different distribution to those that have one missing observation, which is different again to the distribution of subjects with only one measurement observed. A separate model is fitted to each stratum, $f(\mathbf{Y} \backslash \mathbf{X}, \mathbf{M}, \theta)$, and an overall parameter of interest is computed by taking a weighted average of the stratum-specific parameter estimates, weighted by the proportion of observations in each stratum, $f(\mathbf{M} \backslash \mathbf{X}, \phi)$.

In contrast, the selection model assumes that the complete data are random samples from the same distribution, and that a selection of subjects drop out according to their response values. The distribution of the complete observations is f($\mathbf{Y}\backslash\mathbf{X}, \theta$), and the subjects that drop out are a selection of these complete data, with missing data mechanism, f($\mathbf{M}\backslash\mathbf{Y}, \mathbf{X}, \phi$).

2.6.7 Single imputation methods

Single imputation methods are in fairly common usage in the analysis of medical studies. Each missing observation is replaced with an imputed value to create a single "complete" dataset which is then analysed as if it were the original data. Common choices of imputed values are the last observed value on the subject (called "Last observation carried forward (LOCF)"), mean imputation in which the missing observation is replaced with the mean value for that variable, or in the case of dichotomous data, the "Mean equals failure" approach, in which missing observations are assumed to have failed.

These single imputation methods are sometimes sold as being conservative. For example, LOCF is believed by some to be conservative because on average it will reduce any slope that is estimated up until the point of dropout. All single imputation methods are statistically invalid. They have the potential to distort the covariance structure, and will tend to under-estimate the variance[55]. Even if the dropout mechanism is MCAR this is still clearly the case. If the standard errors are under-estimated, the method is not necessarily conservative and can in fact be anti-conservative. The next section explains the use of multiple imputation as a valid tool that allows us to estimate the additional uncertainty in the model due to the missing data.

2.6.8 Multiple imputation and inverse probability weighting

The pattern-mixture and selection models are approaches to full parametric modelling of the dropout mechanism. Multiple imputation (MI) and inverse probability weighting (IPW) are alternative methods that approximate these parametric models. Both techniques are valid provided the imputation model (in MI), or the dropout probability model (in IPW), is correctly specified. Inverse probability weighting is described in section 2.6.3. Multiple imputation draws samples from a distribution defined by a model for the dropout mechanism. Multiple values are imputed for each missing observation so that the between- and within-imputation components of variation can be incorporated in the estimates of the parameter standard errors using Rubin's rules [31]. Carpenter et. al. provide a theoretical and practical comparison of multiple imputation and inverse probability weighting [32].

2.7 Summary

This chapter is an overview of common approaches to the analysis of repeated measurements data, and an introduction to the main issues when there are missing observations in the data. Random effects models, generalised estimating equations and the summary statistics method have been described and the approaches compared. Different classifications of missing data have been defined, and the sensitivity of the models to each category of missing data has been discussed. Approaches to dealing with missing data are introduced. The theory in this chapter is referred to throughout the thesis, and many of the issues raised here are dealt with more thoroughly in later chapters.

Chapter 3

Sensitivity of marginal models, random-effects models and summary statistics methods to missing data

3.1 Introduction

The aim of this chapter is to explore the extent to which dropout affects the parameter estimates of methods for repeated measurements data. A thorough investigation is carried out using simulation studies, together with evidence from the literature. From the findings of the research, recommendations are made to the applied statistician, for the analysis of incomplete repeated measurements. The impact of dropout on the results of the asthma clinical trial are discussed in chapter 5, in the context of this work.

In theory, likelihood-based methods such as the random-effects model are unbiased for data that have missing observations missing at random (MAR). This follows from noting that when the data are MAR or missing completely at random (MCAR), the likelihood factorises into parameters of the distribution of the response, θ, and parameters of the missing data mechanism, ϕ, as described in section 2.6.2. The likelihood, based on only the observed data, is only fully efficient if these vectors of parameters, θ and ϕ, are distinct, i.e. there is no overlap in the two parameter vectors. Because GEE models and the summary statistics method are not based on likelihood theory, in general both approaches require missing data to be MCAR.

Theory also tells us that when the data are gaussian, generalised estimating equations reduce to the score equations in likelihood-based models and, provided the correct correlation structure is applied to the GEE model, the two types of model become equivalent [4]. According to Diggle, Liang and Zeger, equivalent marginal and random effects models for gaussian data produce identical regression parameters but, where there is missing data this statement will not always hold [9]. Park [2] demonstrated that generalised estimating equations do not always reduce to the score equations when the data are incomplete, even if the missing data are MCAR. This is because the estimation of the covariance matrix in the GEE model is no longer equivalent to that in maximum likelihood estimation when the clusters are of unequal size.

Liang and Zeger state in their seminal 1986 paper that if the correlation structure in a GEE model is correctly specified an assumption of MCAR "may not be necessary" [4]. This implies that if the wrong working correlation matrix is used in a GEE model of the simulated data, biased estimates may be produced when the data are not MCAR. With binary response data, the GEE of Liang and Zeger, therefore, require missing data to be MCAR, whereas MAR data will not cause the random effects model to be biased. Research has found [34], at least in limited scenarios, that GEE show little bias if the correlation is well estimated. There have been extensions to Liang and Zeger's GEE that attempt to improve the estimation of the correlation parameters, and this is discussed in section 3.4.2. The summary statistics method requires a stronger assumption of MCAR, for all types of response.

In practice, the extent to which missing observations cause bias to parameter estimates and affect their clinical and statistical significance is not clear. For example, it is not known how robust the GEE model is to misspecification of the working correlation matrix in the presence of missing data. It is also unclear how sensitivity to missing observations compares between data with cluster-level and cluster-varying covariates, what proportion of missing data leads to substantial bias under MAR and MNAR, and how the random effects model, summary statistics method and GEE compare in different scenarios.

Through simulation studies the effect on parameter estimates of missing data with various missing mechanisms was investigated. GEE models, random effects models and the summary statistics method were compared under various scenarios. The GEE model was fitted with several working correlation structures to investigate the sensitivity of the model to missing data under misspecification of the correlation structure. The summary statistics method was used as a comparison to the more sophisticated models, to give a sense of the maximum size of bias that could result from the missing data process. In addition, the summary statistics method is an approach that is employed in the analysis of repeated measurements from medical studies. A simple unweighted summary statistics method was implemented, rather than an approach that weights by a function of cluster size or variance within clusters. As discussed in section 2.4, the appeal of the unweighted summary statistics approach is its simplicity, which is not shared by the weighted summary statistics approach.

MAR and MNAR missing data mechanisms were compared for gaussian and binary data, with both cluster-level and cluster-varying covariates. By definition it is not possible for any method to be robust to every MNAR missing data mechanism, because the values of the missing observations are unknown and could potentially take any value. In practice, however, observations within clusters are correlated, and because of this it is possible that the bias caused by MNAR data may be small. The effect of MNAR dropout was investigated with various levels of intra-cluster correlation. The sample size and the proportion of observations missing were altered to investigate their impact on the parameter estimates. The strength of the ICC was considered important as it governs the amount of information carried by each observation in a cluster, and the effect of its size on the parameter estimates was also investigated.

For binary data, sensitivity of the parameter estimates to missing data under different levels of event probability, p, were compared. The models were fitted using the statistical package, Stata version 8. Two methods of estimation were compared for the random effects logistic regression model: Gauss-Hermite quadrature in Stata 8, and penalised quasi-likelihood (PQL) in MLwiN, as PQL has been known to produce biased results [7]. The fitting of binary logistic random effects models is described further in section 2.5.3. The correlation coefficient in the GEE was estimated using the standard Liang and Zeger method, as a function of the Pearson residuals of the data.

No simulation study into the impact of missing data on these methods of analysis of clustered data has been carried out in such depth.

The results of the simulation studies and explanations of the various findings are given in section 3.4. Conclusions of the study are summarised in section 3.6, and from the findings, recommendations are made to applied statisticians analysing incomplete clustered data.

The following scenarios were investigated:

Gaussian data

MAR data

1. Cluster-level covariate
2. Reducing the sample size
3. Cluster-varying covariate

MNAR data

4. Selection model (SM) simulated data
5. Pattern mixture model (PMM) simulated data
6. PMM simulated, data include period × treatment interaction
7. SM simulated, reducing the intra-cluster correlation to 0.2

Binary data

MAR data

8. Probability of event, p=0.5
9. Reducing the sample size
10. Probability of event, p=0.1

MNAR data

11. Success probability, p=0.5

3.2 Gaussian data

3.2.1 MAR observations

Scenario 1: Cluster-level covariate

Data simulation

The data were simulated based on a repeated measurements clinical trial, comparing two treatments over time. One thousand sets of longitudinal data were simulated with the following distribution:

$$y_{ij} = \alpha + u_i + \beta treat + e_{ij} \tag{3.1}$$

$$u_i \sim N(0, \sigma_u^2)$$

$$e_{ij} \sim N(0, \sigma_e^2)$$

where y_{ij} is the response of the ith subject on their jth visit. All complete datasets contain 4 observations on 200 subjects, unless otherwise stated. Values of β, σ_u^2 and σ_e^2 were chosen so that, on average, the statistical significance of β, the treatment effect, would be fairly strong. The size of $\beta/s.e.(\beta)$ is 2.5.

Observations were removed from the complete datasets with subjects dropping out of the study without returning. At least one observation was available for all subjects, with dropout occurring at periods 2, 3 or 4. Datasets were constructed with varying degrees of missingness, from 5% to 30% of observations missing. The results displayed are for datasets with 20% of observations missing. The intra-cluster correlation (ICC), as defined in equation 2.2, is a measure of the correlation between observations within clusters.

Unless otherwise stated, the data are simulated to have an ICC of 0.5, by setting the between- and within-cluster variances to be equal.

The same 1000 complete datasets were used to construct 1000 datasets with incomplete observations. The probability of an observation being missing was dependent on the value of z_{ij}, a function of previous observations in the cluster, with the following missing data mechanism:

$$logit(p_{ij}) = \alpha_j + \phi z_{ij} \tag{3.2}$$

where p_{ij} is the probability that the ith subject drops out at time j. The choice of z_{ij} depends on whether the missing data are MAR or MNAR. In the MAR scenarios $z_{ij} = y_{i,j-1}$, i.e. the previous value of the response predicts dropout. Where the missing data are MNAR, the current value of the response predicts dropout, $z_{ij} = y_{ij}$. The value of ϕ is large enough that subjects that drop out are the subjects with the lowest values of z_{ij}. In other words, the probability of dropout is strongly related to the response.

In the datasets with 20% of observations missing, on average the frequency of subject dropout at periods 2, 3 and 4 is given in table 3.1.

Table 3.1: Mean frequency of missing data patterns in datasets with 20% observations missing

Pattern	Dropout at period	Proportion subjects
XXXX	-	0.55
X . . .	2	0.1
XX . .	3	0.15
XXX .	4	0.2

Model fitting

The sensitivity of random effects models, GEE models and the summary statistics method to MAR data was compared. The data were simulated to have a random intercept but no other random effects, therefore a random intercept model was fitted for the random effects model. The correct correlation structure for the GEE model is exchangeable, which is equivalent to the random intercept model.

The summary statistics method is a simple approach used widely in medical research. The set of measurements from each cluster is reduced to an appropriate single summary measure, and these summary measures are analysed as if they were the raw data [11]. This removes the clustering effect from the data. The choice of summary measure, or *summary statistic*, depends on the research question, and the nature of the response. For example, the mean response over the cluster could be chosen as the summary statistic if clinical interest is in the average difference in response between two treatment groups. Time to maximum response, area under the curve, or rate of change of response are other possible choices of summary statistic.

In this scenario, the summary statistics method was implemented by reducing the vector of observations for each subject to the mean of observations for each subject, as the overall efficacy of the treatments over time is of interest. The mean in the summary statistics method is directly comparable to the estimates of the random effects and GEE models. The measurements are equally spaced in time, and therefore there is no need to weight the observations on a subject according to their spacing. The treatment effect was estimated to be the difference in means between the two treatment groups.

The relevant hypothesis test to test for a non-zero treatment effect, was a two sample t-test.

Results

The results of the models fitted to incomplete gaussian MAR data with a cluster-level covariate, compared to the same 1000 complete datasets, are summarised in figure 3.1. The bias is computed by comparing the estimate of the treatment effect to the estimate obtained for the complete data from the random effects model (or, equivalently, the GEE exchangeable model). The bias compared to this sample treatment effect reflects the bias introduced by the missing observations rather than including the sampling error. The efficiency is the ratio of the empirical standard error of the parameter estimate, $\hat{\beta}$, from a correctly specified random effects model fitted to the complete data, compared to the empirical standard error from each model fitted to the incomplete data. The next row of the table is the size of the mean bias compared to the empirical standard error of the estimate. The final line of the table contains the proportion of datasets where the parameter estimate is statistically significant at the 5% level, and the parameter estimate is in the correct direction. In other words, a parameter estimate that is negative when the true parameter is positive would not be counted as significant. For the complete data, with a correctly specified random effects or GEE exchangeable model this proportion was 0.71.

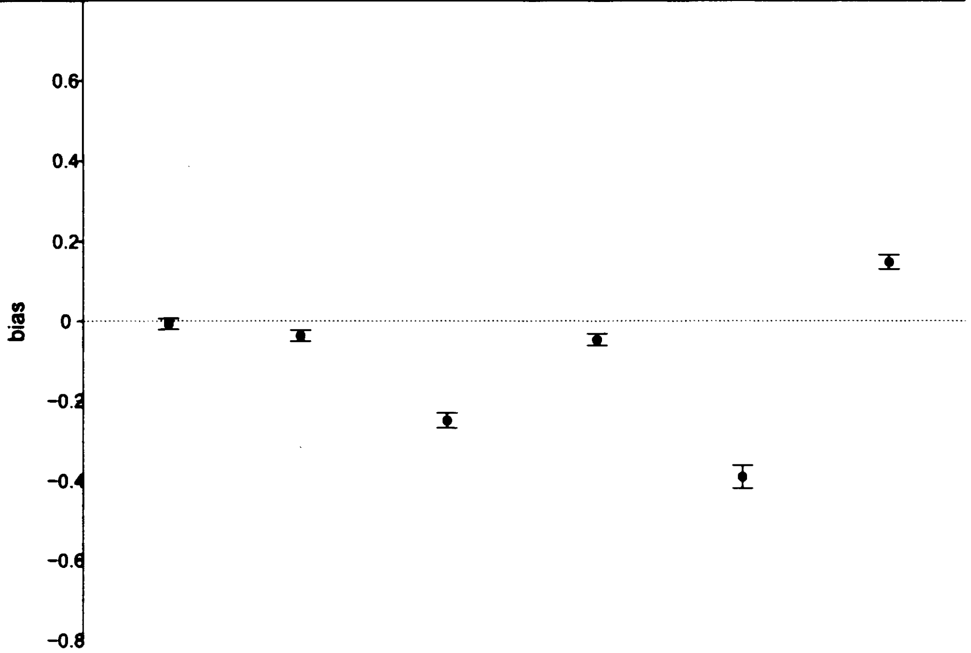
Plotted below the table are the mean bias and 95% confidence intervals for the mean bias, for each model. These confidence intervals are an estimated range for which 0.95 of all mean biases would lie within, if many sets of 1000 datasets were simulated. The intervals can be used to assess the statistical significance of any bias in the estimates. The proportion of the 1000 parameter estimates significant at the 5% level allows us to assess what effect the bias would be likely to have on decisions made by clinicians as a result of the findings of a similar repeated measurements study with missing data. This, in contrast to the statistical significance of the bias, demonstrates whether the bias is large enough to affect clinical decisions. This can be considered "clinical significance" as opposed to "statistical significance". The mean bias of the parameter estimate compared to its standard error gives further insight into the size of the bias, in this case providing information about its size compared to uncertainty in the parameter estimate due to sampling error.

The table below the bias plot summarises the bias and efficiency in the random effects, σ_u^2 and σ_e^2, and the effect of any bias on the ICC.

The results demonstrate that the random effects model is unbiased with 20% observations MAR, for both fixed and random effects, for a cluster-level covariate. The GEE model is significantly biased even when the correlation structure is correctly specified as exchangeable, or an unstructured correlation matrix is used. The size of this bias is very small compared to the the standard error of the parameter, and has no effect on the proportion of datasets where the estimate is statistically significant. The GEE model performs worst when the wrong correlation structure is applied, and the proportion of datasets where the estimate is significant at the 5% level is reduced quite substantially, The summary statistics method is also significantly biased. The size of the mean bias for the GEE models with incorrectly specified correlation structure, and the summary statistics method, is substantial in

Figure 3.1: Scenario 1. The effect of 20% MAR missing observations on gaussian data with a cluster-level covariate.

	Hierar-	**GEE Model**				**Summ.**
	chical	**Exch**	**Indep**	**Unstr**	**A-R**	**Stats.**
Efficiency (%)	91	93	109	94	104	76
Mean bias/s.e.	-0.009	-0.045	-0.329	-0.057	-0.506	0.163
Proportion sig.	0.67	0.67	0.64	0.67	0.55	0.66



Random effects estimated in random effects model

ICC		Efficiency (%)		Bias (95% C.I.)	
Comp. data	Inc. data	σ_u^2	σ_e^2	σ_u^2	σ_e^2
0.498	0.497	84	84	-0.104 (-0.40,0.20)	0.041 (-0.11,0.19)

comparison to the empirical standard error of the estimate.

Increasing proportions of missing data were simulated and, as expected, the bias in the parameter estimates tends to increase with the proportion of missing observations. The bias becomes substantial in comparison to its standard error with around 10% of observations missing when a GEE model with incorrect correlation is implemented. With as little as 5% of observations missing the GEE are significantly biased, even with the correct correlation structure chosen.

The summary statistics method is less efficient than the more sophisticated GEE and random effects models. The efficiency of the random effects and GEE models is not substantially affected by the missing data, and actually increases for the independence autoregressive GEE, despite the decrease in the sample size. This increase in efficiency is discussed in section 3.4.4.

Scenario 2: Reducing the sample size

The same models were fitted to the simulated data with the number of cases in each dataset reduced to 100 and to 40. All parameter values were kept the same, so that the standard error became large in comparison to the size of the treatment effect.

Reducing the sample size had no effect on the findings. As with the larger datasets, the bias tends to increase as the proportions of missing observations increases. The random effects model and correctly specified GEE model had unsubstantial bias while the greatest bias is caused by the GEE model with independent or autoregressive correlation, and the summary statistics method.

Scenario 3: Cluster-varying covariates

The treatment effect in the above models is termed a cluster-level covariate because it does not vary within clusters. In this scenario, the sensitivity of the models to missing data was explored when the covariate varies within clusters. Time period was introduced into the model as an extra covariate that varies within clusters. As the data are longitudinal, the subjects are the clusters and time varies within subject.

Data simulation

Time period, equal to j, is treated as a continuous variable, and the response is simulated to change linearly over time. The model from which the data are simulated is as follows:

$$y_{ij} = \alpha + u_i + \beta_1 treat + \beta_2 time + e_{ij} \tag{3.3}$$

$$u_i \sim N(0, \sigma_u^2)$$

$$e_{ij} \sim N(0, \sigma_e^2)$$

The value of β_2 was chosen so that the size of $\beta_2/s.e.(\beta_2)$ was 2.5, as it was for the cluster-level covariate, treatment, in section 3.2.1.

Model fitting

A random effects model and GEE models were fitted to the data, with the same covariates that the data were simulated from. A summary statistics method was also applied to the data. This time, because the parameter of interest is change in the outcome over time, a suitable summary statistic is an estimate of the slope. The slope was estimated by ordinary least squares. In the incomplete data some subjects have only one observation, that taken at period 1, and for these subjects there is no estimate of the slope. The estimate of the parameter of interest, $\hat{\beta}_2$, is the mean of all the slopes. The model standard error of β_2 is the standard definition of the standard error of a mean, i.e. the sample standard deviation of the estimates of the slope divided by the square root of the number of slope estimates. The results are summarised in figure 3.2.

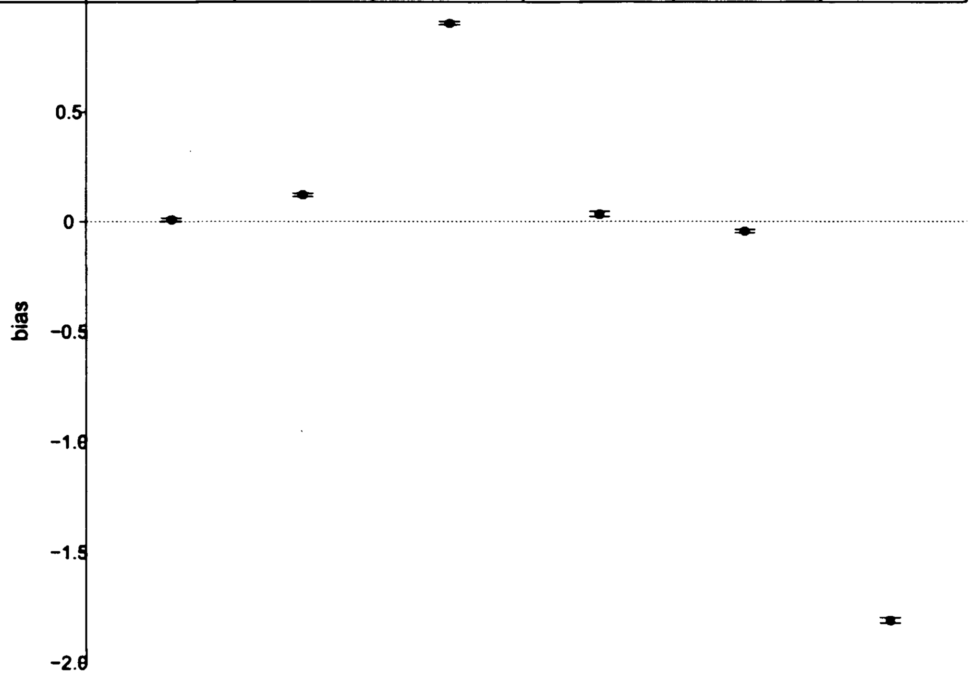
Results

When the covariate varies within cluster, the bias is in general greater than when the covariate is fixed over the cluster. The random effects model is virtually unbiased, but this bias is borderline significant.

The performance of the GEE model is again poor in comparison to the random effects model, especially when the wrong correlation structure is chosen. With an exchangeable or unstructured working correlation matrix the bias is significant but smaller than the standard error of the parameter. The bias in the correctly specified GEE is significant even with 5% of observations missing. With the wrong correlation structure, in the presence of missing data, the estimates become extremely biased. For example, with 20% of observations missing, the parameter estimate obtained from the independence

Figure 3.2: Scenario 3. MAR missing observations in gaussian data with a cluster-varying covariate.

	Hierar-	**GEE Model**				**Summ.**
	chical	**Exch**	**Indep**	**Unstr**	**A-R**	**Stats.**
Efficiency (%)	61	60	59	41	57	36
Mean bias/s.e.	0.046	0.607	4.495	0.138	-0.215	-7.047
Proportion sig.	0.53	0.77	1.00	0.61	0.46	0.00



Random effects estimated in random effects model

ICC		Efficiency (%)		Bias (95% C.I.)	
Comp. data	Inc. data	σ_u^2	σ_e^2	σ_u^2	σ_e^2
0.495	0.520	81	83	1.94 (1.6,2.2)	-0.595 (-0.74,-0.45)

GEE model is more than three times the true value of the parameter. The summary statistics method performs worst of all. The bias is significant with as little as 5% of observations missing at random.

The efficiency of the methods is substantially decreased by the missing data, even for the random effects model which is unbiased, but especially for the unstructured GEE and the summary statistics method. The proportion of parameters significant at the 5% level was, again, 0.71 in the complete data. This proportion is affected by the missing data in all models, even in the random effects model, which is unbiased but less efficient because of dropout.

3.2.2 Missing data that are missing not at random

Scenario 4: Selection Model simulated data

The incomplete datasets so far have contained missing observations that are missing at random (MAR). Because the data are gaussian, in theory both random effects and marginal models should be robust to this missing data mechanism. However, neither type of model can, in theory, handle missing data that are missing not at random (MNAR). It is unclear, however, to what extent MNAR dropout will affect the parameter estimates because of the correlation within subjects. The models were fitted to datasets with a MNAR missing data mechanism to test how sensitive they are in practice to MNAR missing data.

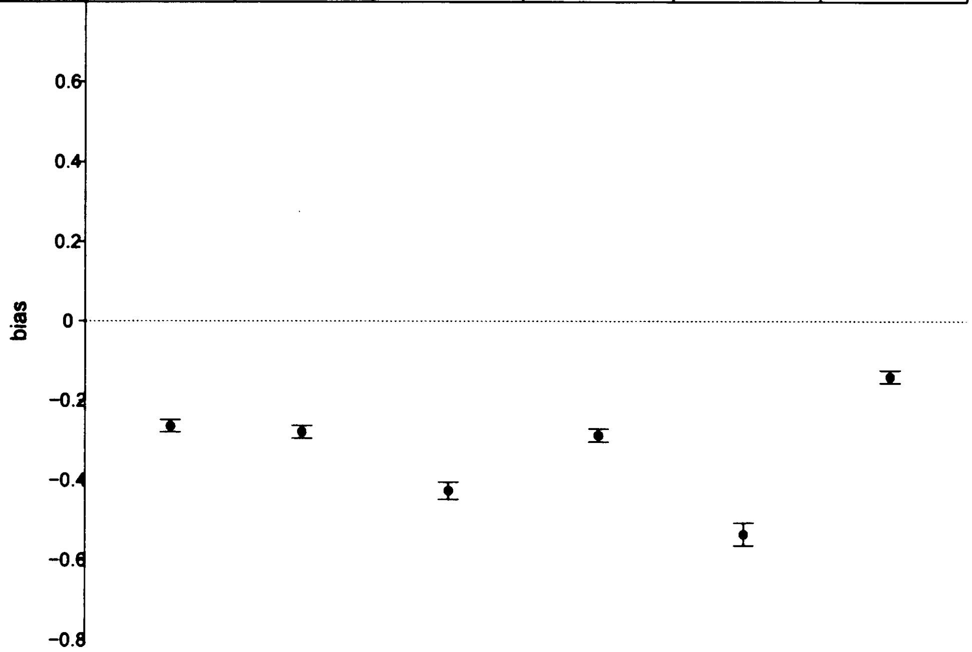
The same complete datasets were used to construct datasets with a MNAR missing mechanism. The probability of an observation being missing was dependent on the value of that observation; in equation 3.2 $z_{ij} = y_{ij}$.

The same models were fitted to the data here as to the data with a MAR missing data mechanism.

Results

Figure 3.3: Scenario 4. MNAR selection model simulated data

	Hierar-chical	GEE Model				Summ. Stats.
		Exch	Indep	Unstr	A-R	
Efficiency (%)	116	117	125	119	118	102
Mean bias/s.e.	-0.361	-0.382	-0.604	-0.397	-0.737	-0.179
Proportion sig.	0.67	0.67	0.63	0.67	0.55	0.66



Random effects estimated in random effects model

ICC		Efficiency (%)		Bias (95% C.I.)	
Comp. data	Inc. data	σ_u^2	σ_e^2	σ_u^2	σ_e^2
0.498	0.447	103	91	-6.76 (-7.0,-6.5)	-2.64 (-2.8,-2.5)

There is generally more bias with a MNAR mechanism than with MAR data. All models, including the random effects model, are significantly biased when the data are MNAR, even with as little as 5% of observations missing. The findings are the same for fixed and random effects. There is very little difference in the performance of the models; in the scenario simulated, the summary statistics method produces the least biased estimates, although this bias is significant and substantial. The bias due to MNAR dropout is so great that the choice of correlation structure in the GEE becomes relatively unimportant.

The efficiency of all models increases when the data are incomplete, even though there are less available data. Notice that the within- and between-variances in the random effects models are negatively biased, demonstrating that there is less overall variability in the data. This issue is discussed further in section 3.4.4.

Scenario 5: A Pattern mixture model MNAR mechanism

Data simulation

The selection model simulated MNAR data attempts to replicate a clinical situation where a subject's response is unusually low (or high) at the time of dropout. An example of where this might occur is a study in which clinicians withdraw patients if their response falls below (or rises above) a pre-specified threshold, or a patient's condition suddenly deteriorates to the extent that they fail to attend clinic. An alternative clinical scenario can be considered where subjects that drop out of the study come from a different distribution to subjects that complete the study. This can be considered a missing covariate problem; there is a covariate that is missing from the analysis which causes the distribution of the subjects that drop out of the study to be different to the distribution of the completers. In certain clinical scenarios this may be a more realistic dropout mechanism than the selection model mechanism in scenario 4. For example, it may be that subjects that drop out of the study are characterised by a different average response and / or different reaction to the treatment to subjects that complete the study. The data were simulated from the following models for the completers and dropouts:

$$y_{ij} = \alpha_c + u_i + \beta_c treat + e_{ij} \qquad completers$$

$$y_{ij} = \alpha_d + u_i + \beta_d treat + e_{ij} \qquad dropouts$$

$$u_i \sim N(0, \sigma_u^2)$$

$$e_{ij} \sim N(0, \sigma_e^2)$$

The population parameters were defined so that:

$\beta_c = 10 \times \beta_d$

The overall population treatment effect, $\beta_{combined}$, averaged over all subjects, can be computed as the weighted mean of the two values of beta for the two distributions, weighted by the proportion of subjects with each distribution. This value was chosen to give $\beta/s.e.(\beta)$ =2.5, as it was in the previous scenarios.

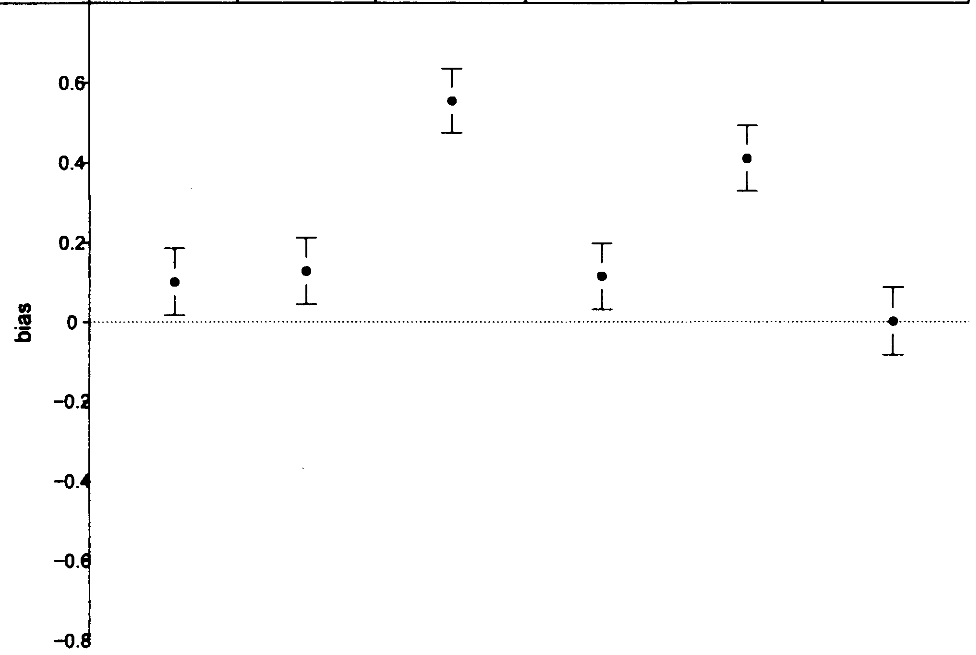
The dropout mechanism here is different to the previous scenarios, as the data are simulated in such a way that it is predetermined which subjects will drop out. Amongst the dropouts, those with the lowest response at time 2 drop out at time 2, those remaining with the lowest response at time 3 drop out at time 3, and the rest drop out at time 4.

Results

The random effects model is significantly biased with this MNAR missing data, but the bias is less than in the selection model simulated MNAR data. The GEE model is also significantly biased, even when the correct correlation structure or an unstructured correlation is used. Again, the bias in the random effects follows the same trend as the bias in the fixed effects. Notice that the models fitted here are wrong, in that there are two distribution for the data, and only one distribution is modelled. This increases the overall variability in the data, and affects the variance components and normality assumptions of the model, even in the complete data. The summary statistics method under this MNAR mechanism is unbiased. This is explained in section 3.4.3. In the straightforward unweighted summary statistics method adopted here, the number of observations within clusters is not adjusted for, and therefore, the standard error is estimated incorrectly.

Figure 3.4: Scenario 5. MNAR pattern mixture model simulated data

	Hierarchical	GEE Model				Summ. Stats.
		Exch	Indep	Unstr	A-R	
Efficiency (%)	113	114	134	113	128	104
Mean bias/s.e.	0.092	0.118	0.552	0.105	0.400	0.002
Proportion sig.	0.77	0.78	0.93	0.78	0.88	0.70



Random effects estimated in random effects model

ICC		Efficiency (%)		Bias (95% C.I.)	
Comp. data	Inc. data	σ_u^2	σ_e^2	σ_u^2	σ_e^2
0.715	0.677	97	87	-10.73 (-11.2,-10.2)	-0.253 (-0.39,-0.12)

Scenario 6: MNAR with a time-by-treatment interaction

More realistically the model for the data will have interaction terms as well as main effects. Subjects in the treatment group were given a time effect of

β_2 while subjects in the control group had no time effect, as follows:

$$y_{ij} = \alpha_c + u_i + \beta_c treat + \beta_2 time \times treat + e_{ij} \qquad \textit{completers}$$

$$y_{ij} = \alpha_d + u_i + \beta_d treat + \beta_2 time \times treat + e_{ij} \qquad \textit{dropouts}$$

$$u_i \sim N(0, \sigma_u^2)$$

$$e_{ij} \sim N(0, \sigma_e^2)$$

Again *time* is equal to j. The size of β_c and β_d are the same as in scenario 5, and the dropout mechanism is also the same. As in scenario 5, the parameter of interest is the overall treatment effect. The random effects model fitted to the data is:

$$y_{ij} = \alpha + u_i + \beta_1 treat + \beta_2 time \times treat + e_{ij} \tag{3.4}$$

while the GEE model fitted is:

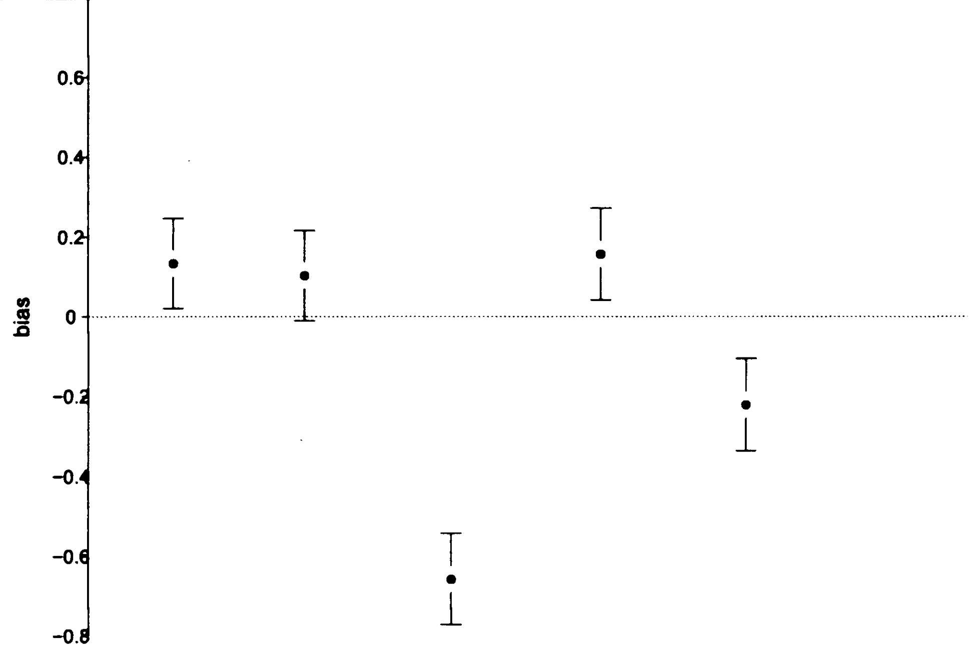
$$E(y_{ij}) = \alpha + \beta_1 treat + \beta_2 time \times treat + e_{ij} \tag{3.5}$$

Results

The overall findings are very similar to those of scenario 5, with all models biased. The performance of the GEE model was worse than the random effects model, especially when the correlation structure is misspecified. The bias is in the opposite direction to scenario 5 for the GEE independence and auto-regressive models. The other difference in the results from scenario 5

Figure 3.5: Scenario 6. Bias due to MNAR data with a period-by-treatment interaction.

	Hierar-	**GEE Model**				**Summ.**
	chical	**Exch**	**Indep**	**Unstr**	**A-R**	**Stats.**
Efficiency (%)	104	104	98	98	95	175
Mean bias/s.e.	0.095	0.073	-0.455	0.108	-0.151	1.991
Proportion sig.	0.61	0.47	0.31	0.49	0.43	1.00



Random effects estimated in random effects model

ICC		Efficiency (%)		Bias (95% C.I.)	
Comp. data	Inc. data	σ_u^2	σ_e^2	σ_u^2	σ_e^2
0.705	0.677	94	86	-7.44 (-7.9,-6.9)	-0.011 (-0.15,0.13)

is that the efficiency is generally decreased by the missing data, rather than increased.

The summary statistics method is a poor analysis approach when an interaction term is introduced, even with complete data, because it does not take into account any period effect and overestimates the difference in means of the two treatment groups.

Scenario 7: Changing the correlation within clusters

All of the data so far have had an intra-cluster correlation coefficient (ICC) of 0.5. This is the proportion of the total variance that is between-cluster variance. Lowering the ICC means that there is more variation within clusters and less variation between clusters. Removing an observation from a cluster removes a greater amount of information from the data than when the ICC is lower, because with a higher ICC the observations that remain within the same cluster are more similar to the missing observation. We therefore expect missing data to cause a greater degree of bias.

Data simulation

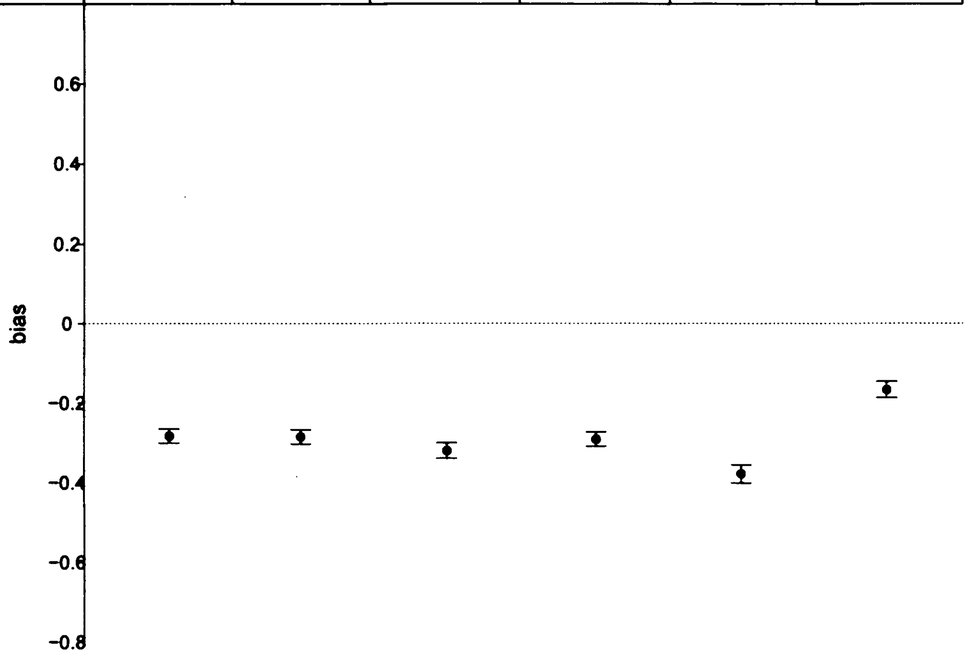
Further datasets were simulated with an ICC of 0.2. The same total variance was used but this time the between-cluster variance and within-cluster variance had the ratio 1:4. The incomplete datasets were constructed in exactly the same way as those in scenario 4.

Results

The results are summarised in figure 3.6. When the ICC is reduced, the bias in every model is much greater in comparison to the standard error of the

Figure 3.6: Scenario 7. MNAR data with a reduced ICC.

	Hierar-chical	**GEE Model**				**Summ. Stats.**
		Exch	**Indep**	**Unstr**	**A-R**	
Efficiency (%)	140	140	141	141	133	120
Mean bias/s.e.	-1.114	-1.119	-1.180	-1.132	-1.240	-0.854
Proportion sig.	0.62	0.61	0.60	0.62	0.55	0.59



Random effects estimated in random effects model

ICC		Efficiency (%)		Bias (95% C.I.)	
Comp. data	Inc. data	σ_u^2	σ_e^2	σ_u^2	σ_e^2
0.198	0.152	116	101	-3.91 (-4.1,-3.7)	-6.80 (-7.0,-6.6)

parameter. The absolute size of the bias is smaller than when the ICC is 0.5, but as in scenario 4, all models produce significantly biased estimates. The random parameters are again biased. As in scenario 4, the extent of the bias is so severe that there is little difference in bias between the different models.

The choice of correlation structure has less of an effect on the bias when the ICC is low.

3.3 Binary data

When the data are not gaussian, the random effects model and the GEE model are no longer equivalent, even when the appropriate correlation structure is chosen. Theory tells us that while the random effects model can handle missing observations that are MAR, GEE models require the stricter assumption of MCAR, as explained in section 2.6.2. It is expected that there will be bias in estimates from a GEE model fitted to binary incomplete data that do not meet the MCAR assumptions.

Estimates from a GEE model have a population-averaged interpretation while coefficients of a random effects model are cluster-specific, as described in section 2.5.4. For gaussian data, the estimates from a GEE model are the same as those estimated from a random effects model. However, for non-gaussian data the estimates are no longer equivalent. When the data are binary, population-averaged coefficients are interpreted as the average log odds ratio of the event for a unit change in the covariate, having adjusted for all other covariates. In contrast, the interpretation of the cluster-specific estimate from a random effects model is the log odds of the event for a change in the covariate, for a specific cluster with its particular values of the other covariates, averaged over all clusters. Equation 2.7 is an approximation derived by Zeger et. al. to convert a cluster-specific coefficient from a random effects logistic regression to a population-averaged coefficient.

3.3.1 Data simulation

The GEE is fitted to data simulated on the population-averaged scale, i.e. the logistic relationship between the success probability and the covariate is marginal, as in equation 2.6. Parallel datasets must be simulated on a cluster-specific scale to test the robustness of the random effects model to missing data, with a cluster-specific relationship between the success probability and the covariate, as given in equation 2.5.

Population-averaged data simulation

Population-averaged correlated binary data were simulated using an approach suggested by Oman and Zucker [35]. The data have an exchangeable correlation structure and are generated by simulating two independent standard Normal variables, ϵ_{ij} and ϵ_i, and a third independent binomial variable, U_{ij} with success probability γ. There are, again, 200 subjects with 4 observations for each subject. A continuous variable, S_{ij}, correlated within subject, is constructed as follows:

$$S_{ij} = (1 - U_{ij})\epsilon_{ij} + U_{ij}\epsilon_i \qquad (3.6)$$

From this continuous variable, a binary variable with success probability, p_i, is defined by:

$$\begin{aligned} Y_{ij} &= 1 \qquad \text{if } S_{ij} < \theta_{ij} \\ &= 0 \qquad \text{otherwise} \end{aligned} \tag{3.7}$$

where θ_{ij} is the quantile of the standard Normal distribution corresponding to the marginal success probability of the binary variable, p_i. The correlation between observations within-cluster is equal to γ^2. This correlation occurs as a result of the variable, ϵ_i, which takes the same value for all observations within the same subject.

The covariate, treatment effect, was introduced to the data by defining the value of p_i for the ith subject from the following relationship:

$$logit(p_i) = \alpha + \beta treat_i \tag{3.8}$$

Cluster-specific simulated data

Simulation of cluster-specific data is more straightforward, as follows:

$$Y_{ij} \sim Bernoulli(p_i) \tag{3.9}$$

$$p_i = \frac{exp(\alpha + u_i + \beta treat_i)}{1 + exp(\alpha + u_i + \beta treat_i)} \tag{3.10}$$

For both the population-averaged and cluster-specific simulations, the strength of β was chosen so that the size of $\frac{\beta}{se(\beta)}$ was 2.5, as in the gaussian data.

The missing data mechanism is the same as for the gaussian data, as described by equation3.2. The parameter ϕ governs the strength of the relationship between the probability of dropout and the response, and this is the same for gaussian and binary data.

3.3.2 MAR data

Scenario 8: Success probability p=0.5

Model fitting

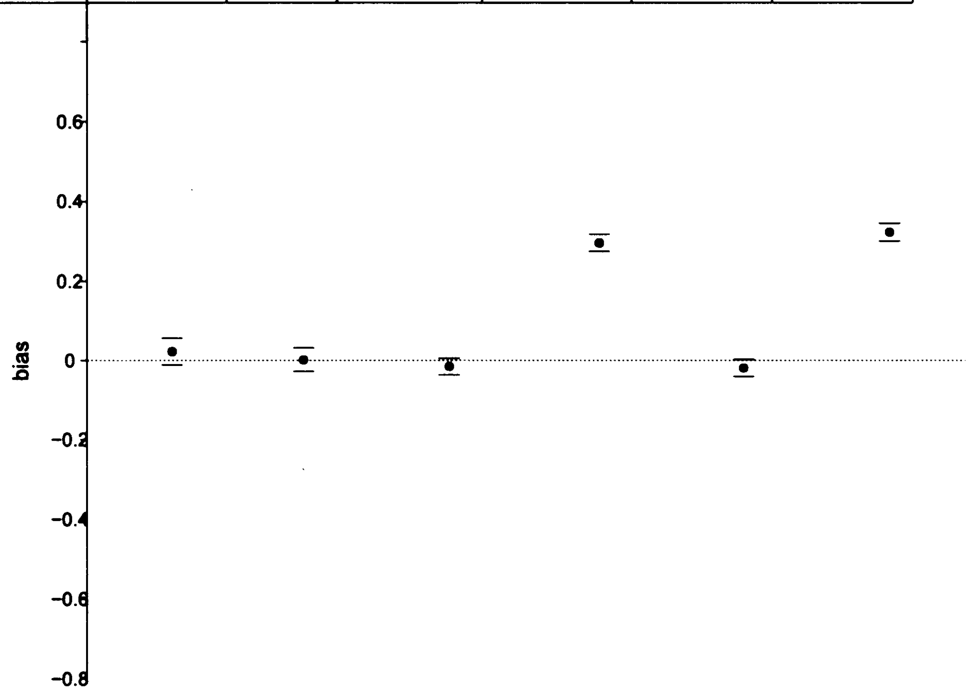
The random effects model was fitted to the binary data using both MLwiN and Stata, as the two packages use different methods of model fitting. These methods, described in section 2.5.3, are second-order penalised quasi-likelihood and quadrature respectively. There is concern that the quadrature method is unreliable for fitting random effects models to non-gaussian data because convergence can be difficult to achieve, especially if the number of quadrature points is not large [7]. Stata uses 12 quadrature points by default but this number can be increased at the expense of longer convergence time. In this study the number of quadrature points was kept at the default setting of 12.

Results

The random effects models implemented in Stata and MLwiN are not significantly biased in the fixed effects, although the bias in the random effects is significant. The GEE model is unbiased when the correlation structure is correctly specified or an unstructured correlation matrix is used. The GEE

Figure 3.7: Scenario 8. Binary MAR data, success probability p=0.5

	R.E. model		GEE Model			
	MLwiN	**Stata**	**Exch**	**Indep**	**Unstr**	**A-R**
Efficiency (%)	117	94	85	76	84	70
Mean bias/s.e.	0.063	0.006	-0.059	1.111	-0.074	1.159
Proportion sig.	0.64	0.70	0.72	0.24	0.73	0.22



Random effects estimated in random effects model

MLwiN		Stata	
Efficiency σ_u^2 (%)	Bias σ_u^2 (95% C.I.)	Efficiency σ_u^2 (%)	Bias σ_u^2 (95% C.I.)
83	-0.145 (-0.20,-0.09)	76	0.142 (-0.01,0.30)

model with incorrect correlation structure produces the most bias and least efficient estimates.

Note that there is no estimate of the within-cluster variance, σ_e^2, for binary data. In a random effects logistic regression model it is usual to constrain the level 1 random effects, e_{ij}, to have mean zero and variance 1, resulting in binomial variation.

Scenario 9: Reducing the sample size

As with gaussian data, the sample size did not affect the findings of the study. Again, the random effects model in MLwiN and the GEE model with correctly specified correlation were unbiased, but the other models were significantly biased.

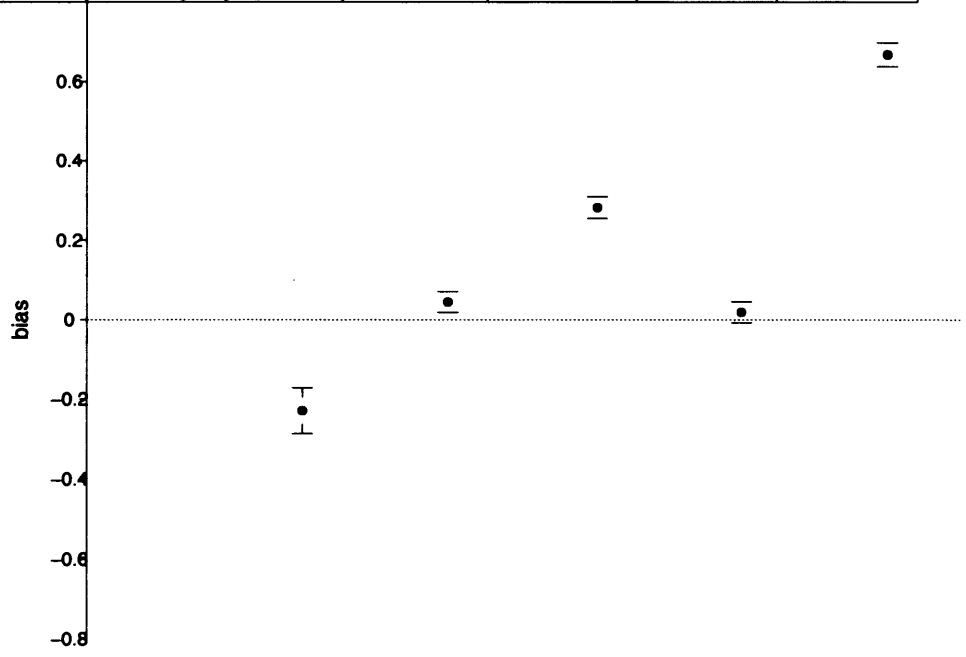
Scenario 10: Changing the success probability, p

The results in figure 3.8 are for binary data with success probability p=0.2 instead of the previous p=0.5. The missing data mechanism is MAR, as in scenario 8.

When the success probability of the binary response variable is reduced from 0.5 to 0.2 the bias due to dropout in the GEE models increases. With this smaller success probability, the bias is statistically significant, even when the correct correlation structure is applied. GEE models with incorrectly specified correlation structure are extremely biased. The random effects model fitted in Stata is also significantly and substantially biased, for both the fixed and random effects. There are problems with model convergence in MLwiN, which is why no results are shown for this model. In almost 40% of simula-

Figure 3.8: Scenario 10. Binary MAR data with success probability 0.2

	R.E. model		**GEE Model**			
	MLwiN	**Stata**	**Exch**	**Indep**	**Unstr**	**A-R**
Efficiency (%)	-	86	98	83	94	58
Mean bias/s.e.	-	-0.555	0.148	0.865	0.062	1.718
Proportion sig.	-	0.65	0.67	0.35	0.71	0.10



Random effects estimated in random effects model

MLwiN		Stata	
Efficiency σ_u^2 (%)	Bias σ_u^2 (95% C.I.)	Efficiency σ_u^2 (%)	Bias σ_u^2 (95% C.I.)
		67	0.541 (0.37,0.71)

tions, either the model fails to converge, or the model becomes un-estimable because the covariance matrix of the random effects becomes negative definite or all the random parameters become zero during the iterations.

3.3.3 Missing observations that are missing not at random

Scenario 11

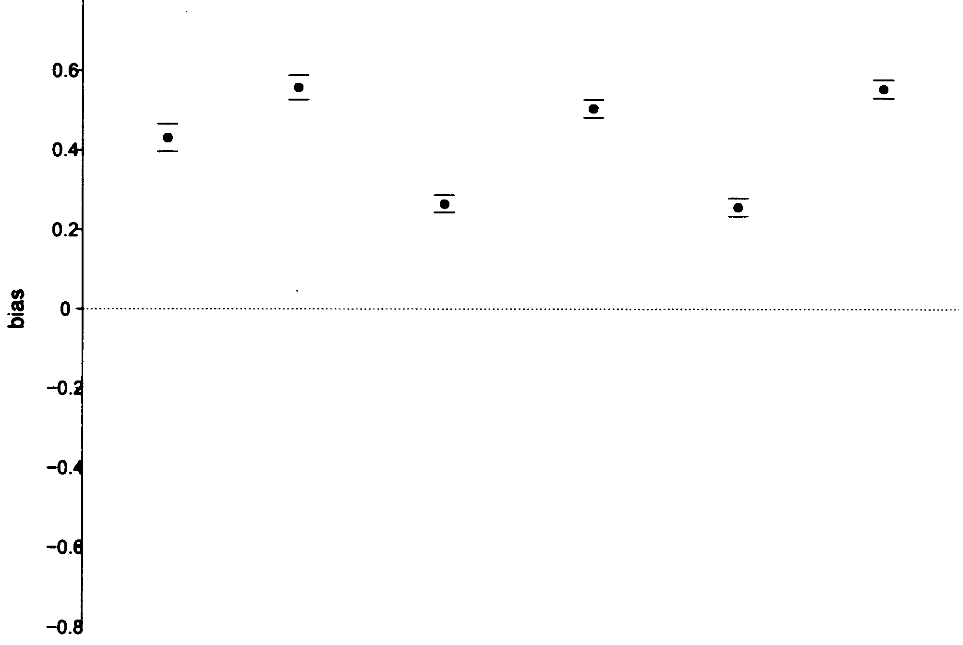
Refer to figure 3.9 for a summary of the results. All models perform very poorly when the missing data are MNAR. For example, the estimates obtained from a random effects model fitted in Stata and the GEE exchangeable model are both biased on average by an amount comparable to the size of their standard error. And worse than this, a GEE model with auto-regressive correlation structure produces a mean estimate biased by almost twice its standard error. The size of the bias in the random effects model is greater here than for MNAR gaussian data, but is no greater in the GEE models. The extent of the bias is so great that no model could be recommended when the data are binary and observations are MNAR.

3.4 Discussion of findings

The findings are summarised in tables 3.2 and 3.3 for gaussian and binary response data respectively. In each scenario and for each model, a • in the column "Biased" indicates that with 20% of observations missing, the bias

Figure 3.9: Scenario 11. Binary MNAR data, success probability 0.5.

	R.E. model		**GEE Model**			
	MLwiN	**Stata**	**Exch**	**Indep**	**Unstr**	**A-R**
Efficiency (%)	127	84	77	71	75	64
Mean bias/s.e.	1.175	1.515	1.001	1.831	0.958	1.908
Proportion sig.	0.24	0.14	0.29	0.09	0.30	0.06



Random effects estimated in random effects model

MLwiN		Stata	
Efficiency σ_u^2 (%)	Bias σ_u^2 (95% C.I.)	Efficiency σ_u^2 (%)	Bias σ_u^2 (95% C.I.)
83	-0.012 (-0.07,0.05)	105	-0.056 (-0.28,0.16)

was at least 0.1 times the size of the standard error, and was statistically significant. And efficiency of less than 80% is indicated by a • in the "Ineffic." column.

Table 3.2: Summary of findings for gaussian response data

	Scenario	R.E.		Exch.		Ind.		Unstr.		A-R		Sstat.	
		Biased	Ineffic.	Biased	Ineffic.	Biased	Ineffic.	Biased	Ineffic.	Biased	Ineffic.	Biased	Ineffic.
	MAR												
1	Cluster-level					•				•		•	
3	Cluster-varying		•	•	•	•	•	•	•	•	•	•	•
	MNAR												
4	SM	•		•		•		•		•		•	
5	PMM			•		•		•		•			
6	period×treat					•		•		•		•	
7	ICC=0.2	•		•		•		•		•		•	

Table 3.3: Summary of findings for binary response data

	Scenario	R.E.				Exch.		Ind.		Unstr.		A-R	
		MLwiN		Stata									
		Biased	Ineffic.	Biased	Ineffic.	Biased	Ineffic.	Biased	Ineffic.	Biased	Ineffic.	Biased	Ineffic.
	MAR												
8	p=0.5							•	•			•	•
9	p=0.2	–	–	•		•		•				•	•
	MNAR												
11	p=0.5	•		•		•	•	•	•	•	•	•	•

Studies of the degree of missingness show that in the random effects model or GEE with correctly specified correlation, the bias is unsubstantial with up to

20% missing observations when missingness is MAR. With a MNAR dropout mechanism, however, the parameter estimates are affected substantially with as little as 5% of observations missing. This demonstrates that the correlation within clusters is no protection against MNAR missing data, and that any assumption that the missing data are MAR, and not MNAR, should not be made without serious consideration.

Reducing the ICC does increase the bias of the parameter estimate relative to its standard error, as expected. The absolute size of the bias, however, is smaller when the ICC is lower. This follows from noting that data of the same sample size, and the same total variance (i.e. the sum of the within- and between-cluster variances) but lower ICC produce parameter estimates with smaller standard errors, at least when estimating the effect of a cluster-constant covariate. The missing data, therefore, cause a smaller absolute bias in the mean response in each treatment arm, and therefore a smaller bias in the treatment effect. This result cannot be generalised to cluster-varying covariates because the ICC will affect the estimates of a cluster-varying covariate differently; large correlation within clusters is likely to lead to better estimation of effects within clusters.

3.4.1 Random effects model

The random effects model was found to be robust to MAR missing data for both cluster-level and cluster-varying covariates. Although the bias in the presence of a cluster-varying covariate was statistically significant, the size of the bias compared to the standard error of the parameter was minimal. The only exception to this robustness to MAR is in the logistic regression model when p is small, when the model fitted in MLwiN often failed to converge, and the quadrature method implemented in Stata 8 was substantially biased. These findings agree in general with the theory that likelihood estimation is robust to MAR data, as discussed in section 2.6.2. The findings were the same for both fixed and random effects. As expected, the more severe missing data mechanism, MNAR, leads to significant bias in estimates from the random effects model, even with as little as 5% of observations missing. The bias in the random effects model, compared to the standard error of the estimate, as in all approaches, increases as the ICC decreases. This is because each observation carries more information when the correlation within clusters is reduced, therefore removing observations causes more bias. As expected, the extent of the bias also increases with the proportion of observations missing. Reducing the number of clusters in the datasets did not affect the findings of the study.

Although the quadrature method, implemented in Stata 8 for the logistic random effects model, was robust to MAR for p=0.5, the estimates were significantly biased when the average success probability was reduced to 0.2. The adequacy of the method of model fitting was checked using a feature in Stata 8 that re-fits the model with two different numbers of quadrature points, and compares the parameter estimates between the two model fits. If the parameter estimates from each model fit have a relative difference of greater than 1×10^{-4} the quadrature method is not considered reliable [36].

For both the complete datasets and the incomplete datasets in scenarios 8 and 10, the relative difference in the model parameters with 8 and 16 quadrature points was of the order of 0.1. This demonstrates that the quadrature method did not fit the model reliably, and explains why the random effects model fitted in Stata gave biased results, in scenario 10, with p=0.2. Up to 30 quadrature points can be used to fit the logistic random effects model in Stata 8. With 30 quadrature points, the relative difference in the parameter estimates with 26 and 34 quadrature points is of the order 10^{-10}, demonstrating a much more reliable model fit. In fact, with the maximum number of quadrature points, MAR data do not cause significant bias in scenario 10.

Stata have released version 9 since these simulation studies were carried out. Stata 9 uses adaptive Gauss-Hermite quadrature in its estimation of random effects logistic regression models, using the procedure *glamm*. This transforms the integrand so that it is sampled on a more appropriate range, and has been shown to be a better numerical integration method than non-adaptive quadrature [37]. The *glamm* procedure also extends to random coefficients other than a random intercept, which the standard random effects model is Stata does not allow.

There were convergence problems with MLwiN when the success probability was small. Goldstein [7] found by simulation that when the average probability is very small (or very close to 1), and there are therefore many clusters where all elements are zero (or one), the estimates may not converge, and when they do, they may be biased.

3.4.2 GEE models

GEE models on gaussian data are expected to be unbiased under MAR when the correlation structure is correctly specified and the correlation coefficient is estimated without bias. Here the GEE model was found to be fairly robust to MAR observations if the correct correlation structure was applied, or an unstructured correlation matrix was used. Under these conditions the bias was small but statistically significant. The estimate was more biased but just as efficient as the estimate from the random effects model. For binary response data, the GEE model is fairly robust to MAR observations only if the correct correlation structure is applied and the average event probability, p, is 0.5. If the wrong working correlation matrix is applied and / or the event probability is small, the GEE model is significantly biased but the size of the bias compared to the standard error of the estimate is small, and is smaller than that estimated from the random effects model implemented in Stata 8.

If the missing data are MNAR, GEE perform poorly, for both gaussian and binary data, for cluster-level and cluster-varying covariates. This is also the case for the random effects model; all models are very biased in the presence of as little as 5% of observations MNAR.

MAR dropout in GEE with gaussian response

The following description of GEE demonstrates why GEE are, in theory, asymptotically robust to MAR dropout for gaussian data, when the correct correlation structure is chosen. Following this proof is an explanation of why GEE with the wrong working correlation matrix are biased under MAR dropout.

The score-like equations solved in GEE models, for N clusters, are [4]:

$$\sum_{i=1}^{N} \mathbf{U_i} = \sum_{i=1}^{N} \mathbf{D_i^T V_i^{-1} S_i} = \mathbf{0} \tag{3.11}$$

where

$$\mathbf{V_i} = \frac{\mathbf{A}_i^{1/2}\mathbf{R}_i(\alpha)\mathbf{A}_i^{1/2}}{\phi}$$

In this representation of the score equations $\mathbf{A}_i$ is a diagonal variance matrix of $\mathbf{Y}_i$, $\mathbf{R}_i(\alpha)$ is the working correlation matrix, and $\mathbf{V}_i$ is the working covariance matrix of $\mathbf{Y}_i$. $\mathbf{S}_i$ is a vector of residuals for the ith cluster, $(\mathbf{Y}_i - \boldsymbol{\mu}_i)$. $\mathbf{D}_i = \frac{\partial \mu_i}{\partial \beta}$ can be considered the "covariate weighting" because it weights how useful the elements of $\mathbf{Y}_i$ are for estimating β from the value of $\mathbf{X}_i$.

Consider the situation where the full response vector, $\mathbf{Y}$, is partitioned into $\mathbf{Y_1}$ of dimension n_1, which is fully observed for all subjects, and $\mathbf{Y_2}$ of dimension n_2, which is observed only for some subjects. If $\boldsymbol{\mu} = E(\mathbf{Y}_{1i} \mid \mathbf{X}_i)$ then $\mathbf{S}_{1i}$ is a vector of residuals $\mathbf{Y}_{1i} - \boldsymbol{\mu}_{1i}$ and analogously, $\mathbf{S}_{2i} = \mathbf{Y}_{2i} - \boldsymbol{\mu}_{2i}$. Let $\mathbf{V}_{1i}$ be the working covariance matrix in the fully observed partition of the data, and let $\mathbf{D}_{1i} = \frac{\partial \mu_{1i}}{\partial \beta}$. Then the contribution to the estimating equations from a subject that is only partially observed is:

$$\mathbf{U}_i^O = \mathbf{D}_{1i}' \mathbf{V}_{1i}^{-1} \mathbf{S}_{1i} \tag{3.12}$$

Define $\mathbf{M}_i$ to be an indicator variable, equal to one if a subject is partially observed, and zero for completely observed subjects. The full GEE, had all observation been obtained is then:

$$\sum_{i=1}^{N} \mathbf{U}_i = \sum_{i=1}^{N} [\mathbf{M}_i \mathbf{U}_i + (1 - \mathbf{M}_i)\mathbf{U}_i] \tag{3.13}$$

while the estimating equations for the incomplete data, ignoring the missing data are:

$$\sum_{i=1}^{N} \mathbf{U}_i = \sum_{i=1}^{N} [\mathbf{M}_i \mathbf{U}_i^O + (1 - \mathbf{M}_i)U_i] \tag{3.14}$$

so that the difference in the estimating equations on the complete compared to the incomplete data lies in whether the contribution from the subjects that drop out is $\mathbf{U}_i$ or $\mathbf{U}_i^O$.

Assume that the working correlation matrix, $R(\alpha)$, is correctly specified. For normal linear regression $\mathbf{A}_i$ is constant across j and can be re-written:

$$\mathbf{A}_i = \sigma^2 \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & \dots \\ \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & 1 \end{pmatrix} \tag{3.15}$$

and then:

$$\begin{aligned} E(\mathbf{U}_i \mid \mathbf{M}_i = 1) &= E(\mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{S}_i \mid \mathbf{M}_i = 1) \qquad (3.16) \\ &= \frac{\sigma^2}{\phi} \mathbf{D}_i^T \begin{pmatrix} Corr(\mathbf{Y}_1) & Corr^T(\mathbf{Y}_1, \mathbf{Y}_2) \\ Corr(\mathbf{Y}_1, \mathbf{Y}_2) & Corr(\mathbf{Y}_2) \end{pmatrix}^{-1} E \begin{bmatrix} \mathbf{S}_1 \mid \mathbf{M}_i = 1 \\ \mathbf{S}_2 \mid \mathbf{M}_i = 1 \end{bmatrix} \end{aligned}$$

and from multivariate theory we note that the expectation of the unobserved $\mathbf{S}_2$ conditional on the observed $\mathbf{Y}_1$, assuming the missing data are MAR, is:

$$E(\mathbf{S}_2 \mid \mathbf{M}_i = 1) = Corr(\mathbf{Y}_1, \mathbf{Y}_2) Corr(\mathbf{Y}_1)^{-1} E(\mathbf{S}_1 \mid \mathbf{M}_i = 1) \qquad (3.17)$$

so that, from 3.16:

$$E(\mathbf{U}_i \mid \mathbf{M}_i = 1) = \qquad (3.18)$$

$$\frac{\sigma^2}{\phi}\mathbf{D}_i' \begin{pmatrix} Corr(\mathbf{Y}_1) & Corr^T(\mathbf{Y}_1,\mathbf{Y}_2) \\ Corr(\mathbf{Y}_1,\mathbf{Y}_2) & Corr(\mathbf{Y}_2) \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{I}_{n_1} \\ Corr(\mathbf{Y}_1,\mathbf{Y}_2) Corr(\mathbf{Y}_1)^{-1} \end{pmatrix} E[\mathbf{S}_1 \mid \mathbf{M}_i = 1]$$

which can be shown to be equal to:

$$E(\mathbf{U}_i \mid \mathbf{M}_i = 1) = \frac{\sigma^2}{\phi} D_i' \begin{pmatrix} Corr^{-1}(\mathbf{Y}_1) \\ \mathbf{0}_{n_2,n_1} \end{pmatrix} E[\mathbf{S}_1 \mid \mathbf{M}_i = 1] \qquad (3.19)$$

where $0_{n_2,n_1}$ is a matrix of zeros of dimension $n_2 \times n_1$. The proof of 3.19 is given in appendix A. Compare this to the expectation of the ith contribution to the GEE of the observed measurements, ignoring the missing data:

$$\begin{aligned} E(\mathbf{U}_i^O \mid \mathbf{M}_i = 1) &= E(\mathbf{D}_{1i}'\mathbf{V}_{1i}^{-1}\mathbf{S}_{1i} \mid \mathbf{M}_i = 1) \qquad (3.20) \\ &= \frac{\sigma^2}{\phi}\mathbf{D}_{1i}' Corr^{-1}(\mathbf{Y}_1) E[\mathbf{S}_1 \mid \mathbf{M}_i = 1] \end{aligned}$$

so that $E(\mathbf{U}_i^O \mid \mathbf{M}_i = 1) = E(\mathbf{U}_i \mid \mathbf{M}_i = 1)$ and the GEE on the incomplete data is asymptotically unbiased if the missing data are MAR and the correct correlation structure is chosen.

Intrinsic weighting in GEE

Equation 3.19 demonstrates that GEE are robust to MAR missing data if the correct correlation structure is chosen. The intrinsic weighting in GEE explains the bias in GEE with different working correlation matrices.

To ease notation, $\mathbf{V}_i^{-1}$ is represented by $\mathbf{C}_i$, which is referred to as the "inverse covariance weighting", and the score equations in 3.11 become:

$$\sum_{i=1}^{N} \mathbf{U_i} = \mathbf{D}_i^T \mathbf{C}_i \mathbf{S}_i = 0 \tag{3.21}$$

If the covariate is fixed at the cluster level, the elements of $\mathbf{D}_i$ are constant within the cluster and $\mathbf{D}_i$ becomes:

$$\mathbf{D}_i = d_i \begin{pmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{pmatrix} \tag{3.22}$$

and the score equations are reduced to:

$$\begin{aligned} \mathbf{U}_i \quad &= d_i(1,1,\ldots,1) \begin{pmatrix} C_{i11} & C_{i12} & \ldots & C_{i1n_i} \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ C_{in_i1} & \ldots & \ldots & C_{in_in_i} \end{pmatrix} \begin{pmatrix} S_{i1} \\ S_{i2} \\ \cdot \\ \cdot \\ \cdot \\ S_{in_i} \end{pmatrix} && (3.23) \\ &= d_i(\textstyle\sum_{k=1}^{n_i} C_{ik1}S_{i1} + \sum_{k=1}^{n_i} C_{ik2}S_{i2} + \ldots + \sum_{k=1}^{n_i} C_{ikn_i}S_{in_i}) && (3.24) \\ &= d_i(W_{i1}S_{i1} + \ldots + W_{ij}S_{ij} + \ldots + W_{in_i}S_{in_i}) && (3.25) \end{aligned}$$

where C_{kj} is the element of the inverse covariance matrix corresponding to the jth and kth elements of $\mathbf{Y}_i$. Equation 3.25 demonstrates that the weight from the covariance matrix for the jth residual, W_{ij}, is the sum of the jth column of the inverse covariance weighting matrix, $\mathbf{C}_i$. The vector $\mathbf{W_i}$ is referred to here as the "correlation weighting".

As stated earlier, for gaussian data, $\mathbf{A_i}$ is constant across j and can be rewritten:

$$\mathbf{A}_i = \sigma^2 \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & \dots \\ \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & 1 \end{pmatrix} \tag{3.26}$$

Therefore, for gaussian data with only a cluster-level covariate, $\mathbf{C}_i$ is a scalar multiple of the inverse working correlation matrix. For exchangeable, independence and autoregressive GEE models, the working correlation matrix is a function of the correlation parameter, $\hat{\alpha}$, estimated at each iteration of the GEE algorithm. The expected correlation weighting, $\mathbf{W_i}$ is calculated for exchangeable, independence and autoregressive GEE models, as a function of $\hat{\alpha}$, in table 3.4.

Here, the simulated data are exchangeable, and equation 3.19 demonstrates that if exchangeable GEE are applied to these data with MAR dropout, and the correlation coefficient is estimated without bias, the regression estimates will be asymptotically unbiased. In scenario 1, the GEE parameter estimates were much more biased when the wrong correlation structure was chosen.

Equation 3.25 expresses the GEE as a weighted sum of the residuals. If these weights do not match those in the exchangeable GEE, and there is bias in

Table 3.4: Intrinsic weighting within and between clusters for the GEE model and summary statistics method

Correlation structure	**Correlation weighting,** W_{ij}	**Weighting for cluster,** $W_{i.}$
Independence	$\begin{pmatrix} w_{i1} \\ w_{i2} \\ w_{i3} \\ \cdot \\ \cdot \\ \cdot \\ w_{in_i} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{pmatrix}$	n_i
Exchangeable	$\begin{pmatrix} w_{i1} \\ w_{i2} \\ w_{i3} \\ \cdot \\ \cdot \\ \cdot \\ w_{in_i} \end{pmatrix} = \frac{1}{1+(n_i-1)\hat{\alpha}} \begin{pmatrix} 1 \\ 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{pmatrix}$	$\frac{n_i}{1+(n_i-1)\hat{\alpha}}$
Auto-regressive	$\begin{pmatrix} w_{i1} \\ w_{i2} \\ w_{i3} \\ \cdot \\ \cdot \\ \cdot \\ w_{in_i} \end{pmatrix} = \frac{1}{1+\hat{\alpha}} \begin{pmatrix} 1 \\ 1-\hat{\alpha} \\ 1-\hat{\alpha} \\ \cdot \\ \cdot \\ 1-\hat{\alpha} \\ 1 \end{pmatrix}$	$\frac{n_i-(n_i-2)\hat{\alpha}}{1+\hat{\alpha}}$
Summary Statistics	$\begin{pmatrix} w_{i1} \\ w_{i2} \\ w_{i3} \\ \cdot \\ \cdot \\ \cdot \\ w_{in_i} \end{pmatrix} = \frac{1}{n_i} \begin{pmatrix} 1 \\ 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{pmatrix}$	1

elements within clusters and/or in whole clusters, the parameter estimates will be biased. Table 3.4 demonstrates that the weights between elements and/ or between clusters in the independence and autoregressive GEE are different to those in the exchangeable GEE. The MAR dropout mechanism causes subjects to drop out immediately after an unusual response, causing bias in the elements of incomplete cluster as a whole. The correlation between elements in a cluster means that there is also bias in the clusters. This explains the different bias in the GEE when different working correlation structures are chosen.

The unstructured GEE with MAR data gives similar bias to the exchangeable GEE, because the correlation structure is estimated from the data. When the data are exchangeable, it is expected that the unstructured correlation matrix will be estimated as similar to the exchangeable correlation matrix. By examining the working correlation matrix of the unstructured GEE model, it was established that this was indeed the case.

Estimation of the correlation coefficient, α

Park [2] demonstrated that generalised estimating equations do not always reduce to the score equations when the data are incomplete, even if the missing data are MCAR. He attributes this to bias in the estimation of the covariance matrix when the cluster sizes are unequal.

Since Liang and Zeger first published their GEE approach, advances have been made in estimating parameters of the correlation matrix in the presence of missing data. The standard GEE use the unconditional Pearson residuals to estimate α. One alternative approach proposed by Carey et. al. [38] and Lipsitz and Fitzmaurice [39] is to estimate α using conditional residuals, i.e. $\{Y_{is}Y_{it} - E(Y_{it} \mid Y_{is} = y_{is}, X_i)\}$ in estimating the correlation between observations s and t $(t > s)$ in the ith cluster, as opposed to $\{Y_{is}Y_{it} - E(Y_{it}Y_{is} \mid X_i)\}$ in the standard GEE. An alternative approach for estimating the correlation parameters is through multivariate normal estimating equations. Lipsitz et. al. [40] suggested this approach for binary outcomes, which ensures that the correlation matrix is non-negative definite. Fitzmaurice et. al. [34] demonstrate through simulation studies that, with binary response data, these advances in estimating the correlation parameters result in negligible bias in α, and as a result, in the regression parameters, when missing data are MAR. They found that the standard GEE of Liang and Zeger, fitted to the same data, gave biased estimates of α and $\boldsymbol{\beta}$.

Perhaps then, a further possible explanation of the sensitivity of GEE models to missing data is that there is bias in the estimation of α. The simulation study mentioned above, by Fitzmaurice et. al., investigated the bias in the correlation parameter for binary response data only. The GEE with correctly specified correlation structure is biased in the presence of MAR dropout, even with gaussian response data. This disagrees with the result in equation 3.19. Also, the GEE reduce to the score equations under these conditions, and the random effects model is unbiased. Bias in the estimation of α would explain this anomaly.

Note that the data are simulated to have an exchangeable correlation but that the autoregressive GEE model fits a correlation of $\gamma^{|j-k|}$ between observations at time-points j and k. For example, the autoregressive model attempts to fit the following correlation structure to a cluster of size 4:

$$\begin{pmatrix} 1 & \gamma & \gamma^2 & \gamma^3 \\ \gamma & 1 & \gamma & \gamma^2 \\ \gamma^2 & \gamma & 1 & \gamma \\ \gamma^3 & \gamma^2 & \gamma & 1 \end{pmatrix} \tag{3.27}$$

when the true correlation structure is:

$$\begin{pmatrix} 1 & \alpha & \alpha & \alpha \\ \alpha & 1 & \alpha & \alpha \\ \alpha & \alpha & 1 & \alpha \\ \alpha & \alpha & \alpha & 1 \end{pmatrix} \tag{3.28}$$

Depending on how the particular algorithm estimates γ for autoregressive correlation, $\hat{\gamma}$ could be very biased if the true correlation is exchangeable. However, Stata's autoregressive GEE model estimates γ using only pairs of observations that are one time-period apart. Therefore in Stata, $\hat{\gamma}$ estimated from an a-r GEE model will be unbiased (if less efficient) compared to $\hat{\alpha}$ estimated on the same data from the exchangeable GEE, when the true correlation structure is exchangeable.

The estimation of the correlation coefficient was checked by examining the estimates across all 1000 datasets. By correlation coefficient, we refer to α in the exchangeable GEE, γ in the auto-regressive GEE, and all elements, α_{jk} of the unstructured correlation matrix. For example, in the unstructured GEE, a cluster of size 4 has a vector of 6 correlation parameters, defined by:

$$\begin{pmatrix} 1 & \alpha_{21} & \alpha_{31} & \alpha_{41} \\ \alpha_{21} & 1 & \alpha_{32} & \alpha_{42} \\ \alpha_{31} & \alpha_{32} & 1 & \alpha_{43} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & 1 \end{pmatrix} \qquad (3.29)$$

For gaussian response data, with cluster-varying or cluster-level covariates, the GEE model with exchangeable, unstructured or a-r correlation estimated this parameter without bias when the data were complete. The estimate was significantly biased by the missing data, with very little difference in the estimates between exchangeable, unstructured and autoregressive correlation structures. Whether the covariate was fixed at the cluster level or was cluster-varying made very little difference either. 95% confidence intervals for the correlation coefficient are given in table 3.5. This bias explains the bias in the GEE with gaussian data and a MAR dropout mechanism, even when the correlation structure is correctly specified.

GEE models with binary data

The severity of the bias in GEE models is similar for gaussian and binary data for MAR and MNAR data, especially if the success probability in the binary data is not 0.5. In general, GEE models require missing observations to be MCAR. As explained by equation 3.19, for gaussian response data, there are circumstances where the generalised estimating equations are robust to MAR. GEE are not expected to be robust to MAR missing data for binary data, and there is indeed bias in the GEE logistic model, even when the correct

Table 3.5: 95% confidence intervals for the correlation coefficient, estimated in each GEE, for the complete and incomplete gaussian data with α=0.5

Complete/ Incomplete	Correlation structure	95% CI
Cluster-level covariate		
Complete	Exchangeable	(0.42,0.57)
	Unstructured	(0.42,0.57)
	A-R	(0.42,0.57)
Incomplete	Exchangeable	(0.32,0.51)
	Unstructured	(0.27,0.50)
	A-R	(0.29,0.48)
Cluster-varying covariate		
Complete	Exchangeable	(0.42,0.57)
	Unstructured	(0.37,0.61)
	A-R	(0.42,0.57)
Incomplete	Exchangeable	(0.31,0.51)
	Unstructured	(0.28,0.51)
	A-R	(0.29,0.48)

correlation structure is chosen. This is the case even when the estimation of α is unbiased, when the success probability, p, is 0.5. When p=0.2, the estimation of α becomes biased, as shown in table 3.6, and the bias in the GEE increases for all correlation structures. The bias in α is a possible explanation for this increase in bias in the regression parameters.

3.4.3 Summary statistics method

Cluster-level covariates

For MAR data and cluster-level covariates, the size of the bias is similar in size to that of the GEE model with mis-specified correlation structure. The exchangeable GEE has been shown to be robust to MAR dropout. Just

Table 3.6: 95% confidence intervals for the correlation coefficient, estimated in each GEE, for the complete and incomplete binary data with α=0.5

Complete/ Incomplete	**Correlation structure**	**95% CI**
Success probability p=0.5		
Complete	Exchangeable	(0.42,0.57)
	Unstructured	(0.37,0.61)
	A-R	(0.41,0.58)
Incomplete	Exchangeable	(0.43,0.61)
	Unstructured	(0.36,0.63)
	A-R	(0.39,0.59)
Success probability p=0.2		
Complete	Exchangeable	(0.39,0.60)
	Unstructured	(0.33,0.66)
	A-R	(0.38,0.61)
Incomplete	Exchangeable	(0.20,0.50)
	Unstructured	(0.21,0.60)
	A-R	(0.17,0.50)

as the bias in the independence and autoregressive GEE was explained by examining the weights within and between clusters, the same approach can be used with the summary statistics method. Table 3.4 demonstrates that the weights of the summary statistics method do not match those of the exchangeable GEE, explaining the bias in the summary statistics method.

Cluster-varying covariates

With cluster-varying covariates, the summary statistics method is the most biased of all approaches. The bias is positive in contrast to that for the GEE and random effects models. The summary statistics method implemented here computes the time effect for each subject separately by estimating the slope over time of the responses, and then takes the mean of all of these estimates. The formula for the ordinary least squares estimate of the slope is:

$$\hat{\beta}_{time} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{3.30}$$

When all four responses are observed, $\mathbf{x}^{\mathbf{T}} = (1, 2, 3, 4)$ with a resulting $\overline{x}$ of 2.5 and:

$$\hat{\beta}_{time} = \frac{1}{5}(-1.5y_1 - 0.5y_2 + 0.5y_3 + 1.5y_4) \tag{3.31}$$

If y_4 is missing then $\hat{\beta}_{time}$ becomes:

$$\hat{\beta}_{time} = \frac{1}{2}(y_3 - y_1) \tag{3.32}$$

with the value of y_2 having no effect on the estimate of the slope. This occurs here because the observations are at equally spaced timepoints. If y_4 is missing then $\bar{x} = 2$. The result of this is that when $i = 2$ in equation 3.30, $(x_i - \bar{x})$ becomes zero, and there is no contribution from y_2 to the estimate of the slope.

If both y_3 and y_4 are missing then $\hat{\beta}_{time}$ becomes:

$$\hat{\beta}_{time} = y_2 - y_1 \tag{3.33}$$

If subjects have only one observation, no estimate of the slope is available. Subjects that drop out after period 3 have a slope estimated from the first and third observations only. With MAR data, the third observation is unusually low for that subject, and is on average lower than the first observation, so the estimate of the slope becomes negatively biased. The same applies to subjects that drop out after period 2. Their second observation is unusually low and their slope is estimated to be the difference between their second and first observations. Again, the estimate of the slope becomes negatively biased.

The bias in the summary statistics method is much greater for the cluster-varying covariate than for the cluster-level covariate. Consider a simple scenario with no time effect where the response of a particular subject that drops out after period 3 varies randomly about $\bar{y}$. There is a MAR missing data mechanism and the response at period 3 is biased by an amount $-\delta$, leading to the subject dropping out. The subject has expected response $y^T = (\bar{y}, \bar{y}, \bar{y} - \delta, \bar{y})$. The expected estimate of the slope for that subject's incomplete cluster is $E(\frac{1}{2}(y_3 - y_1)) = -\frac{\delta}{2}$. The expected slope for the complete cluster is $-\frac{\delta}{10}$, so the expected bias in the slope for that subject is $-\frac{4\delta}{10}$. This compares to a bias of $(\bar{y} - \frac{\delta}{4}) - (\bar{y} - \frac{\delta}{3}) = \frac{\delta}{12}$ in the estimate of the mean for that subject. Following the same argument, consider a cluster of size 2, in which the second observation is biased by an amount $-\delta$. The expected bias in the period effect for this cluster is $-\frac{11\delta}{10}$ compared to an expected bias of $-\frac{\delta}{4}$ in the mean of the cluster. This demonstrates that the bias in the slope is greater for the summary statistics method than the bias in the treatment effect.

MNAR data

In the fairly limited scenarios here, the summary statistics method performs better than the more sophisticated random effects and GEE models, with MNAR data. When the data were simulated under a pattern mixture model framework the summary statistics method was unbiased. This is because the observations remaining in an incomplete cluster were unbiased estimators of the mean of the complete data for that cluster, and clusters are weighted equally, regardless of their size. The incomplete clusters have a different mean to the complete clusters, but all clusters are given the same weights as they would have been had the data been completely observed.

It should be noted that although the summary statistics method did perform well under these limited scenarios, it is a fairly inflexible tool that cannot cope with several covariates, any interaction terms, more complicated correlation structures etc.. The summary statistics approach is very sensitive to missing data when the covariate of interest is cluster-varying, as demonstrated in figure 3.2. Also, even if the approach does provide an unbiased estimator of the parameter of interest, the unweighted summary statistics method, used here, does not adjust for the number of observations per cluster, and will therefore give a biased estimate of the standard error.

3.4.4 Model efficiency in incomplete data

For gaussian data, the loss of extreme observations in MNAR dropout leads to a reduction in variability, to such an extent that even though there is a loss of information, the efficiency of the estimates increases. This effect is seen in all models, in all MNAR scenarios. The negative bias in the within- and between-cluster variances demonstrates this reduction in variability, caused by MNAR dropout. The effect is most noticeable in scenarios 5 and 6, when the models fitted to the data are wrong, in that the completers and dropouts have different distributions. In a scenario where the data come from a mixture of different normal distributions, it is easy to see why dropout of extreme values would lead to a reduction in variability.

There is slight efficiency gain in the estimation of a cluster-level effect with MAR dropout, in independence and auto-regressive GEE, but not in exchangeable or unstructured GEE models fitted to the same data. Refer to figure 3.1. There seems to be an advantage to modelling the data as independent, or with weak correlation within clusters in this scenario. This gain in efficiency, however, is at the expense of substantial bias in the parameter estimate.

With a binary response, MAR and MNAR dropout leads to an efficiency gain in the random effects model fitted in MLwiN, whereas all other models have a reduced efficiency. Again, this efficiency gain is at the cost of a large bias in the parameter estimate.

3.5 Comparison to other simulation studies

Park [2] investigated the sensitivity of GEE models and random effects models to missing data that are MCAR for gaussian response data. He found through simulating longitudinal gaussian data with 30 and 50 clusters and 4 observations per cluster, that with 30% of observations MCAR, GEE estimates were more biased than random effects estimates for cluster-varying covariates. The data were constructed with an exchangeable correlation structure and the GEE model was fitted with the correct correlation matrix. When the proportion of missing observations was 20% or less, GEE and random effects estimates were comparable. Although the missing data mechanism in Park's study was MCAR, the findings are not inconsistent with the findings of this study for cluster-varying covariates, except in this study, with dropouts MAR, the estimates of random effects and GEE exchangeable models became biased with lower proportions of missing observations. This is not surprising as the MAR missing data mechanism is more likely to introduce bias than the MCAR mechanism.

Touloumi et. al. [3] compared the bias and efficiency of random effects models and GEE models in the presence of missing data, when estimating the effect of time-varying covariates. Normal longitudinal data with 200 subjects and a maximum of 14 observations per subject were simulated with random intercepts and random slopes. The correct random effects model was fitted but an exchangeable correlation matrix was used for the GEE model, which is not correct when there is a random slope as well as a random-intercept. The performance of the two models was compared for incomplete datasets with MCAR, MAR and MNAR missing data mechanisms. The proportion of missing observations increased as the missing data mechanism changed from MCAR to MAR to MNAR, confounding the influence of the missing data mechanism and proportion of missing observations. Data with a MCAR mechanism had 20% of observations missing, MAR data had 50% and 60% of observations missing, while those with a MNAR mechanism had 60% and 80% missing.

The findings of their study demonstrated that the GEE method requires missing data to be MCAR to avoid bias, while random effects models were unbiased for both the MCAR and MAR missing data mechanism. Note, however, that this could also be due to the increased proportion of missing observations in the MAR datasets. The findings for both the GEE model and random effects model agree with this study; with gaussian data the random effects model is robust to missing data that are MAR whereas the GEE model is biased when the wrong working correlation matrix is chosen. In our study, we compared GEE with correctly specified correlation to GEE with mis-specified correlation, and found that the bias due to MAR dropout is unsubstantial when the appropriate correlation structure is chosen.

Little and Raghunathan [41] compared the summary measures method to maximum likelihood models for longitudinal Normal data with 30% of observations missing with MCAR, MAR and MNAR mechanisms. They estimated the effect of cluster-varying covariates. They found that the summary measures technique compared with ML models when data were missing completely at random but that ML methods were much less biased than summary statistics methods for incomplete data with MAR or MNAR mechanisms. This agrees with our findings for the data with cluster-varying covariates; the bias in the summary statistics method was very biased with MAR dropout when estimating a slope.

The work in this thesis builds on these previous studies, investigating the impact of dropout on parameter estimates in a much more comprehensive range of scenarios: The robustness of correctly specified GEE is compared to GEE with the wrong working correlation matrix; cluster-level covariates are compared to cluster-varying covariates; change in the ICC is investigated; binary and gaussian response data are compared; and MAR and MNAR dropout are compared under different degrees of missing data.

3.6 Conclusions and recommendations based on findings

When missing data are MNAR there is no reliable method of analysis for clustered data, even when the proportion of observations missing is very small. This highlights the importance of investigating the cause of dropout in an attempt to determine whether missing data are MCAR, MAR or MNAR. It is, therefore, important to record the cause of dropout for any subject that goes missing from a study.

Missing observations in data with a low ICC should be of greater concern than when the ICC is moderate, as it is possible that any bias is larger.

We prove that GEE for gaussian data with unbiased estimation of the correlation parameters are unbiased under MAR dropout. In the scenarios investigated in the simulation studies, the random effects model or GEE model with correctly specified or unstructured correlation matrix gave reliable results for gaussian response data. Greater bias in GEE with the wrong working correlation matrix have been shown to be due to the intrinsic weighting in GEE. The effect of choosing the wrong correlation structure in GEE, on the bias caused by the missing data, makes the unstructured GEE a sensible choice of model if there is any doubt as to the correct correlation structure of the data.

The bias in the correctly specified GEE seems to be due to bias in the estimation of the correlation coefficient. Further research is needed to investigate the advanced methods of estimation of α, discussed in section 3.4.2, especially when the response is gaussian. For example, it is not known whether these new estimation approaches would reduce the bias in GEE caused by MAR dropout with gaussian data, with the correct correlation structure, and also for different correlation structures.

The summary statistics is only a reliable method if missing data are MCAR. In particular, the summary statistics approach should be avoided when estimating a slope, if there are even a small number of missing observations.

Incomplete binary response data result in biased estimates when the success probability is as small as 0.2, even for MAR data. Penalised quasi-likelihood can have convergence problems. If using quadrature estimation in Stata, it is recommended that the number of quadrature points is increased to the maximum, and the adequacy of the model fit is then checked using the *quadchk* command. With a success probability of 0.5, binary data with a MAR missing data mechanism can be reliably analysed using a random effects model fitted by second-order penalised least squares, or by the quadrature method in Stata, or using a GEE model with correctly specified correlation structure. The *glamm* procedure in Stata 9, which uses adaptive quadrature in its estimation, should be investigated for any improvements in convergence and accuracy.

Chapter 4

Review of Tests of Missing Completely at Random

4.1 Introduction

The previous chapter carries a strong message about the implications of ignoring missing observations when analysing repeated measurements data, and highlights the importance of investigating the missing data mechanism.

Of the categories of missing data described in section 2.6.1, missing completely at random (MCAR) has the strongest assumptions. The summary statistics approach and generalised estimating equations (GEE), in general, require an assumption of missing completely at random (MCAR), in order to provide unbiased estimates. In section 2.6.3 is an explanation of why, when data are gaussian, GEE models with correctly specified correlation structure are robust to missing at random (MAR) data. However, as demonstrated in the previous chapter, GEE with MAR dropout may lead to substantial bias, even when the data are gaussian, if the correlation structure is misspecified. Therefore, GEE models require an assumption of MCAR when the data are non-gaussian, or when the data are gaussian and the correlation structure is unknown. The summary statistics method also requires an assumption of MCAR.

Several tests of MCAR have been published, adopting various strategies and applicable to various types of data. No formal comparison of these tests has been made. In this chapter, these tests are compared qualitatively in terms of their methodology, the types of data the tests can handle, how the tests explore the relationship between the covariates of interest and the probability of dropout, and accessibility of the tests to the user. A quantitative comparison of the tests is made in section 4.6, in which the rate of type I error and the power of the tests are compared on simulated datasets with various missing data mechanisms.

4.2 Available tests of MCAR

The following tests of MCAR were identified in the literature. There are similarities between many of the tests, and three groups of test emerge: *stratification tests*, *dropout tests* and *GEE tests.*

Stratification tests: These tests stratify the data by missing data pattern and test for heterogeneity between the strata. In Little's test, the mean response is compared between the strata, while in both of Park's tests the model of interest is fitted to each stratum and the model parameters are compared between strata.

1. Little (1988) [42]

2. Park et. al. (1993) [43]

3. Park (1997) [44]

   Dropout tests: These tests of MCAR dropout test if some function of the previous responses is a predictor of dropout. In both Ridout's and Diggle's tests the user chooses a function of the previous responses that is a plausible predictor of dropout, whereas in both of Listing's tests the function of the previous responses is defined to be the value of the response immediately before dropout. Ridout and Diggle both take account of the dependence of the response on a categorical covariate, such as treatment group, by stratifying the data by the levels of this covariate, but Listing's tests ignore the effect of any covariates.

4. Diggle (1989)[45]

5. Ridout (1991) [46]

6. Listing and Schlittgen (1998) [47]

7. Listing and Schlittgen (2003) [48]

   GEE tests: Not strictly tests of MCAR, these methods test the ignorability of the missing data in the GEE framework.

8. Chen and Little (1999) [49]

9. Qu and Song (2002) [50]

The methodology of the tests is described below.

Stratification tests

1. The earliest test of missing completely at random for multivariate data found in the literature is that of Little (1988). One simple approach to testing for MCAR for univariate gaussian data is to compare the mean observed response for responders and non-responders, using a t-test. Little extends this approach to multivariate gaussian data by testing whether responses for each missing data pattern are a random sample from a multivariate normal distribution with the same mean and covariance matrix across all patterns. The test statistic is a likelihood ratio test statistic, which is asymptotically χ^2 distributed under the null hypothesis. The assumption that the covariance matrix is constant across all patterns can be relaxed but the author believes this relaxed test is likely to be more sensitive to assumptions such as normality. The means by missing data pattern and the within-cluster covariance matrix are obtained by maximum likelihood estimation, which is implemented using an EM algorithm.

2. Park et. al. (1993) extends Little's approach to a test for categorical data. The data are stratified by missing data pattern and the model of interest, $f(\mathbf{Y} \mid \mathbf{X})$ is fitted to each stratum, where $\mathbf{Y}$ is the response variable and $\mathbf{X}$ is a matrix of the covariates of interest. A generalised Wald test is then carried out to test whether the stratum-specific regression parameters are homogeneous. Park et. al. used weighted

least squares to fit multivariate multinomial models, but suggest that generalised estimating equations could be used to estimate the stratum-specific parameters. Whereas Diggle and Ridout adjust for important predictors of the response by stratifying into homogeneous groups, this approach is more flexible, easily handling several covariates, continuous covariates, or even interaction between covariates, by including them in the model of interest. The test is therefore likely to be sensitive to selection of the model of interest.

3. Park (1997) proposed a test for MCAR which uses generalised estimating equations. Missing data patterns are identified and indicator variables, $\mathbf{Z}$, are constructed to label subjects by the pattern to which they belong. A GEE model is fitted, $f(\mathbf{Y} \mid \mathbf{X}, \mathbf{Z})$, of the same structure as the model of interest, but with the indicator variables, $\mathbf{Z}$, as covariates, as well as covariates of interest, $\mathbf{X}$. A generalised wald test is then used to test if all parameters corresponding to the missing data pattern indicators are zero. The relationship between the covariates of interest and the missing data mechanism can be examined by including interaction terms of missing data pattern and the covariates and testing whether the corresponding parameters are zero. The way the test deals with dependence of missingness on the covariates is discussed further is section 4.4. Like Park's earlier approach, this test is flexible to data with several covariates and / or continuous covariates.

A major advantage of Park's test is that it produces parameters of the missing data mechanism that can be interpreted in a meaningful way; the parameter corresponding to a particular missing data pattern is the expected difference in the response for subjects with that missing data pattern, compared to the subjects that have complete data, after adjusting for the other covariates.

Dropout tests

4. Diggle's method tests whether, within each treatment group, or equivalent homogeneous group, and at each time-point, the subjects about to drop out are a random sample of all subjects that have not yet dropped out. At time-point j, a response function, $h_j(y_{i1}, ..., y_{ij})$, is defined as some function of all values of the response observed up until time j. For each homogeneous group, and at each time-point, the mean value of the response function of subjects about to drop out, $\overline{h_j}$, is compared to the mean value of the response function of all subjects that remain in the study at that time-point, $\overline{H_j}$. The choice of response function depends on the hypothesised missing data mechanism. For example, if it is suspected that a succession of low (or high) measurements leads to patient dropout, the mean response may be chosen as the response function.

The test statistic for the jth homogeneous group, $\overline{h}_j$, asymptotically, has a normal null sampling distribution, with mean $\overline{H_j}$, and variance equal to the variance of all values of the response function in group j

multiplied by a function of the number of subjects to drop out and the number of subjects remaining at time j. If the sample size is small, the null sampling distribution of $\overline{h}_j$ must be obtained exactly. Hypothesis tests within all P groups at all except the last of the J time-points lead to $(J-1)P$ p-values. The resulting p-values are either inspected by eye, or a Kolmogorov-Smirnov test is carried out to test whether the p-values are a random sample from a Uniform $[0,1]$ distribution. It is likely that the test of uniformity for the p-values will lack power because the number of p-values will be small.

5. Ridout's test is based on Diggle's test, and aims to be more powerful, more straightforward, and more flexible to complications in experimental design. As in Diggle's test, a response function, h_j, is defined for each time-point, j, to be a function of the subject's response up to that time-point. If k indexes homogeneous groups, grouped by treatment group and time-point for example, in the kth group, T_k is the sum of the response function for subjects that drop out between the current and the next time-point. Instead of assuming that the mean of the values of the response function within each homogeneous group is normal and carrying out a separate hypothesis test for each group, Ridout suggests constructing one single statistic, T, to be the sum of the T_ks over all groups. Under the null hypothesis of MCAR, the distribution of this single test statistic should be a better approximation to the normal distribution than each of the test statistics in Diggle's test. If the expected

value and variance of T_k under the null hypothesis, conditional on the number of dropouts and the number of available observations in group k, are M_k and V_k respectively, then the mean and variance of T under the null hypothesis are $M = \sum_k M_k$ and $V = \sum_k V_k$ respectively. The test statistic, T, is then standardised and compared to the standard Normal distribution.

6. Listing and Schlittgen argue that the response of subjects about to drop out should be compared to that of subjects that remain in the study until the end, the *completers*. The rationale for this is that any gradual decrease (or increase) in response at each dropout point should not be masked. Although he is not explicit that this is what his test does, Park also compares each group of dropouts to the completers in his 1997 test. Listing and Schlittgen's test also differs from those of Diggle and Ridout in that the comparison is made on the value of the response immediately before dropout, rather than a function of the response chosen by the user. At each time-point, the differences in means between the subjects that drop out at that time-point and the completers, D_t, is computed. The test statistic, D, is a weighted mean of these D_ts, weighted by the the number of subjects that drop out at each time-point. D is asymptotically normal with mean zero under the null hypothesis. The data are assumed to be multivariate normal and the variance of the test statistic is a function of elements of the covariance matrix, estimated by fitting a multivariate normal

distribution to the data. One limitation of the test is that it will be lacking in power if few patients continue to the end of the study. The procedure does not adjust for the dependence of the response on any covariates.

7. Listing and Schlittgen (2003) proposed a non-parametric test for MCAR dropouts. Their test is based on the Wilcoxon summed rank test and compares the value of the outcome variable for subjects that are about to drop out with those that continue to be observed. In contrast to Listing's earlier test, the comparison is between subjects about to drop out and *all* other subjects remaining in the study, rather than comparing the subjects that are about to drop out with subjects that remain in the study until the end. Wilcoxon rank test statistics are computed for the difference between the dropout group and the group of remaining subjects at each point in time. The test statistic used is simply the sum of the individual rank test statistics at each time-point, which is asymptotically distributed as a standard Normal distribution under the null hypothesis. A limitation of the procedure is that, like Listing's earlier test, it does not extend to more than one treatment group.

GEE tests

8. Chen and Little developed a test using generalised estimating equations. Unlike Park's 1997 test, which fits one GEE model to all the data and tests whether parameters corresponding to the missing data pattern are zero, the data are stratified by missing data pattern, and

the model of interest, $f(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\theta})$ is fitted to each stratum using GEE. In some strata the model parameters, $\boldsymbol{\theta}$, may not be identifiable and maximum identifiable parameters are estimated instead. A generalised Wald test is used to detect differences between stratum-specific estimates of $\boldsymbol{\theta}$. The test is not strictly a test of MCAR, but a test of ignorable missingness in the GEE model, testing whether the expectation of the estimating equations taken over the incomplete data is zero. The null hypothesis is that the stratum-specific estimating equations, unadjusted for missing data, all have expected value zero. The authors note that the test will suffer from loss of power if any of the strata are sparse, and suggest that missing data patterns are grouped to avoid this. Even with this grouping, the power of the test will be sensitive to the numbers of subjects in the strata.

9. Qu and Song developed a test of ignorable missing in GEE based on the approach proposed by Chen and Little. The null hypothesis is again that the estimating equations have expected value zero, despite the missing values in the data. The procedure involves computing one test statistic, avoiding the need to solve separate GEE for each missing data pattern. The test is designed to be superior to Chen and Little's in the case of a small number of observations per missing data pattern because it does not require the assumption of normality of the stratum-specific regression parameters necessary for the Wald test. Qu et. al. extended the use of a function called the Quadratic Inference Function,

Q, to the GEE framework [51] which provides an alternative approach to GEE for marginal models. Q is the sum of the scores for each missing data pattern, weighted according to the variance of the score for each pattern. The values of the model parameters that minimise the quadratic inference function, $\boldsymbol{\theta}$, are the solutions to the estimating equations. The test statistic for the test of ignorable missingness is Q evaluated at $\boldsymbol{\theta}$, $Q(\hat{\theta})$. Under the null hypothesis the test statistic is zero, and is asymptotically distributed as χ^2.

4.3 Types of data the tests can handle

The tests of Little (1988), Park (1993), Park (1997), Chen (1999) and Qu (2002) apply to all types of clustered data, including repeated measurements data. The other tests, those developed by Diggle (1989), Ridout (1991), Listing (1998) and Listing (2003), apply only to longitudinal data. The former five tests apply to dropout, where once a subject leaves the study they do not return, as well as intermittently missing data, where a subject returns to the study at a later period after an observation has been missed. All other procedures, those designed for longitudinal data, test for MCAR dropout only.

The tests for MCAR vary in the types of outcome they can handle. Those of Diggle (1989), Ridout (1991) and Listing (1998) only apply to gaussian data. Park's test (1993) was developed specifically for categorical data, while the procedures developed by Park (1997), Chen (1999) and Qu (2002) can handle data that fit into the GLM framework. The test proposed by Listing in 2003 is non-parametric and is designed for non-gaussian quantitative data.

4.4 The influence of covariates on missing data

Little and Rubin's categorisation of missing data considers only the relationship between response variables and the missing data mechanism. There is a further category of MCAR, as defined in section 2.6.1, where missing data are classified as *covariate-dependent* missing if the missing data mechanism is independent of the response variables, conditional on the covariates. Models that are robust to MCAR data are robust to covariate-dependent missing data, provided the relevant covariates are included in the model.

Only Park's 1997 test for MCAR explicitly tests for the dependence of the missing data mechanism on the covariates. By including the covariates of interest as well as indicators of the missing data pattern in the GEE model, the procedure tests for covariate-dependent missing, returning a result of MCAR if the response is independent of the missing data pattern, after adjusting for the covariates of interest. The procedure also allows the relationship between the missing data mechanism and the covariates to be investigated by including an interaction term between the covariates and indicator variables for the missing data pattern in the regression model. This tests whether, not only is the response independent of the missing data pattern, but also whether the effect of the covariates on the response is modified by the missing data pattern.

Although none of the authors of the other tests state that their procedures consider covariate-dependent missing a category of MCAR, stratifying the data into homogeneous groups is one method of taking account of covariate-dependent missingness [9], if the covariates are categorical. Diggle introduced the idea of stratifying the data into homogeneous groups, in his test for MCAR, and this approach was adopted by Ridout (1991), but the only covariate used to stratify the data, apart from the missing data pattern, is treatment group. If there are many levels of the covariate and / or several covariates, the data within each group may become sparse, and stratification by homogeneous group becomes un-feasible.

4.5 Accessibility of the tests to the user

None of the tests in the literature can be carried out directly using tests and models available in standard statistical software packages. In each test, calculation of the test statistic involves more than a trivial amount of computation. Some of the test procedures are simpler to implement and / or understand than others.

The following tests are straightforward to understand and reasonably simple to implement: Little (1988), Ridout, Park (1993), Park (1997), Listing (1998), Listing (2003). These tests all involve test statistics most applied statisticians would be familiar with, and the procedures are straightforward.

The test proposed by Diggle involves two steps; firstly a set of several hypothesis tests, and then a hypothesis test on the resulting p-values. This is unnecessarily cumbersome, when there are now several alternative simpler procedures, and is likely to lack in power because of the reduction in information at the first step. Chen's test is not simple to understand because it involves finding maximum identifiable parameters. The approach suggested by Qu is not very accessible because the user must familiarise them-self with the use of the quadratic inference function.

4.6 Quantitative comparison of the tests

Those tests deemed sufficiently straightforward to be useful to applied statisticians were compared quantitatively using simulated datasets. The simulation study was limited to gaussian repeated measurements data with dropout rather than intermittent missingness. The Park (1993) test was eliminated as it was developed for categorical data, and the Listing (2003) test was not included in the simulation study either, as it is a non-parametric test for non-gaussian data. The four tests from the literature that were compared quantitatively are:

1. Little (1988)
3. Park (1997)
5. Ridout (1991)
6. Listing and Schlittgen (1998)

4.6.1 Data simulation

The tests of MCAR were run on simulated hierarchical Normal data. The dataset simulated was longitudinal with 100 subjects and 4 repeated measurements per subject. 1000 datasets were simulated for each scenario, and the same 1000 datasets were used to compare the tests, in each scenario. The model used to simulate the data is as follows:

$$y_{ij} = \alpha + u_i + \beta_1 treat_i + e_{ij} \tag{4.1}$$

$$u_i \sim N(0, \sigma_u^2)$$

$$e_{ij} \sim N(0, \sigma_e^2)$$

where *treat*$_i$ indicates which the treatment group the subject has been assigned to. Parameter values were chosen to be similar to those estimated from the repeated measurements asthma clinical trial [33]. Half of the subjects were randomised to the treatment group. The covariate, *treat*$_i$ relates to the probability of a subject dropping out of the study. The mean proportion of subjects dropping out at each time-point is given in table 4.6.1.

Table 4.1: Distribution of missing data patterns

Dropout time	**Pattern**	**Proportion subjects**
2	X . . .	0.08
3	XX . .	0.07
4	XXX .	0.07
5	XXXX	0.78

Subjects were simulated to drop out of the study in a selection model framework, as defined in section 2.6.6.

The following scenarios were simulated:

1. (a) MCAR dropout

    (b) MCAR covariate-dependent dropout

2. MAR dropout. Dropout dependent on immediately previous response with weak dependence on treatment group

    (a) Weak MAR dropout

    (b) Moderate MAR dropout

    (c) Strong MAR dropout

3. MAR dropout. Dropout dependent on immediately previous response with strong dependence on treatment group

    (a) Weak MAR dropout

    (b) Moderate MAR dropout

    (c) Strong MAR dropout

4. MAR dropout with probability of dropout dependent on mean of all previous values of response - moderate MAR dropout with weak dependence of dropout on treatment group.

Missing data simulation

A dichotomous variable, D_{ij} is defined to indicate whether a subject drops out at time j, given that they were present at time $j-1$. The probability of the ith subject dropping out at period j is then $p_{ij} = P(D_{ij} = 1 \mid D_{i,j-1} = 0)$. z_{ij} is a function of the previous values of the response, equal to the immediately previous response of the ith subject, $y_{i,j-1}$, in scenarios 2 and 3, and the

Table 4.2: Values of the parameters of the dropout mechanism in the different scenarios

Scenario	**Strength of MAR mechanism**	ϕ_1	ϕ_2
2a	Weak	0.005	0.2
2b	Moderate	0.01	0.2
2c	Strong	0.02	0.2
3a	Weak	0.005	0.8
3b	Moderate	0.01	0.8
3c	Strong	0.02	0.8
4	Moderate	0.01	0.2

mean of all previous responses of the ith subject, $\overline{y_{i,1,..,j-1}}$, in scenario 4. The model for dropout is then:

$$logit(p_{ij}) = \alpha_j + \phi_1 z_{ij} + \phi_2 treat_i \tag{4.2}$$

Three values of ϕ_1 are chosen to give three levels of increasing departure from MCAR dropout: Weak, moderate and strong MAR. ϕ_2 governs the influence of the covariate, *treat*, on the probability of dropout. The value of ϕ_2 is small in comparison to its standard error in scenario 2, and large compared to its standard error in scenario 3. The values of these parameters are summarised in table 4.2.

4.6.2 Results

Table 4.3 displays the results of the Ridout, Listing and Park tests run on 1000 datasets in each scenario. "Ridout previous" refers to Ridout's test with the immediately previous value of the response as the response function, h_j, that predicts dropout, while "Ridout mean" is the test with the mean of all previous values of the response as the predictor of dropout.

It should be noted that a type I error, in the context of tests of MCAR, corresponds to the false conclusion that an assumption of MCAR cannot be made. The null hypothesis in these tests is that the missing data are MCAR, and a small p-value represents evidence against this hypothesis. Counter to convention, a high type I error rate, therefore, corresponds to a "conservative" test. Conversely, a powerful test has a low rate of incorrect conclusions that dropout is MCAR. Conventionally, type I errors are considered more serious than type II errors, because the former involves drawing an incorrect conclusion, while the latter means that the opportunity to draw a conclusion was missed. In these tests, the power of the tests is more important than their type I error rate. The consequence of assuming MCAR when the data are MAR is that the choice of methods for the analysis is more limited. This is much less serious than incorrectly concluding the missing data are MCAR; applying a method that requires missing data to be MCAR to data with a MAR dropout mechanism could give biased results.

Park's test offers the option of including in the GEE, interaction terms between model covariates and missing data pattern indicators. The covariate of interest in these data is treatment group. Introducing these interaction terms allows us to test for heterogeneity in the treatment effect, and not just the intercept, between missing data patterns. The marginal model in the test of MCAR is then:

Table 4.3: Proportion of datasets where the tests are statistically significant under the various scenarios

Scenario	**Missing**	**Strength**	**Ridout previous**	**Ridout mean**	**Listing**	**Park**	**Little**
1 (a)	MCAR	-	0.116	0.128	0.062	0.230	0.180
1 (b)	MCAR	-	0.143	0.164	0.067	0.258	0.209
2 (a)		weak	0.842	0.878	0.169	0.211	0.116
2 (b)	MAR	moderate	0.948	0.955	0.408	0.330	0.185
2 (c)		strong	0.994	0.992	0.880	0.677	0.475
3 (a)		weak	0.894	0.915	0.222	0.193	0.106
3 (b)	MAR	moderate	0.965	0.973	0.489	0.335	0.205
3 (c)		strong	0.998	0.997	0.908	0.698	0.520
4	MAR	moderate	0.869	0.917	0.299	0.340	0.977

$$\begin{aligned} E(y_{ij} \mid treat_i, \mathbf{z}_i) = \quad & \alpha + \beta_1 treat_i + \beta_2 z_{2i} + \beta_3 z_{3i} + \beta_4 z_{4i} \\ & + \beta_5 z_{2i} treat_i + \beta_6 z_{3i} treat_i + \beta_7 z_{4i} treat_i \end{aligned} \tag{4.3}$$

where z_{2i}, z_{3i} and z_{4i} are variables that indicate subject dropout at periods 2, 3 and 4 respectively. The null hypothesis of the test of MCAR is $\beta_1 = \beta_2 = ... = \beta_7 = 0$. The size and power of this extended Park test for scenarios 1(a) to 4 is in table 4.4, in the column headed "Park covariate test". Here the tests are applied to data simulated from the same model as before, given in equation 4.1, but with a sample size of 200 clusters, as explained in the following paragraph.

The problem with this model is that there are many parameters to estimate and, although the generalised estimating equations did, in general, converge, the robust covariance matrix was not always invertible, and the Wald test

Table 4.4: A comparison of the two alternative Park tests

Scenario	Missing data	Strength	Park	Park covariate test
1 (a)	MCAR	-	0.110	0.492
1 (b)	MCAR	-	0.111*	0.456*
2 (a)		weak	0.177	0.495
2 (b)	MAR	moderate	0.393	0.664
2 (c)		strong	0.896	0.947
3 (a)		weak	0.192	0.543
3 (b)	MAR	moderate	0.378	0.670
3 (c)		strong	0.870	0.947
4	MAR	moderate	0.374	0.648

* 500 clusters needed to make the test statistic obtainable in the majority of datasets.

statistic could not, therefore, always be obtained. It was much more likely to be non-singular for a larger number of clusters however, and results are obtained for this model for data with 200 clusters. These results are compared to the Park test carried out above, with a fixed treatment effect across the missing data patterns, also with 200 clusters. In scenario 1 (b), 200 clusters were often insufficient to make the covariance matrix non-singular, and data were simulated with 500 clusters for this scenario.

4.6.3 Summary of findings

All tests were carried out at the 5% level. Listing's tests has approximately the correct significance level. Ridout's and Little's tests have an inflated significance level, with 10 to 18% of datasets incorrectly identified as not MCAR. Park's 1997 test had particularly high false positive rates, with approximately one quarter of MCAR datasets found to be significantly different from MCAR for the standard procedure, with this proportion rising to almost 0.5 for the test with interaction terms between treatment group and missing data indicators.

The most powerful test is Ridout's, which is particularly sensitive at detecting even small departures from MCAR, with over 80% of datasets being correctly identified as not MCAR when a weak departure from MCAR is simulated. Little's test is lacking in power when the probability of dropout in the simulated data is dependent on the immediately previous value of the response, but is the most powerful of the tests when the mean of the previous observations predicts dropout, as in scenario 4.

In scenario 3 the dropout mechanism is strongly dependent on the treatment group. This has little effect on the results of any of the tests, compared to scenario 2 where the treatment group dependence is weak. Dependence of dropout on the treatment group, in scenario 2 (b), also had little effect on the size of the tests.

Choosing the mean of all previous responses as the predictor of dropout rather than the immediately previous response increased the power of Ridout's test only very slightly. This increase was marginally the greatest when the data were simulated to have a dropout dependent on the mean of all previous responses, in scenario 4.

Introducing interaction terms between the missing pattern indicators and treatment group increased the sensitivity of Park's test to departures from MCAR. Park's test statistic, however, can be unobtainable, because of the number of model parameters involved. This is particularly a problem when interaction terms between the missing pattern indicators and treatment group are included in the model, and when the number of clusters in the data is not large. Including these interaction terms in Park's model increased the type I error rate, rising to almost 0.5 for both MCAR scenarios.

4.6.4 Discussion of findings

The simulation study shows that all four tests do, to a greater or lesser extent, discriminate between MAR and MCAR data. For the data simulated, Ridout's test far out-performs the others. Along with Park's test, Ridout's test adjusts for the treatment group. This is not, however, the feature of the test leading to its high power; ignoring treatment group increases the bias between completers and dropouts and therefore increases the power. The distinguishing feature of Ridout's test is that it conditions on the number of dropouts in each missing data pattern. Conditioning on this nuisance parameter will reduce the size of the estimated variance of the test statistic compared to a test that does not condition on the number of dropouts, and seems a plausible reason for the high power of the test.

The lack of power in Listing's test is not surprising, as it does not make use of all the available data at each time-point, comparing the response of the subjects about to drop out only to the subjects that complete the study. Listing et. al. designed the test to be sensitive to a gradual increase or decrease in the response of the dropouts over time. The dropout mechanism used to simulate the missing data, defined in equation 4.2 was applied first to the data at period 2, and then to all those subjects that had not dropped out by period 3, and then to all those remaining at period 4. This means that the response of subjects dropping out earlier in the study tends to be more biased than those dropping out later, or in other words, the bias in the dropouts decreases over time. This is the type of bias that Listing's test was designed to detect.

Park's test is marred by its high false positive rate, particularly when covariate by dropout indicator interactions are included in the model. The problem seems to be multiple testing; significance in any one of the pattern indicators gives a significant result for the test of MCAR. Including interaction terms in the model increases the number of model parameters that are used in the test from three to six, and the type I error rate increases. When these interaction terms are not included in the model, the test suffers from a lack of power. In addition to the high type I error rate, the test statistic is frequently unobtainable when the sample size is not large. This is mostly a problem for the model that includes interaction terms.

Park's and Listing's test are appealing because they provide parameter estimates or test statistics that can be interpreted meaningfully. Park's approach provides parameter estimates that are estimates of the difference in response between subjects with each missing data pattern, and the completers, after adjusting for other covariates in the model. Listing's test provides an estimate of the mean difference between the response of the dropouts immediately before they drop out, and the response of the completers. An associated standard error is provided with this estimate. The test statistic proposed by Ridout is an estimate of the sum of all values of the response function for subjects that drop out. This is not easily interpreted. Although Ridout's test does not have a test statistic which can be meaningfully interpreted, it is the most powerful of the three tests compared quantitatively.

Dependence of dropout on treatment group had little effect on the results of the tests. Little's and Listing's tests make no adjustment for treatment group and it was expected that the power of these tests would increase when dropout was covariate-dependent because the additional bias in the response due to treatment group would not be adjusted for. In fact, there was a slight increase in power from scenario 2 to scenario 3 for the both tests but it is minimal. The type I error rate of both tests were also expected to increase when there is covariate-dependent MCAR, for the same reason. Again, there is an increase in size for both tests, but it is insubstantial.

With an intra-cluster correlation coefficient of 0.5, the results of Ridout's test were very similar whether the mean of all previous responses or only the previous response was chosen as the predictor of dropout. The largest increase in power would be expected for scenario 4, when the mean of the previous responses was used as the predictor of dropout in the data simulation. Even in this scenario, the increase in power was unsubstantial. The correlation within clusters was evidently sufficiently high to make the previous response as good a predictor of dropout as all previous responses.

Little's test improves dramatically when the simulated dropout is dependent on the mean of previous responses, rather than the immediately previous response, while this has little effect on Ridout's test. The power of Ridout's test is so high when dropout depends on the previous response, that no great increase in power can be expected.

It should be noted that all of the approaches require an assumption of MAR in order to test the missing data for MCAR, because MNAR, by definition, cannot be tested for. A scenario could exist in which there is no relationship between the observed data and the probability of dropout, but where the missing data are MNAR. But intra-cluster correlation makes this unlikely. A scenario where there was no dependence of dropout on the observed data would only have MNAR missing data if patients had unbiased observations while they remained in the study, followed by a sudden bias in their response. A dramatic change in circumstances would be needed for this to happen.

4.6.5 Comparison to the findings of other simulation studies

Listing and Schlittgen used simulation studies to investigate the power and size of both of their tests, as did Park and Lee for their 1997 test, to investigate the small sample properties of their procedure. With many simulations run by the authors (1000 for the Listing test and 10 000 for the Park test), the size of the test was correct when the significance level was set to 0.05. The type I error rate was found to be higher for the data simulated here, especially for Park's test.

Park and Lee carried out simulation studies to test the small sample properties of their test. They used binary data with missing data mechanisms that were quite far from the assumptions of MCAR. The odds of success on the outcome variable was either 1.5 or 4.5 times greater for the incomplete data than the odds for subjects in the complete group. With these missing data mechanisms the power of the test, found from simulating 1000 datasets, was less than 0.4 for datasets with 10% of cases incomplete and up to 100 cases per dataset. With 100 cases in each dataset, and an odds ratio of incomplete versus complete data of 4.5, the power of the test was greater than 0.9 when at least 30% of cases were incomplete. A comparison between Park's simulation study and this work is difficult because Park simulated a binary response. However, it is noted that Park found the type I error rate of the test to be correct, whereas it was found to be too high in these simulations. The findings about the power of the test from the two simulation studies are not in disagreement.

4.6.6 Conclusions and recommendations

From the findings of the quantitative comparisons of the tests, the recommended test of MCAR for dropout in gaussian repeated measurements data is Ridout's test. It is the most powerful without having a high type I error rate. Its power does not seem to depend on which function of the previous responses predicts dropout, and it is a conservative test, with a probability of incorrectly concluding that data are not MCAR of more than 10% and a power of at least 85% in these scenarios.

Little's and Park's tests cannot be recommended for data such as these, because of their lack of power; with moderate departures from MCAR, Little had a power of less than 20%. Although Park's test is more powerful when interaction terms between dropout pattern indicators and covariates of interest are added to the model, where the test statistic is obtainable, the type I error rate is then unacceptable. Listing's test, although anti-conservative in these scenarios, is a viable alternative to Ridout's test. Here, it correctly identified moderate departure from MCAR in at least 40% of datasets, rising to about 90% for strong departure from MCAR.

In addition to these formal tests of MCAR, a visual comparison of the response of the dropouts compared to subjects that remain in the study is useful. This is demonstrated on the asthma clinical trial data in section 5.3.

Chapter 5

A repeated measurements asthma clinical trial with patient dropout

5.1 Introduction

As explained in chapter 1, a motivation for the work in this thesis was a longitudinal asthma clinical trial, from which a substantial number of patients dropped out. In this chapter, the data from the trial are analysed using the random effects model, GEE and summary statistics method, described in chapter 2. In chapter 4, several formal tests of MCAR were compared, and from the results of a simulation study, one particular test was recommended. The missing data mechanism in the asthma trial is investigated, both visually, and using this recommended test of MCAR. The conclusions about the missing data mechanism are used, together with the findings from chapter 3, to evaluate the likely bias caused by dropout in the asthma clinical trial.

The data are from a double-blind randomised clinical trial to compare the safety and efficacy of beclomethasone dipropionate (BDP), salmeterol xinafoate (SM) and placebo in corticosteroid-naive children with mild to moderate, chronic, stable asthma [33]. The primary outcome was peak expiratory flow rate (PEFR), a measure of lung function that tends to be low in asthmatics. PEFR was measured daily, with averages taken over a two-week run-in period and then over four three-monthly periods once treatment was started. Over 20% of children dropped out before the end of the trial. Baseline measures were taken on age, height and treatment centre. A total of 241 children, aged 6 to 14 years, were included in the study, 233 of which had a baseline PEFR measurement and at least one post-treatment measurement. Of these 233 patients, 77 received placebo, 80 received BDP and 76 SM.

5.2 Model fitting

Suitable methods to compare the overall efficacy of the treatments include the random effects model, GEE model or the simple summary statistics method. A random-intercept model is fitted, which has the same correlation structure as an exchangeable GEE model. Unstructured generalised estimating equations model the intra-cluster correlation structure from the data. The closer the estimated intra-cluster correlation structure is to the true structure, the greater the model efficiency. However, the unstructured GEE involves the estimation of a greater number of parameters, and therefore potentially reduces efficiency. The simulation studies in chapter 3 demonstrate that incorrectly specifying the working correlation matrix in GEE increases the bias caused by any missing data that are not MCAR, and that choosing an unstructured correlation can limit this bias. Both unstructured and exchangeable GEE are fitted to the data. These random-effects and GEE models are compared to the summary statistics method.

The random-effects model and GEE model, for the jth response on the ith child, y_{ij}, with treatment group indicators, BDP_i and SM_i, and other baseline measures, $\mathbf{x}_i$ as covariates, are summarised in equations 5.1 and 5.2 respectively. A linear time effect, $time_{ij}$, equal to the time-period, j, was also modelled in the random-effects and GEE models. In the summary statistics approach the repeated measures of the ith child are reduced to a mean response over time, m_i, and these summary measures are then modelled as if they were univariate response data. A normal linear regression model is fitted to the mean response, adjusting for baseline covariates, as given in equation 5.3. It is not possible to model a time effect in the summary statistics approach, when the repeated measures are reduced to a mean. As discussed in section 2.4, the choice of summary measure depends on the research question, and the shape of the data. Here, an average response over time is required; the measurements are equally spaced, so there is no need to weight the measurements according to their distance apart, by taking an area under the curve, for example, and the unweighted mean is chosen.

$$E[y_{ij} \mid u_i] = \alpha + u_i + \beta_{bdp} BDP_i + \beta_{sm} SM_i + \beta_{\mathbf{x}} \mathbf{x}_i + \beta_t time_{ij} \tag{5.1}$$

$$u_i \sim N(0, \sigma_u^2)$$

$$E[y_{ij}] = \alpha + \beta_{bdp} BDP_i + \beta_{sm} SM_i + \beta_{\mathbf{x}} \mathbf{x}_i + \beta_t time_{ij} \tag{5.2}$$

$$E[m_i] = \alpha + \beta_{bdp} BDP_i + \beta_{sm} SM_i + \beta_{\mathbf{x}} \mathbf{x}_i \tag{5.3}$$

Table 5.1: Parameter estimates of the various models fitted to the asthma clinical trial data

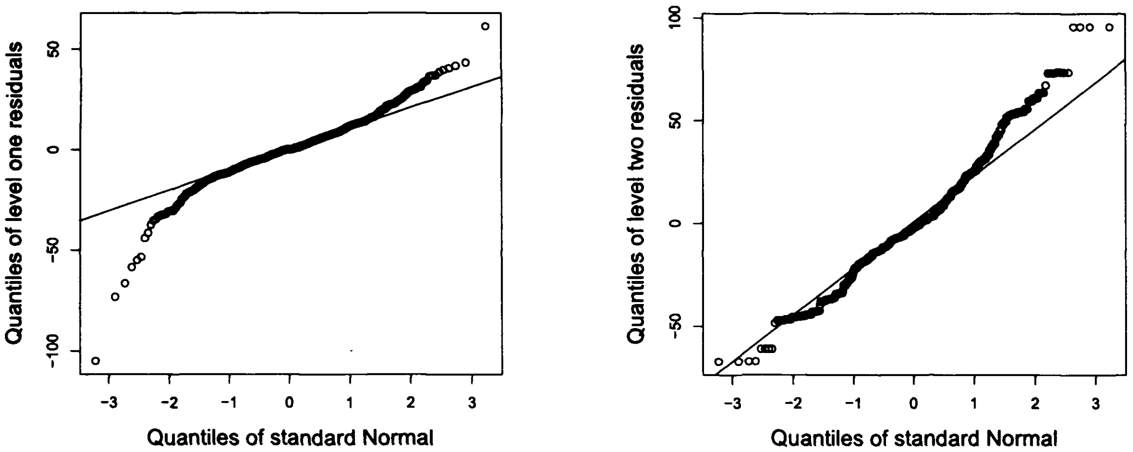
	Estimate (95% confidence interval)	
	Random effects (5.1)	GEE unstr. (5.2)
BDP–placebo	12.53 (4.07,20.99)	10.84 (2.56,19.12)
SM–placebo	18.46 (9.81,27.10)	16.90 (8.44,25.37)
baseline PEFR	0.697 (0.606,0.788	0.715 (0.626,0.804)
age	4.84 (1.58,8.10)	4.26 (1.07,7.44)
height	0.566 (-0.056,1.19)	0.614 (0.01,1.22)
time	7.36 (6.32,8.39)	5.48 (4.69,6.27)
σ_u	25.22 (22.62,27.82)	-
σ_e	16.41 (15.47,17.34)	-
ICC	0.703 (0.650,0.751)	-
	GEE exch. (5.2)	Summ. stats. (5.3)
BDP–placebo	12.55 (3.90,21.20)	13.98 (5.17,22.80)
SM–placebo	18.47 (9.63,27.30)	19.15 (16.17,28.12)
baseline PEFR	0.697 (0.604,0.790)	0.700 (0.604,0.795)
age	4.84 (1.51,8.17)	4.45 (1.05,7.84)
height	0.567 (-0.07,1.20)	0.632 (-0.014,1.277)
time	7.34 (6.35,8.34)	-
σ_u	-	-
σ_e	-	-
ICC	-	-

These models were fitted in Stata with baseline PEFR, age and height as baseline measures, $\mathbf{x}_i$. The parameter estimates from the models are summarised in table 5.1.

Normal plots of the level one and level two residuals in the random-effects model, in figure 5.1, show approximate normality, although there seems to be some kurtosis.

The choice of model does not affect the significance of either treatment effect.

Figure 5.1: Normal quantile quantile plots of the level one and level two residuals in the random-effects model 5.1



The parameter estimates are affected slightly by which model is fitted, while the width of the confidence intervals is similar across the models. Notice that the random intercept model and exchangeable GEE give virtually identical estimates, because, at least for complete data, the two models are equivalent. The unstructured GEE estimated the correlation between the first observation and the other three observations to be very high, approaching 1, while the correlation between all other observations was estimated to be close to 0.5. This does not suggest exchangeable correlation, and explains why the estimates from the unstructured and exchangeable GEE do not agree.

Interestingly, in these data the response is already a summary measure, as it is the mean of daily peak flow rate averaged over four three-monthly periods. This will reduce the within-subject variance, and lead to a higher ICC than if the daily measures had been used as the reponse. Note from the findings of chapter 3 that missing data are more of a concern when the ICC is low, because with less correlation within clusters, the loss of each observation corresponds to a greater loss of information.

The summary statistics method, though crude, gives fairly similar results to the more sophisticated models, and leads to the same conclusions about the efficacy of the treatments. A time effect cannot be modelled when the repeated measurements are reduced to a mean. In these data, ignoring the trend over time has no great impact on the results, but this in no way generalises to all scenarios. For example, see scenario 6 in chapter 3, in which there is a period-by-treatment interaction, which the summary statistics method cannot model.

5.3 Missing data

The frequency of missing data patterns is summarised in table 5.2. Overall, 12.2% of observations were missing, with more observations missing from the placebo group than either treatment group: 16.9% of observations in the control group were missing, compared to 7.2% in the SM group and 12.5% in the BDP group.

Investigation of the missing data mechanism helps to assess the likely impact of dropout on the parameter estimates. In chapter 4, a test of MCAR proposed by Ridout, is found to be the most powerful of all the published tests of MCAR, and has an acceptable type I error rate. It can be used to test for dependence of either the immediately previous value of the response, or the mean of all previous values, on the probability of dropout. This test was

Table 5.2: Frequency and percentages of missing data patterns in the data

Pattern	Cases	% cases
X . . .	27	11.6
XX . .	12	5.2
XXX .	9	3.9
XXXX	185	79.4
Total	233	100

run on the asthma clinical trial data, and there was no statistically significant evidence that either the previous response (p=0.71), or the mean of all previous values of the response (p=0.66), are associated with the probability of dropout.

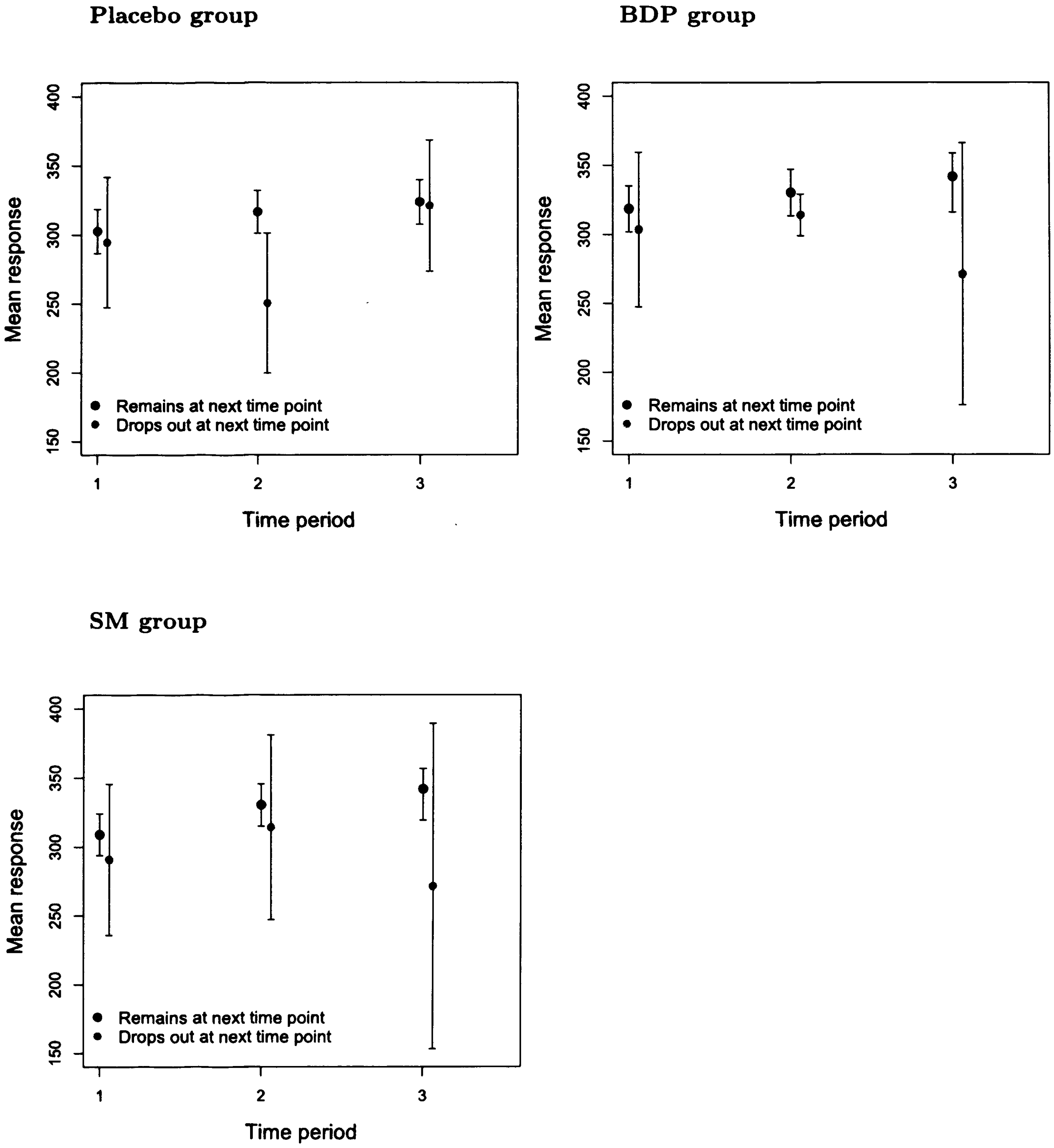
In section 2.6.4 it was suggested that inspection of the observed data helps to distinguish between MCAR and MAR dropout. Carpenter et. al. [27] suggest plotting the means (± 2 standard errors) for the subjects that drop out at the next time-point compared to those that do not. Such plots of the asthma data are shown in figure 5.2, separately for each treatment group. They reinforce the results of Ridout's test, showing that although the means of subjects just before they drop out are systematically lower than the means of subjects that remain in the study, there is no evidence that this difference is statistically significant.

There is no evidence that the missing data are not MCAR. As discussed in section 2.6.4, it is not possible, by definition, to test for MNAR dropout. The missing data would only be MNAR, however, if there were a sudden change in the symptoms of the dropouts, after their last observation was recorded. The main cause of patient withdrawal (46% of all patients that withdrew) was asthma exacerbation, with less frequent causes being adverse events, noncompliance or protocol violation, and a quarter of dropouts withdrawing for unknown reasons. With so many patients dropping out due to asthma exacerbation, it is difficult to rule out MNAR dropout, especially as the subjects that dropped out of the study did not have particularly low PEFR before dropping out, and therefore the observed data are insufficient to estimate the missing data.

Chapter 6 introduces a Bayesian model for repeated measurements data with MNAR dropout, when the cause of dropout is recorded for each subject. In the asthma clinical trial data, the dropout cause is not available at the patient level, and the Bayesian model cannot be applied. In chapter 6, data are simulated based on the asthma clinical trial data, in an attempt to make the simulations as realistic as possible.

In light of the findings of chapter 3, 12% of observations missing from repeated measurements data, that are possibly MNAR, could cause substantial bias to the parameter estimates. If the dropout mechanism is MNAR, the parameter estimates are unreliable. The unstructured GEE are possibly slightly more reliable than the exchangeable GEE, but with MNAR dropout, no model that ignores dropout is unbiased. The high ICC in the observed data potentially reduces the bias caused by any MNAR dropout, as discussed in section 3.2.2, but this relies on the within-cluster correlation between the observed and unobserved data also being high, which is a strong assumption.

Figure 5.2: A visual comparison between the mean response of subjects about to drop out, and those that remain in the study at the next time point



Placebo group

BDP group

SM group

In conclusion, it is not satisfactory to ignore the possibility of MNAR dropout in the analysis of these data. Ideally, the causes of dropout would be recorded at the patient level, and a Bayesian model incorporating this information would then be possible. In the absence of this information, a sensitivity analysis should be carried out, in which dropout is modelled, and several plausible values are chosen for the dropout parameters.

Chapter 6

Multiple causes of dropout in repeated measurements data

6.1 Introduction

Data from a repeated measurements asthma clinical trial were analysed in chapter 5. Over 20% of patients dropped out of the study before the end of the trial, and investigation of the dropout led to the conclusion that MNAR dropout was a definite possibility, especially as the main cause of dropout was exacerbation of symptoms. The simulation studies in chapter 3 demonstrate that, if dropout was MNAR, this quantity of missing observations would cause serious bias in the parameter estimates of the models. An alternative approach is needed, in which the bias is adjusted for, and the additional uncertainty due to missing data is taken into account, by modelling the dropout mechanism. Molenberghs et. al. [52] distinguish very clearly between uncertainty due to sampling error and uncertainty due to "ignorance" of the values of missing observations.

In clinical trials researchers are encouraged, or even required by protocol, to record cause of patient dropout. This information will usually be found in the publication of the trial, and will often be used in any discussion of the likely impact of the missing data on the parameter estimates. Three different causes of dropout were recorded in the asthma clinical trial: asthma exacerbation; adverse event and noncompliance or protocol violation. The cause of dropout helps clinicians to make judgements about the likely effect of the missing data. If this information is used informally in discussions about the missing data, it is only sensible that it should be used in any model of the dropout mechanism. Clinicians were interviewed about the information they would feel able to provide about the response of patients after dropping out of a study. See section 6.6. Both clinicians said that the single most important piece of information that would help them to give an opinion on this, was the cause of dropout.

Modelling MNAR dropout is necessarily subjective, and experts should be consulted when constructing the dropout model. In this chapter, a Bayesian model is proposed, for which prior distributions are elicited from clinicians, on the bias in the response of patients that drop out, separately for each cause of dropout. The Bayesian framework is not routinely used in the analysis of clinical trials, but Spiegelhalter et. al.[53] advocate its use, on the grounds that evidence external to the trial is already used informally in making decisions based on the evidence from the trial. For example, the degree of scepticism a clinician has about the new treatment is used in combination with the findings of the trial, when deciding whether or not to recommend the new treatment. Also, quantitative opinions are elicited from experts, or drawn from previous studies, when carrying out sample size calculations. This quantitative evidence could be used to construct prior distributions for a Bayesian analysis. The authors argue that a Bayesian framework allows external evidence such as this to be combined quantitatively with the data from the trial, in a more formal approach.

6.2 Literature on multiple causes of dropout, and Bayesian models for MNAR missing data

Discussion in the literature on incorporating information about causes of missingness into the analysis is very limited. Proposed methods can be crude, and are usually limited to cross-sectional data.

Gould [54] proposed a rather simplistic approach to handling multiple causes of dropout in which the response is replaced with ranks, and missing observations are imputed with ranks, according to their reason for dropout. For example, subjects that drop out of the study because they are "cured" are given a score corresponding to the highest ranking of all subjects in the study, and subjects that drop out due to intolerance to the treatment are assigned the lowest possible score. This method is crude, and limited only to data where the response is ordinal. It makes strong untestable assumptions about the dropout mechanism, and, as in all single imputation methods, ignores the uncertainty in the missing data. See Carpenter et. al. [55] for a further explanation of why single imputation methods are invalid. Lachin [56] proposed a similar ranking method for dropout in clinical trials, which is again crude, and has the same problems as Gould's method.

Rubin [57] wrote an early paper on eliciting prior information from experts on the expected response of subjects with missing observations compared to those subjects that responded. The method incorporates the additional uncertainty in parameter estimates due to the missing data, widening their credible intervals, but does not adjust for any bias due to dropout. It applies only to cross-sectional data, and does not model multiple causes of dropout. More recently, White et. al. [30] developed a Bayesian method that incorporates the bias due to dropout in the model of interest. This model, for cross-sectional data, is described in detail in section 6.4.1.

Dufouil et. al. [58] propose a sensitivity analysis for MNAR dropout, treating death differently to other causes of dropout. In the terminology of the survival analysis, they treat dropout as "failure" and death as "censoring", in a model for the probability of dropout. Their dropout model provides propensity scores in an inverse probability weighting method. Inverse probability weighting is introduced in section 2.6.3. The dropout model has previous values of the response, as well as the current response value, as covariates. The model parameters corresponding to the previous response are estimated from the data, and plausible values of the parameter for the current response are chosen, to form a sensitivity analysis. Beyond differentiating between death and other dropout, the approach does not extend to multiple causes of missing data.

6.3 Methods for Missing Not at Random data

MNAR is the most severe category of missing data, as it implies that the observed data are insufficient to predict the values of the missing observations. Therefore, subjective assumptions must be made about the dropout mechanism, which is then modelled jointly with the model of interest. Due to the subjective nature of the dropout model, the standard approach to dealing with MNAR data is to carry out a sensitivity analysis, in which a range of plausible models for dropout are fitted, including extreme case scenarios. This results in a range of plausible parameter estimates rather than a point estimate for each parameter.

The sensitivity analysis is often used as a test of the assumptions about the missing data mechanism, rather than a tool for analysing incomplete data with a MNAR mechanism. Sensitivity of the parameter estimates to different missing data mechanisms is tested, and where estimates are sensitive to the different mechanisms, the data may be deemed too flawed to be useful. Conversely, where the estimates are found to be reasonably robust to a range of plausible dropout mechanisms, the sensitivity analysis is often used as a justification for ignoring the missing data.

Alternatives to the sensitivity analysis have been proposed recently [29] [30], in which prior distributions are elicited for parameters of the dropout mechanism, and a Bayesian analysis is adopted, resulting in posterior distributions for the parameters of interest which incorporate experts' uncertainty about the dropout mechanism. Whereas the sensitivity analysis provides the clinician with a series of parameter estimates under several different missing data mechanisms, a Bayesian analysis averages over a range of possible missing data mechanisms, effectively weighting according to the prior distributions of the missing data parameters. The credible intervals incorporate, not only the likely bias caused by dropout, but also the additional source of uncertainty in the estimates, from the lack of information about the unobserved data.

A sensitivity analysis may be preferred by some researchers and clinicians because the effects of the various missing data mechanisms remain explicit. Clinicians examining the findings of a study are then able to informally interpret this information alongside their own beliefs about the likely missing data mechanisms. In contrast, the Bayesian approach offers a formal method of incorporating into the analysis, clinicians' beliefs about the uncertainty associated with the missing data. An advantage of this method is that it provides a single estimate of each parameter of interest which attempts to incorporate all sources of uncertainty about the parameters, including the missing data.

So far, these Bayesian methods have only been developed for cross-sectional data with a single cause of dropout. Here, this approach is extended to multiple causes of dropout and to repeated measurements data.

6.4 Modelling the dropout mechanism

6.4.1 A model for single cause of dropout

Joint modelling of the missing data mechanism and the model of interest can be considered in a pattern-mixture model (PMM) or selection model (SM) framework, defined as follows:

$$\text{SM} \quad p(\mathbf{Y},\mathbf{M}\backslash\mathbf{X},\theta,\phi) = p(\mathbf{M}\backslash\mathbf{Y},\mathbf{X},\phi)p(\mathbf{Y}\backslash\mathbf{X},\theta) \tag{6.1}$$

$$\text{PMM} \quad p(\mathbf{Y},\mathbf{M}\backslash\mathbf{X},\theta,\phi) = p(\mathbf{Y}\backslash\mathbf{X},\mathbf{M},\theta)p(\mathbf{M}\backslash\mathbf{X},\phi) \tag{6.2}$$

where $\mathbf{M}$ is a dichotomous variable to indicate whether an observation is missing or observed. $\mathbf{Y}$ is a vector of the response, $\mathbf{X}$ represents the covariates, and θ and ϕ are parameters of the model of interest and parameters of the missing data mechanism respectively.

In the pattern-mixture model framework, subjects with different missing data "patterns" are considered to be sampled from different distributions. For example, subjects that have complete observations have a different distribution to those that have one missing observations, which is different again to the distribution of subjects with only one measurement observed. The data are stratified according to which observations are missing in the response. A separate model is fitted to each stratum, $p(\mathbf{Y}\backslash\mathbf{X},\mathbf{M},\theta)$, and an overall parameter of interest is computed by taking a weighted average of the stratum-specific parameter estimates, weighted by the proportion of observations in each stratum, $p(\mathbf{M}\backslash\mathbf{X},\phi)$.

In contrast, in the selection model framework the distribution of all observations is considered to be $\text{p}(\mathbf{Y}\backslash\mathbf{X}, \theta)$, had we been able to observe them all. The subjects that drop out are assumed to be a selection of these complete data, with missing data mechanism, $\text{p}(\mathbf{M}\backslash\mathbf{Y}, \mathbf{X}, \phi)$.

A pattern-mixture model framework is chosen here, in which subjects with missing observations are considered to be sampled from a different distribution to those with complete observations. In the pattern-mixture model, parameters of the dropout mechanism are biases in the response, whereas dropout parameters in a selection model are odds ratios. Odds ratios are notoriously difficult to interpret, and therefore bias in the response is an easier quantity to elicit than an odds ratio. Consider the simple example of a cross-sectional study with treatment group, $treat_i$, as the only covariate. A pattern-mixture model for the continuous response, y_i of a subject, i, is given in equation 6.3, below.

$$y_i = \begin{cases} \alpha + u_i + \beta \text{treat}_\text{i} + e_i & \text{if } M_i = 0 \\ \alpha + u_i + \delta_1 + (\beta + \delta_2)\text{treat}_\text{i} + e_i & \text{if } M_i = 1 \end{cases} \tag{6.3}$$

where β is the treatment effect for the completers and δ_2 is the bias in the treatment effect of the incompleters compared to the completers of the study. δ_1 is the intercept bias of the incompleters compared to the completers. An overall treatment effect, β^*, is computed by taking a weighted average of β and $(\beta + \delta_2)$, weighted by the proportion of subjects in each stratum, as follows [28] :

$$\beta^* = \pi(\beta + \delta_2) + (1 - \pi)\beta \tag{6.4}$$

where π is the proportion of subjects that have missing observations.

Throughout this chapter, the parameters notated δ are referred to as bias parameters. The word "bias" here is used informally to mean the difference between the dropouts and completers.

In a sensitivity analysis, a range of plausible values would be substituted for δ_1 and δ_2, resulting in a range of possible values for the overall treatment effect, β^*. In a Bayesian framework, prior distributions are elicited for δ_1 and δ_2, leading to a posterior distribution for β^*, from which a point estimate, its standard error, and credible intervals can be obtained.

Carpenter et. al. [27] propose a pattern-mixture model such as this for a sensitivity analysis, in which fixed values are substituted for the parameters of the dropout mechanism, δ_1 and δ_2. Non-informative priors are specified for the parameters of the model of interest and WinBUGS is used to implement Markov Chain Monte Carlo (MCMC) estimation. MCMC is used for the model estimation, not because a Bayesian analysis is required, but because, by using vague priors for all parameters, it is a convenient approach to obtaining very good approximations to maximum likelihood estimates.

White et. al. [30] extend this pattern-mixture model approach to the Bayesian framework, eliciting prior distributions from experts for the parameters of the dropout mechanism, such as δ_1 and δ_2 in equation 6.3. Their model estimates the difference between interventions in a cross-sectional clinical trial scenario, with gaussian response. They assume normality for the prior distributions, and because of the simplicity of the scenario, are able to provide an approximate formula for the estimation of the treatment effect, and its standard error. They show that this approximation agrees well with MCMC estimation in WinBUGS.

6.5 Extending the method to longitudinal data and multiple causes of dropout

6.5.1 Extension to longitudinal data

In the asthma clinical trial data, PEFR tended to increase over time, and the model in equation 6.3 is extended to incorporate a linear time effect, that is not directly of interest. For simplicity, only two treatment arms are modelled, but extension to more than two treatment groups is trivial. Apart from trends over time, the main issue in the analysis of repeated measurements data is the dependence between observations on the same subject, as described in 2.1. A random effects model is used with a random intercept, so that subjects are allowed to have different intercepts, but the slope is the same for subjects within the same treatment group. This could be extended by allowing other parameters in the model to be random. A random-intercept model was used to analyse the asthma data.

In the MNAR scenario, by definition, the distribution of the observations of subjects that drop out of the study is different before and after dropout. Subjects that do not complete the study, *dropouts*, are modelled with bias, δ_1 before dropout, and bias δ_2 after they go missing. A new indicator variable,

g_{ij} is introduced, which indicates whether it is before or after the subject dropped out. The time of the last observation before dropout is represented by t^* in the following description of the model, in equation :

$$\text{Completers} \quad Y_{ij} = \alpha + u_i + \beta_1 \text{treat}_\text{i} + \beta_2 t_{ij} + \epsilon_{ij} \tag{6.5}$$

$$\text{Dropouts} \quad Y_{ij} = \alpha + u_i + \beta_1 \text{treat}_\text{i} + \beta_2 t_{ij} + \delta_1 (1 - g_{ij}) + \delta_2 g_{ij} + \epsilon_{ij}$$

$$g_{ij} = \begin{cases} 0 & \text{if } t_{ij} \leq t^* \\ 1 & \text{otherwise} \end{cases}$$

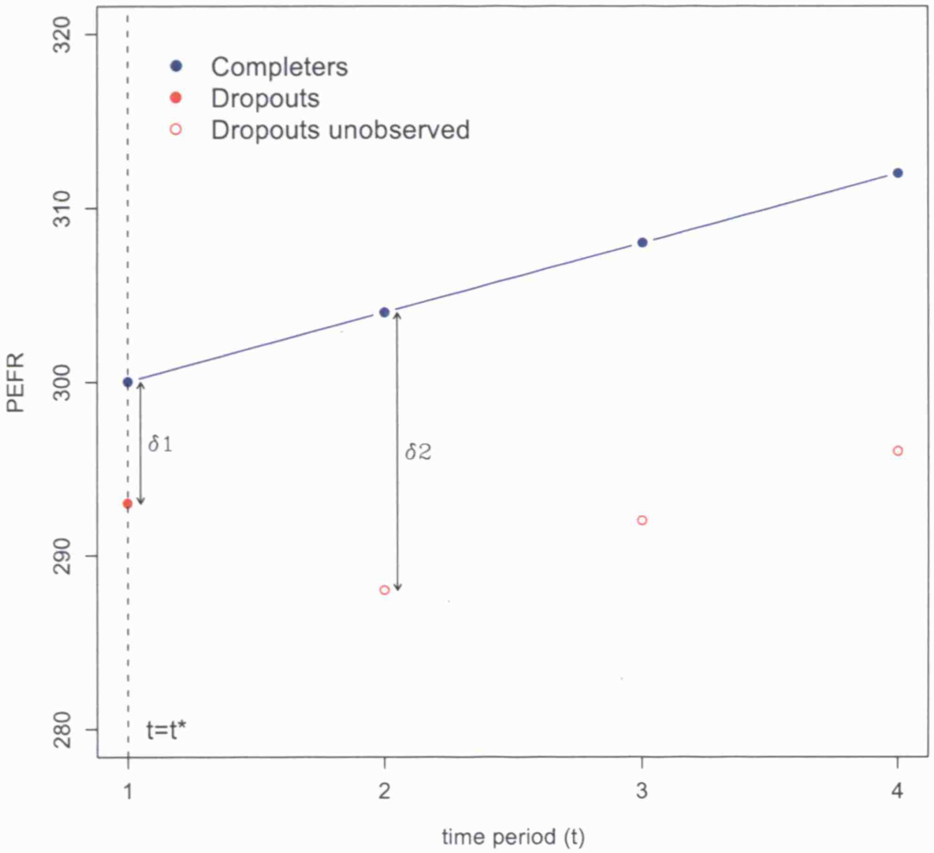
Which can also be written:

$$Y_{ij} = \alpha + u_i + \beta_1 \text{treat}_\text{i} + \beta_2 t_{ij} + \delta_1 (1 - g_{ij}) \text{drop}_\text{i} + \delta_2 g_{ij} \text{drop}_\text{i} + \epsilon_{ij} \tag{6.6}$$

where drop_i indicates whether or not a subject dropped out before the end of the study. The model is represented on the following plot of profiles of the completers and dropouts in one of the treatment arms. For simplicity, focus in this chapter is on dropout after time 1, i.e. $t* = 1$.

In the model described in equation 6.6 the data are stratified into completers and dropouts only. The pattern-mixture model allows the data to be stratified by missing data pattern, for example subjects that drop out after period one are assigned a different stratum to those that drop out after periods two, three, four etc.. It may also be decided to stratify by treatment arm. This may result in very many strata, with few subjects per stratum, in which case it is recommended that data are stratified by groups of missing data pattern, or simply by completers versus dropouts [28]. In the asthma data, stratifying by treatment arm and missing data pattern would result in some strata with

Figure 6.1: Mean response profile of the completers compared to the dropouts, with dropout after time 1



as few as 2 or 3. It would only be sensible to stratify by missing data pattern and / or by treatment arm if the clinicians providing the priors on the dropout parameters believed the parameters were different across dropout times and / or treatment arms.

6.5.2 Extension to multiple causes of dropout

To incorporate cause of dropout into the model, subjects are stratified by cause of dropout, as well as stratifying by missing data pattern. The simplest model would have three strata: the completers (0), subjects that dropped out for reason one (1), and a third stratum of subjects that dropped out for a different reason (2). This scenario would be modelled as follows:

$$(0)\ Y_{ij} = \alpha_i + \beta_1 \text{treat}_\text{i} + \beta_2 t_{ij} + \epsilon_{ij} \tag{6.7}$$

$$(1)\ Y_{ij} = \alpha_i + \beta_1 \text{treat}_\text{i} + \beta_2 t_{ij} + \delta_{11}(1 - g_{ij}) + \delta_{21} g_{ij} + \epsilon_{ij}$$

$$(2)\quad Y_{ij} = \alpha_i + \beta_1 \text{treat}_\text{i} + \beta_2 t_{ij} + \delta_{12}(1 - g_{ij}) + \delta_{22} g_{ij} + \epsilon_{ij}$$

$$\alpha_i \sim N(\mu_\alpha, \sigma_u^2)$$

$$\epsilon_{ij} \sim N(0, \sigma_e^2)$$

Prior distributions must be elicited for $\boldsymbol{\delta_2}=(\delta_{21}, \delta_{22})^T$. These parameters, δ_{21} and δ_{22}, are the mean difference in response between completers and dropouts, after dropout, for each cause. The dropout parameters in the vector, $\boldsymbol{\delta_1} = (\delta_{11}, \delta_{12})^T$, are estimated from the data. This model is represented by a plot of the mean response profiles of subjects in each stratum, in one of the treatment arms, in figure 6.2. It is assumed in this case that dropout bias is the same in both treatment arms.

The overall treatment effect, β_{treat}, is a weighted average of the treatment effects in each stratum, averaged over all time periods. For example, consider the case where subjects that complete the study have four measurements, and all dropouts leave the study after one observation. If C denotes the control group and T the treatment group, π_{C0} and π_{T0} are the proportions of completers in the control group and treatment group respectively.π_{C1} and π_{C2} are the proportions of subjects that drop out for causes 1 and 2 in the control group. π_{T1} and π_{T2} are the equivalent proportions in the treatment group. This information is summarised in table 6.1 below.

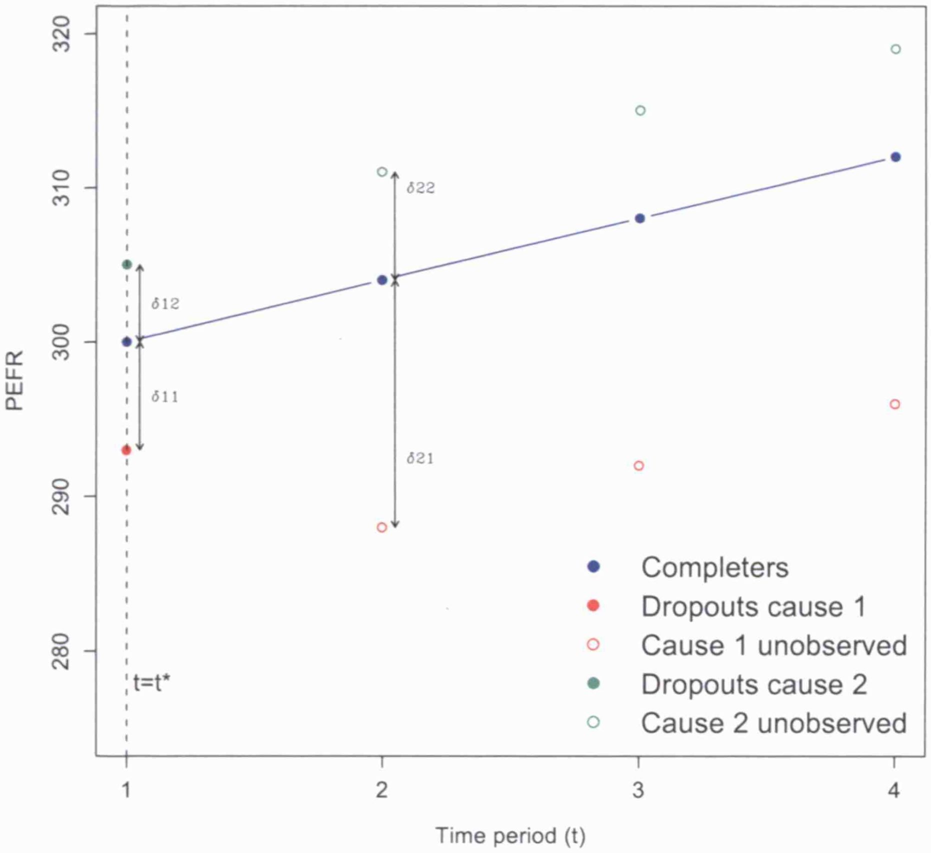
Table 6.1: Notation for the proportions of completers and dropouts for each cause, in each treatment arm, and the expected value of the response in each stratum

		Control group		**Treatment group**	
Stratum		Proportion	$E(Y)$	Proportion	$E(Y)$
XXXX		π_{C0}	μ_α	π_{T0}	$\mu_\alpha + \beta_1$
X . . .	Cause 1	π_{C1}	$\mu_\alpha + \delta_{11} + \frac{3}{4}\delta_{21}$	π_{T1}	$\mu_\alpha + \beta_1 + \delta_{11} + \frac{3}{4}\delta_{21}$
X . . .	Cause 2	π_{C2}	$\mu_\alpha + \delta_{12} + \frac{3}{4}\delta_{22}$	π_{T2}	$\mu_\alpha + \beta_1 + \delta_{12} + \frac{3}{4}\delta_{22}$
Sum		1	-	1	-

Using the summaries in table 6.1, the combined treatment effect is as follows:

$$\begin{aligned}
\beta_{treat} &= E(\overline{y_{T..}} - \overline{y_{C..}}) \\
&= (\pi_{T0}\overline{y_{T0}} + \pi_{T1}\overline{y_{T1}} + \pi_{T2}\overline{y_{T2}}) - (\pi_{C0}\overline{y_{C0}} + \pi_{C1}\overline{y_{C1}} + \pi_{C2}\overline{y_{C2}}) \\
&= [\pi_{T0}(\mu_\alpha + \beta_1) + \pi_{T1}(\mu_\alpha + \beta_1 + \delta_{11} + \tfrac{3}{4}\delta_{21}) + \pi_{T2}(\mu_\alpha + \beta_1 + \delta_{12} + \tfrac{3}{4}\delta_{22})] \\
&\quad -[\pi_{C0}(\mu_\alpha) + \pi_{C1}(\mu_\alpha + \delta_{11} + \tfrac{3}{4}\delta_{21}) + \pi_{C2}(\mu_\alpha + \delta_{12} + \tfrac{3}{4}\delta_{22})] \\
&= \beta_1 + \pi_{T1}(\delta_{11} + \tfrac{3}{4}\delta_{21}) + \pi_{T2}(\delta_{12} + \tfrac{3}{4}\delta_{22}) - \pi_{C1}(\delta_{11} + \tfrac{3}{4}\delta_{21}) - \pi_{C2}(\delta_{12} + \tfrac{3}{4}\delta_{22}) \\
&= \beta_1 + (\delta_{11} + \tfrac{3}{4}\delta_{21})(\pi_{T1} - \pi_{C1}) + (\delta_{12} + \tfrac{3}{4}\delta_{22})(\pi_{T2} - \pi_{C2})
\end{aligned} \qquad (6.8)$$

Figure 6.2: Mean response profile of the completers compared to the patients with each cause of dropout, with dropout after time 1.



6.6 Prior elicitation

To find out how feasible it would be to elicit priors on dropout parameters in a repeated measurements study such as the asthma clinical trial, it was decided to consult a thoracic clinician who is involved in asthma studies. A HIV clinician was also consulted, to investigate the feasibility of the approach in repeated measurements studies more generally. Because of the chronic nature of the disease, studies into HIV tend to result in the collection of repeated measurements.

The clinicians were asked about likely causes of dropout in studies in their respective fields of medicine, how confident they thought they would be in estimating bias due to dropout, and what information they thought would help them to make their estimation.

Discussion with the clinicians was based around the questions below. In answering the questions the thoracic clinician was asked to consider an asthma study in which Peak Expiratory Flow Rate (PEFR), is measured repeatedly over time. The HIV clinician was asked to consider a HIV study where the response is repeated measurements of CD4 count, a measure of how the immune system is functioning. Both clinicians were shown the type of questionnaire that would be used for prior elicitation, given in appendix B.

1. What are the main causes of dropout in asthma / HIV studies and clinical trials?

2. Do you believe the bias in the response of the dropouts is different for different causes of dropout?

3. Would knowing the cause of dropout help you to predict the response of the dropouts after they have dropped out?

4. Would knowing the mean and standard deviation of the completers help you to predict the response of the dropouts?

5. Would knowing the bias in the response of the dropouts before they drop out (δ_1) help you to predict the bias in their response after they drop out (δ_2)?

6. What other information might help you predict the response of the dropouts?

7. Consider one cause of dropout. Do you believe these dropouts follow a different slope to the completers after they drop out? If so, would you be able to predict what this slope would be?

8. Do you have any other comments about estimating the response of dropouts by cause of dropout?

Both clinicians said that knowing the cause of dropout would help them to estimate the dropout bias. The thoracic clinician explained that there are two main causes of dropout in an asthma study. The most common cause is that the patient finds participating in the study too much of an inconvenience and drops out, and the second most common cause is that the patient suffers an acute exacerbation of symptoms and either visits their GP or is admitted to hospital. This statement is compatible with the findings of the asthma clinical trial, in which the main cause of dropout was exacerbation of symptoms, with the second most common cause being unknown. He suggested that a third, less common cause of patient withdrawal would be that a patient's symptoms were so stable that they lost motivation for continuing in the study and withdrew them-self. He was confident that these different causes would lead to very different dropout biases, often in opposite directions, and would be confident in estimating what these biases were likely to be.

The HIV clinician said that knowing the cause of dropout would usually indicate whether or not the patient had continued on treatment. For example, common causes of dropout in HIV studies are death, change of therapy due to tolerability problems, failure to take treatment, or removal of consent. He was confident that knowing whether or not the patient had continued with their treatment would tell him the direction of the dropout bias, and would allow him to predict the size of the bias reasonably accurately.

The consultants were also asked whether they would be able to estimate the correlation of $\boldsymbol{\delta_2}$ with other parameters in the model. Of particular interest was the correlation between $\boldsymbol{\delta_2}$ and $\boldsymbol{\delta_1}$, as it seemed possible that the bias after dropout would depend on the bias before dropout. The HIV clinician told us that knowing the values of $\boldsymbol{\delta_1}$ would not help him to estimate the values of $\boldsymbol{\delta_2}$, once he knew the cause of dropout. The same was true for the other parameters in the model. The thoracic clinician agreed that he would be unable to estimate any correlation between $\boldsymbol{\delta_2}$ once he knew the mean of the response up until dropout and the length of time the subjects remained in the study.

Both clinicians said that the time of dropout would affect their estimate of dropout bias. This means that the model needs to allow different prior estimates of the dropout parameters at each time-point. The thoracic clinician explained that time of dropout was important because earlier dropouts would be likely to have had poor response values. This implied that the response before dropout would be more useful to the clinician than the length of time in the study before dropout. He added that the more observations recorded before dropout, the easier it would be to estimate the dropout bias because the stability of the response before dropout would provide information about the dropout bias.

The clinicians were asked if they would be able to estimate the slope of patients who drop out after dropout. Both clinicians said that although they thought it was plausible that the slopes of the dropouts would change after dropout, neither of them felt able to offer estimates of what these slopes would be.

During discussion with the HIV clinician, it became clear that for HIV studies it would be more sensible to elicit the mean response after dropout rather than the dropout bias. This conclusion was made because the clinician explained that the CD4 count of patients who discontinue their treatment tends to drop towards their baseline response and then stay at approximately that level. This happens at different rates for different patients. It seems it would be sensible to elicit parameters for the mean response after dropout compared to the baseline response, and parameters for the time taken for the response to reach this level. This highlights the need to discuss the likely dropout model with clinicians before constructing the model and carrying out any prior elicitation.

Garthwaite [59] has written a thorough guide on prior elicitation, drawing on findings from a large number of experiments in elicitation. Several experiments have demonstrated that experts are poor at estimating measures of spread compared to measures of location. Research has shown that eliciting credible intervals, whereby the expert is asked to assign weights to a series of possible intervals for a parameter, is more reliable than attempting to directly elicit a variance or standard deviation. But even with credible interval elicitation, subjects tend to underestimate the width of intervals. It has been found that visual aids can help with prior elicitation.

6.7 A simulation study

Talking to the clinicians, the approach seems feasible, and prior elicitation of the dropout parameters promises to provide useful information, with the potential to reduce bias in the model estimates. Equation 6.7 seems to be a sensible model for the asthma clinical trial scenario. To investigate how well the model works under various scenarios, with correctly and incorrectly specified priors for the dropout parameters, it was fitted to simulated incomplete data, based on the asthma clinical trial. Estimates of the treatment effect could then be compared to estimates from the full data, had all data been collected, and estimates from an available case analysis, ignoring the missing data.

Table 6.2: Population parameter values for the simulated data

Parameter		Value
β_1	Treatment effect of completers	20
β_2	Time effect	2
μ_α	Mean intercept of completers	300
σ_u^2	Between-subject variance	500
σ_e^2	Within-subject variance	500
δ_{11}	Cause 1 bias before dropout	-10
δ_{12}	Cause 2 bias before dropout	0
δ_{21}	Cause 1 bias after dropout	-40
δ_{22}	Cause 2 bias after dropout	0

6.7.1 Data simulation

For simplicity, only two treatment arms were simulated, analogous to the placebo group and the salmeterol xinafoate (SM) group in the asthma clinical trial. In the original data there were 153 patients, split equally between these two groups. Data were simulated to have approximately the same number of subjects, with parameter values close to the parameter estimates of the asthma data, and a similar proportion of subjects dropping out. Two causes of dropout were simulated: Cause one was that subjects suffered an acute exacerbation of symptoms, and cause two was that subjects had stable symptoms. According to the thoracic clinician, these are the two main causes of dropout in asthma studies. For each cause of dropout, the data were simulated to have a different bias before and after dropout. The model used to simulate the data is that given in equation 6.7, with the parameter values given in table 6.2. Here, patients drop out after the first time period, but the model generalises to several dropout times.

Subjects that dropped out because they had an acute exacerbation of symptoms (cause 1) had a reduced response before dropout compared to the completers, and then a further drop in response after dropout. Subjects who dropped out because their asthma was stable (cause 2) were simulated to have the same mean response as the completers, both before and after dropout. The number of subjects in each dataset, N, is 150, and the proportion of dropouts from each treatment arm and for each cause of dropout were as follows:

Table 6.3: Proportions of dropouts by cause of dropout and treatment arm

Parameter		**Value**
π_{T1}	Proportion dropouts - cause 1, treatment arm	0.16
π_{T2}	Proportion dropouts - cause 2, treatment arm	0.05
π_{C1}	Proportion dropouts - cause 1, placebo arm	0.32
π_{C2}	Proportion dropouts - cause 2, placebo arm	0.11

6.7.2 Model fitting

The dropout parameters, $\boldsymbol{\delta_2}$, were given Normal prior distributions. Model-fitting was implemented in WinBUGS using MCMC simulation. Bayesian inference treats missing values as parameters that have a joint posterior distribution with the model parameters, conditional on the observed data [60], and initial values are therefore required for all missing observations.

It is important to check for convergence of the iterative simulations and to discard the iterations prior to convergence. The recommended approach to checking for convergence [60] is to run more than one chain of iterations with overly-dispersed initial values and ensure that the chains converge for every parameter of interest in the model. This can be done by eye, from plots of the chains of iterations. In addition, a convergence parameter, $\hat{R}$ is calculated in WinBUGS for each parameter, which tends towards 1 as the simulations converge. This convergence parameter was monitored for all model parameters to ensure convergence had been reached. A large number of iterations were needed to reach convergence, meaning that model fitting was slow; model fitting on 1000 datasets took about 12 hours. Sensible initial values were chosen, as unsuitable initial values can affect the rate of convergence [61].

Once early simulations were discarded, sufficient simulations were obtained to accurately estimate the posterior distribution of each parameter of interest. It was ensured that the Monte Carlo error, an estimate of the difference between the mean of the sampled values and the true posterior mean, was less than 5% of the sample standard deviation for each model parameter.

The multiple causes model with strong priors on the dropout parameters was compared to the same model using weak priors. Strong dropout priors had a prior variance of 1000, resulting in a prior standard deviation of approximately 5 times the standard error of the treatment effect. In contrast, the weak prior distributions were given a prior variance of 1×10^4, resulting in a prior standard deviation of approximately 20 times the standard error of the treatment effect. A model with priors centred on the true values of $\boldsymbol{\delta_2}$ was compared to a model where the priors were given the opposite direction but the same size as the true values, i.e. $-\delta_{21}$ and $-\delta_{22}$.

The multiple causes model, in equation 6.7, fitted to the incomplete data, was compared to an available case analysis of the incomplete data, and a multiple causes model fitted to the complete data. The available case analysis assumes that the missing observations are MAR, and estimates the treatment effect to be $\hat{\beta}_1$ in equation 6.9, ignoring the dropout biases both before and after dropout.

$$Y_{ij} = \alpha_i + \beta_1 \text{treat}_\text{i} + \beta_2 t_{ij} + \epsilon_{ij} \tag{6.9}$$

Under this naive model, a straight line is fitted to each treatment arm, with random intercept. This model is referred to as the naive model as it is much simpler than the multiple causes model that the data were simulated from.

The multiple causes model fitted to the full data provides the inference we would have obtained had we been able to observe all subjects through to the end of the study, and is therefore considered the "gold standard".

As a comparison to our multiple causes model, a pattern-mixture model is fitted which assumes the missing observations are MAR. Dropouts and completers are allowed to be centred about different means, which are estimated directly from the data. This assumes that there is only one cause of dropout, or that all dropouts come from the same distribution. The model fitted is:

$$Y_{ij} = \alpha_i + \beta_1 \text{treat}_\text{i} + \beta_2 t_{ij} + \beta_3 \text{drop}_\text{i} + \epsilon_{ij} \tag{6.10}$$

where drop_i indicates whether or not the subject dropped out. The estimate of the treatment effect in model 6.10 is β_1.

6.7.3 Results

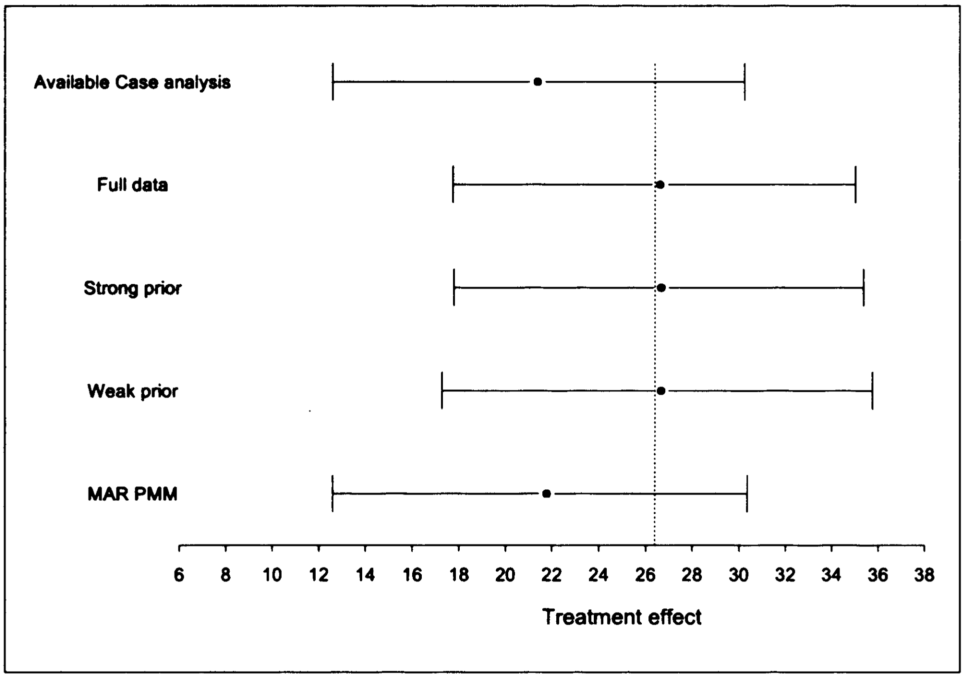
One thousand datasets were simulated for each scenario. Unless otherwise stated, the sample size is 150 cases, and the ICC is 0.5. The scenarios investigated were:

1. Correctly centred priors for dropout parameters
2. Incorrectly centred priors for dropout parameters
3. Reduced correlation within clusters
4. Reduced sample size
5. Introduce random dropout parameters
6. Examine model under virtually flat priors

1. Correct prior means for $\boldsymbol{\delta_2}$

The dotted vertical line in figure 6.3 represents the true population treatment effect, computed using the values in tables 6.2 and 6.3 in equation 6.8. As expected, when the prior distributions for both elements of $\boldsymbol{\delta_2}$ are centred about the true values of the simulated elements of $\boldsymbol{\delta_2}$, the bias is fully recovered by the model. This is true regardless of the strength of the prior

Figure 6.3: **Scenario 1**: Correct prior means for $\boldsymbol{\delta_2}$



distributions. The effect of assigning weak priors to the dropout parameters is that the standard error of the treatment effect is inflated.

The strength of the prior distributions does not affect the point estimate of the treatment effect, it is only the mean of the prior that has an effect. This is because the data provide no information about the mean bias of the dropouts after they have dropped out, $\boldsymbol{\delta_2}$. The posterior distributions of the elements of $\boldsymbol{\delta_2}$, therefore, are the same as the prior distributions, and the contribution to the treatment effect of these parameters, in equation 6.8, is the same regardless of the strength of their priors.

The multiple causes model fitted to the full data results in the narrowest of all credible intervals, as we would expect. There is more information available when fitting the model to the full data.

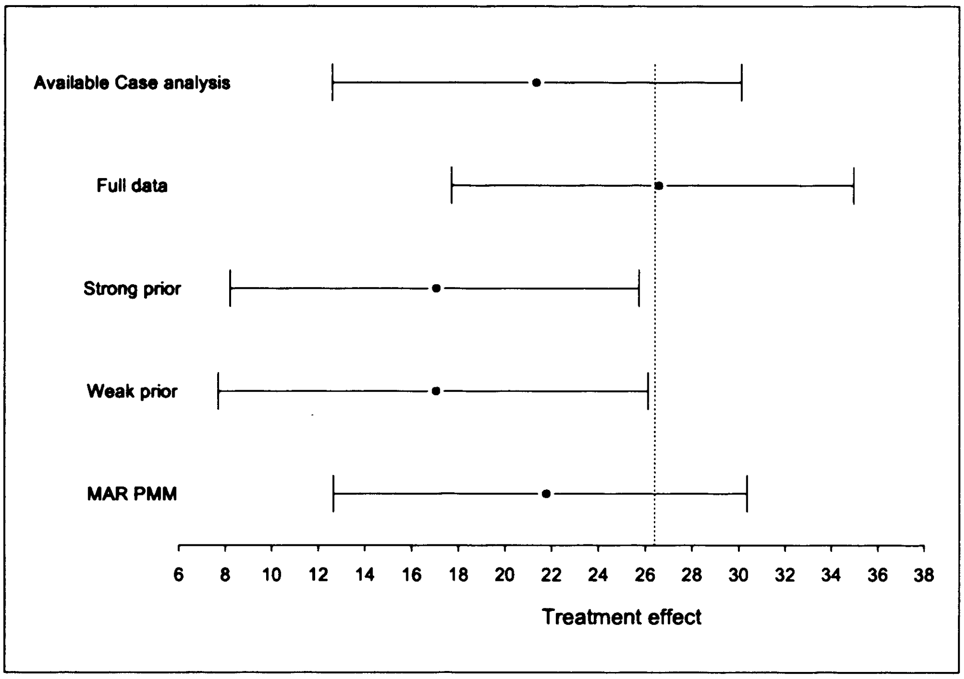
The credible interval from the available case analysis is narrower than those from the multiple causes model, even with a strong prior. This is because the multiple causes model incorporates additional uncertainty in the distribution of the missing data. Note that the available case analysis fits a different model to the model from which the data are simulated; the data actually come from several distributions and this naive model fits only one.

The MAR pattern-mixture model estimates a treatment effect similar to the available case analysis, with a similar standard error. Both models make a MAR assumption. The MAR pattern-mixture model is a different model again, to the model from which the data are simulated, modelling two parallel slopes in each treatment arm, one for the dropouts and one for the completers.

2. Incorrect prior means for $\boldsymbol{\delta_2}$

Figure 6.4 shows the results of the model when the priors are incorrectly specified. The prior distributions for $\boldsymbol{\delta_2}$ were centred about means with opposite signs to the true values, i.e. $-\delta_{21}$ and $-\delta_{22}$. This demonstrates what happens to the treatment effect if the clinician is wrong about the direction of the bias in the dropouts.

Figure 6.4: **Scenario 2**: Incorrect prior means for $\boldsymbol{\delta_2}$

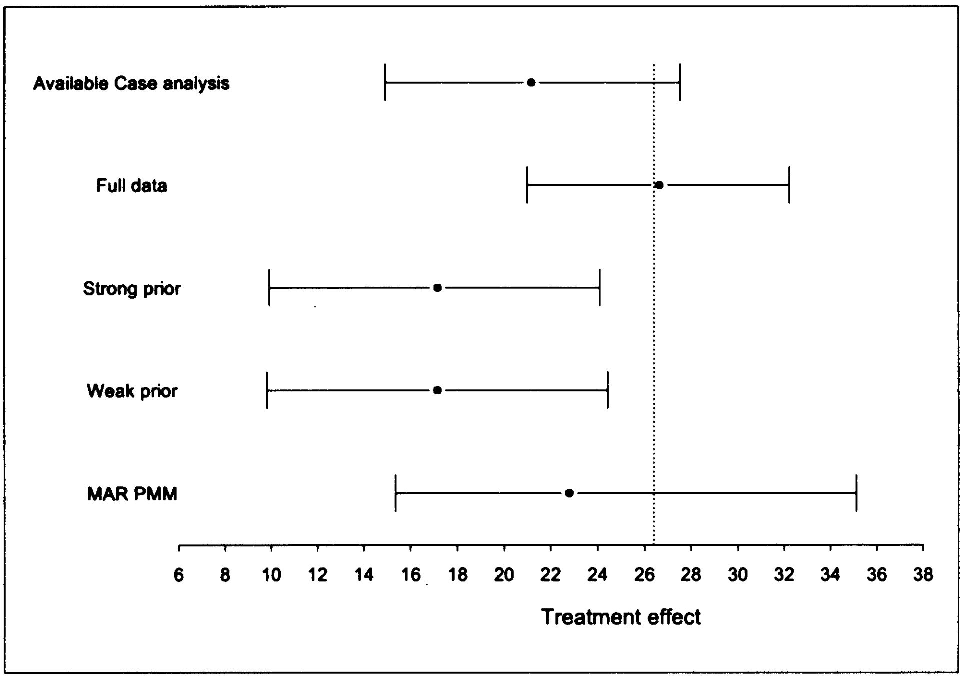


If clinicians' knowledge about the response of dropouts after they drop out is unreliable, the model is a dangerous tool to adopt, especially if clinicians underestimate their own uncertainty about the dropout parameters. Even when a weak prior distribution is assigned to each dropout parameter, the point estimate of the treatment effect is biased, and in this case, its credible interval does not contain the true population value for the treatment effect.

3. Reduced correlation within clusters, ICC=0.1

As in scenario two, the priors for $\boldsymbol{\delta_2}$ are incorrectly centred with means $-\delta_{21}$ and $-\delta_{22}$. Note that the estimates from the multiple causes model would again be unbiased if the priors for $\boldsymbol{\delta_2}$ were correctly specified.

Figure 6.5: **Scenario 3**: Reduced correlation within clusters, ICC=0.1

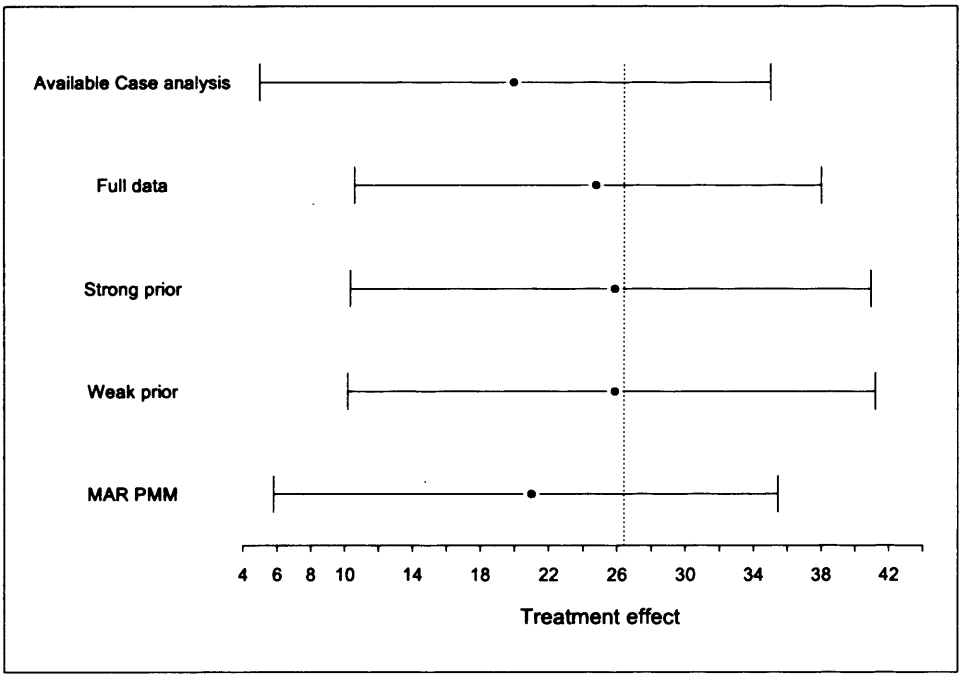


The missing data cause greater bias, compared to the standard error of the estimate, when the correlation within clusters is lower. This is because each observation that is removed, being less similar to other observations within the cluster, carries a greater amount of information. Note that the credible intervals are, in general, narrower for all models when the ICC is low, because of the greater amount of information carried by each observation. The exception is the MAR pattern-mixture model, which has inflated credible intervals compared to the same model fitted to data with an ICC of 0.5. Note, however, that this pattern-mixture model fits parallel straight lines to the data, one for each treatment arm for each cause of dropout, and that this is the wrong model for the data.

4. Reduced sample size, N=50

In all of the following scenarios the correct prior means are specified for $\boldsymbol{\delta_2}$.

Figure 6.6: **Scenario 4**: Reduced sample size, N=50

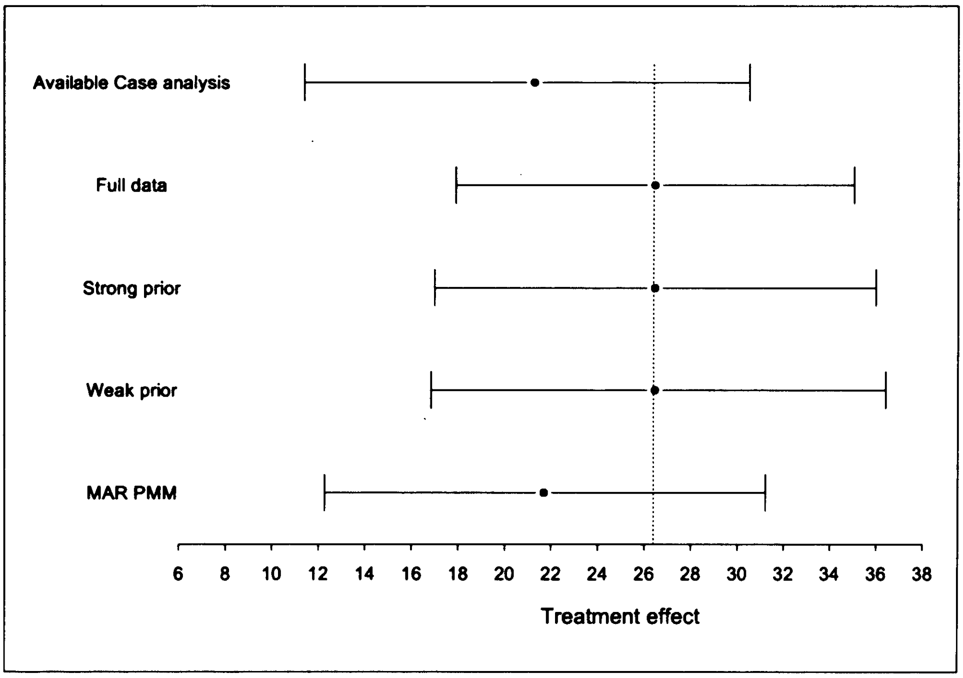


Note the wider confidence intervals because of the reduced sample size. Because of this, the bias is small in comparison to the standard error of the parameter. Once again, the informative priors for the dropout parameters recover the bias caused by the missing data.

5. Random δ_1 and δ_2

For each cause, the bias in the dropouts, both before and after dropout, is simulated to be normally distributed. See figure 6.7 for the results.

Figure 6.7: **Scenario 5**: Random $\boldsymbol{\delta_1}$ **and** $\boldsymbol{\delta_2}$

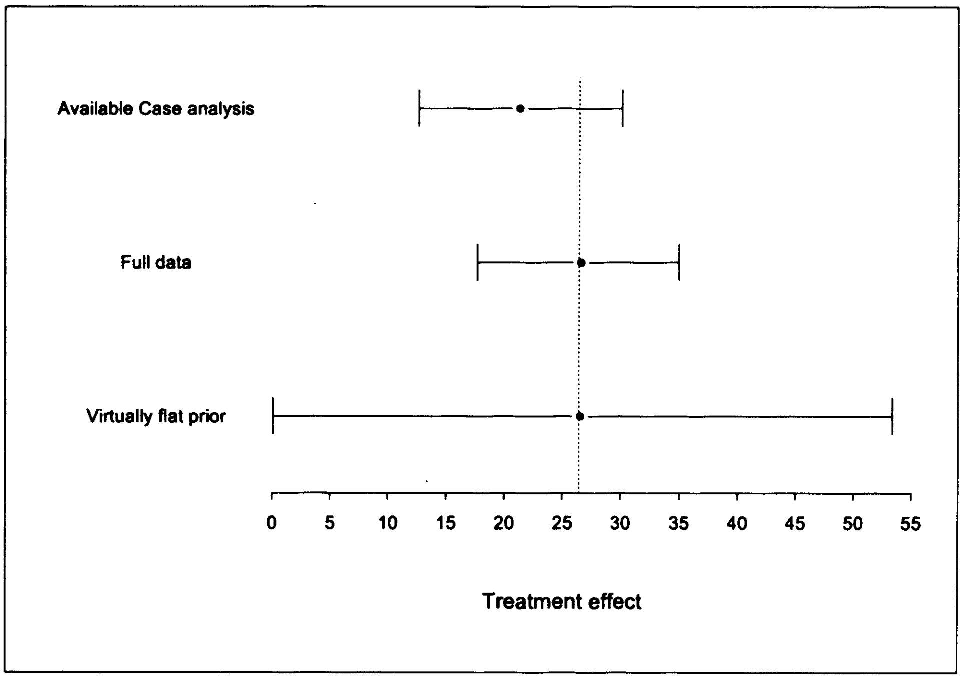


When the data are simulated to have random dropout parameters, the model with correctly centred priors for the dropout parameters, recovers the bias caused by the missing data. The credible intervals are slightly wider for all models, because of the additional uncertainty in the data, introduced by the dropout parameters.

6. Flat priors

In figure 6.8 the variances of the prior distributions tend towards infinity.

Figure 6.8: **Scenario 6**: Flat priors



The prior variances of each of the dropout parameters were set to 1×10^6. As the priors tend towards being uninformative the posterior variance of the treatment effect is inflated substantially. We see very clearly here that the point estimate of the treatment effect is corrected by the same amount regardless of the strength of the dropout priors.

6.7.4 Summary of findings

Correct priors for the dropout parameters recover all the bias due to dropout, as expected. The strength of the priors affects the standard error of the treatment parameter, but not the point estimate. With the strongest priors used here, there is more uncertainty in the parameter estimates than in the correct model fitted to the complete data. There is also more uncertainty than in the available case analysis, because the multiple causes model takes into account the uncertainty in the distribution of the missing data. As the priors tend towards being uninformative, the posterior treatment effect tends towards a distribution with infinite variance. The results of the multiple causes model with incorrect prior means for $\boldsymbol{\delta_2}$ demonstrates the potential danger of the model if clinicians give poor estimates of the dropout parameters and overestimate their certainty about the parameters. The model has been shown to cope with reduced ICC, reduced sample size and random dropout parameters.

6.8 Discussion of findings

6.8.1 Prior elicitation

The prior precision of $\boldsymbol{\delta_2}$ affects the posterior precision of the treatment effect but not the posterior mean. This is because the data provide no information about δ_{21} and δ_{22}, therefore the posterior distributions of δ_{21} and δ_{22} remain centred on the prior means of these two parameters, regardless of the prior variances they are given. When clinicians have little information to offer about the missing data, very weak priors for the dropout parameters result in very wide credible intervals for the treatment effect. However unappealing these very wide credible intervals may be, they provide a more honest representation of the true uncertainty about the model parameters than an available case analysis provides.

The prior distributions for $\boldsymbol{\delta_2}$ would have to be very narrow for the multiple causes model to result in a narrower credible interval for the treatment effect than the estimate from the full data. This would correspond to the clinicians being so certain about the missing values of the response that more could be learnt from the incomplete data together with the priors elicited from the clinicians, than from the data had all subjects been observed until the end of the study. This is implausible, and would suggest that the clinicians were under-estimating their uncertainty in the dropout parameters.

Note that the available case analysis fits a different model to the other approaches, and they are therefore not directly comparable. The multiple causes model does not, therefore, reduce to the available case analysis in the special case when the dropout parameters are centred on zero with zero variance. In this special case, neither would the model reduce to the MAR pattern-mixture model specified here, as the MAR pattern-mixture model fits two parallel slopes in each treatment arm, one each to the completers and dropouts. The multiple causes model with dropout parameters are centred on zero with zero variance would be equivalent to a pattern-mixture model stratified by cause of dropout.

A limitation of the scenarios simulated is that the subjects have the same slope after dropout. This restriction would not necessarily hold in practice. Further investigation would be needed to assess the robustness of the model to changes in slope after dropout. Although the clinicians both said that they would be unable to offer an estimate of the slope of the dropouts after they drop out, it seems likely that they would be able to provide some information about what this slope might be, even if it is a very weak prior. The model should be extended to include a parameter for the slope after dropout.

6.8.2 Generalisability of the results

The results are specific to the particular data that were simulated. Attempts were made to make the simulations as realistic as possible but it is not possible to simulate data to represent every possible scenario. As discussed in section 6.8.1, it is likely that dropouts would have a different slope after they drop out of the study. It would be useful to test the robustness of the model to a change in slope after dropout, firstly for a model that ignores the change in slope, and secondly, for a model with a weak prior on the change in slope.

So far the model has only been used to estimate effects of cluster-level covariates such as a treatment effect. In theory, the model will extend to covariates that vary within the cluster, but obtaining a weighted average of the parameter estimates across causes of dropout would be more complicated.

The model fitted here assumes that, within cause of dropout, dropout bias is the same in each treatment arm. This restriction could be removed, but careful discussion with the clinicians would be needed to decide if this was sensible. The issue is not whether dropout bias is different between the treatment arms, but whether, conditional on cause of dropout, there is a difference. One possible argument in favour of stratifying further by treatment arm is, in the case of exacerbation of symptoms, patients may deteriorate to the same level of PEFR, regardless of treatment group. As it stands, the model has the same treatment effect before and after dropout, and this may be unrealistic.

The model assumes that the dropout parameters are uncorrelated with other parameters in the model. This assumption will not always hold. Although ideally attempts should be made to elicit the correlation between the dropout parameters and other model parameters, the elicitation process proposed is already quite complex, with parameters of bias for different causes of dropout being elicited, as well as parameters of the slope after dropout. White et. al.[30] suggest a method for eliciting correlation parameters by eliciting priors conditional on several values of another model parameter. Prior elicitation for longitudinal data with multiple causes of dropout is already far from straightforward, and adding elicitation of parameters of correlation between various parameters would make the prior elicitation process too complicated.

The multiple causes model, here, is designed for the analysis of an asthma clinical trial. In section 6.6 it emerged from discussion with the HIV clinician, that it would be better to elicit a distribution for the response after dropout, for each cause of dropout, rather than a bias in the response after dropout.

This illustrates that the model must be tailored to the type of data being analysed.

6.8.3 Potential of the multiple causes model as a tool for MNAR data

It is extremely encouraging that clinicians consider the cause of patient dropout to be the single most important factor in estimating the dropout bias. Provided with this information, both clinicians consulted felt very confident about their abilities to predict the response bias in patients that drop out.

It has been assumed that the data on cause of dropout is reliable. In fact, this will not always be the case. It is standard practice to report reasons for patient dropout in clinical trials, and this data may even be required by the protocol. However, it is possible that the quality of this data will not be high, as it can be difficult to obtain information on patients who have dropped out of a study.

The two clinicians interviewed for this study were comfortable with the idea of prior elicitation and were able to provide opinions on the distributions of the dropout parameters using the questionnaire given in appendix B. It is hoped that such a questionnaire, filled out by the clinician in the presence of a statistician, would be a reliable method of prior elicitation. This is something that needs to be further investigated. If there are problems in using this style of questionnaire, in which clinicians are asked to assign weights to possible values of the dropout parameters, time needs to be spent on improving the prior elicitation process.

The causes of dropout that the clincians were asked to consider were simple, clear and unambiguous. In the case of the thoracic clinician, the first cause under consideration was dropout due to an acute exacerbation of symptoms, and the other, dropout due to stable symptoms. In practice, more complex causes of dropout will be recorded. Prior elicitation would then be more difficult.

In the model as it is specified here, data with a large number of missing data patterns and several treatment groups would be stratified into very many small strata. This would particularly be a problem in data with intermittent missingness and studies with a large number of follow-up times. Under these circumstances it is recommended that the strata are limited to completers versus dropouts by each treatment group. This does, however, require the somewhat restrictive assumption that the bias is independent of the time and number of missing observations on a subject.

An alternative use for the approach is as a "reverse Bayesian analysis" tool that would determine which prior distribution for the dropout parameters would mean the parameter of interest was no longer statistically significant. This would be an alternative method of carrying out a sensitivity analysis. A disadvantage of the method is that it would not produce point estimates and credible intervals for the parameters of interest that incorporated the uncertainty due to patient dropout.

The most important message to be learned from the simulation study, however, is that the approach should only be recommended with extreme caution. Prior elicitation should be carried out in such a way that the clinicians' true uncertainty about the dropout bias is reflected in the prior distributions for the dropout parameters. The approach should only be adopted in situations where the clinicians genuinely have knowledge about the likely dropout bias for each cause of dropout. The consequence of mis-specifying the priors in this simulation study was demonstrated with an extreme case. It is inconceivable that a clinician who knew that the cause of dropout was exacerbation of symptoms, would get the direction of the bias wrong, rather than just its size. However, the potential is certainly there for a prior to be mis-specified, if not in the wrong direction, to such a degree that the parameter estimates are extremely biased. To give an idea of the level of inaccuracy in the mean priors that would lead to seriously biased results, assume that the clinician provides the correct direction of the bias. In this example, they would need to provide bias parameters that were at least twice as large as the true bias, in order to cause a treatment effect estimate that was worse than that obtained from the available case analysis.

6.9 Conclusions

This work highlights the great potential of a Bayesian model for MNAR missing data when the causes of dropout are known. When missing data are MNAR, and it is therefore necessary to look beyond the available data in carrying out the analysis, it is not sensible to ignore the causes of dropout. Interviews with clinicians indicate that there are experts available to provide information about the missing response, providing the causes of dropout are recorded, and this expertise should not be ignored.

Simulations demonstrate that the model successfully recovers the bias caused by the missing data, when the elicited priors are unbiased. This is the case under a range of scenarios. The model is only recommended with extreme caution, however. It is highly sensitive to mis-specification of the elicited dropout priors, because the data provide no information on these parameter values.

The work in this chapter provides the incentive to carry out further work on prior elicitation for missing data problems. Perhaps more importantly, though, it demonstrates the value of pursuing patients after they withdraw from a study, so that their cause of dropout can be recorded.

Chapter 7

Discussion

7.1 Sensitivity of existing models to dropout

The assumptions about the robustness of GEE models, random effects models and the summary statistics method to missing data, are not as categorical as is commonly stated in the literature.

As expected, no method for repeated measurements is statistically valid when data are missing not at random. The extent of the bias in the presence of MNAR dropout, with even a small proportion of missing data, is large enough to be of great concern.

The random effects model is robust to MAR data if the response is gaussian, but for binary data, may be biased if the event probability is small (0.2). That is if the model converges: there are convergence problems with the penalised quasi-likelihood method. An important finding of this research is that the quadrature method, although in general biased in the presence of MAR data, is vastly improved by increasing the number of quadrature points. In Stata, the default number of quadrature points is 12. Increasing this to the maximum of 30 produced unbiased estimates under MAR dropout in the scenarios where estimation with 12 quadrature points was significantly biased.

The sensitivity of the GEE model to missing data depends strongly on the accuracy of the working correlation matrix. In contrast to popular belief, the GEE model is not always robust to MAR data when the response is gaussian. Estimates from a correctly specified GEE are significantly but not substantially biased when dropout is MAR, and the bias is considerable when the wrong correlation structure is chosen. This bias is believed to be due to bias in the estimation of the correlation parameters. Further research is needed to investigate this further. Although a distinction is often made in the literature between gaussian and non-gaussian response data, the conclusions were the same for both gaussian and binary data in the scenarios investigated.

If a particular model is chosen in order to reduce bias caused by missing data, it should be remembered that marginal and cluster-specific parameters have a different interpretation, as explained in section 2.5.4.

The summary statistics method relies, as expected, on an assumption of MCAR data. With MAR data, the approach can be extremely biased, especially where covariates vary within the cluster.

Missing data should be of even more concern when the ICC is low because, assuming that the missing data are correlated with the observed data, the size of the bias caused by dropout is potentially higher.

7.2 Tests of MCAR

The importance of identifying the missing data mechanism is highlighted by the findings discussed above. Many tests of MCAR have been proposed over the last twenty years. Some of these procedures are unnecessarily complicated, for example involving several stages. In chapter 4, a qualitative comparison of the various tests assessed their benefits and limitations, in terms of their accessibility to the user, the types of data they can handle, whether general patterns of missing data can be analysed, or only monotone missing, and how the approaches model the relationship between dropout and any covariates.

The tests vary in their strategy of testing for MCAR, and as a result so do their power and significance level. A quantitative comparison was made between those tests that were deemed straightforward enough to be accessible to the user. For gaussian repeated measurements data with dropout, a test by Ridout is recommended as the most powerful test that has a reasonable significance level. This test is implemented on the asthma clinical trial data, together with visual inspection of the data, in chapter 5.

Tests proposed by Listing and by Park are appealing because they provide test statistics or parameter estimates that can be interpreted meaningfully. But, in the scenarios simulated, these tests proved to be either lacking in power or with a high type I error rate. In addition, Park's test statistic was often unobtainable because the covariance matrix of the parameter estimates was non-invertible. This was particularly a problem for smaller sample sizes.

The research highlights the need for a powerful test of MCAR with meaningful parameters or test statistic.

7.3 Model for data with multiple causes of dropout

The model for multiple causes of dropout proved to be potentially a very powerful tool for MNAR data. Clinicians were extremely supportive of the method, and believe the cause of dropout to be the most important factor in predicting the dropout bias.

The results of the simulation studies do, however, demonstrate the potential danger of the model if clinicians are to overestimate their certainty about the bias, and great care is needed to ensure the prior distributions represent the clinicians' true uncertainty about the dropout parameters.

The model has only been tested under limited scenarios. For example, it has been limited to covariates that are fixed at the cluster level, and modelling the effect of cluster-varying covariates has not been attempted. Also, the scenario where the slope of dropouts changes after they drop out has not been investigated. These are issues which need further exploration in order to advocate the model with greater confidence.

As well as a model for MNAR data, the approach could be used as a tool for sensitivity analysis. In this setting, the model would estimate the minimum size of dropout parameters that would cause the parameters of interest to no longer be significant.

7.4 Future Research

The model developed for MNAR data with multiple causes of dropout has should be fitted to real data to test the practicalities of the method, and highlight any unforseen issues in both the model fitting and prior elicitation processes. Death as a cause of dropout is a difficult scenario to model, and is an issue that it would be interesting to explore. So far the model has only been applied to gaussian data, and further work is needed to extend this to categorical and discrete data. The idea of the model as a reverse Bayesian analysis tool for sensitivity analysis is very appealing, and a possible area for future work.

As discussed in section 7.2, a test of MCAR was recommended for gaussian repeated measurements with dropout that is powerful and has an acceptable type I error rate. However, it would be useful to develop an accurate test of MCAR where the parameters or test statistic were meaningfully interpretable.

Simulation studies led to important findings about the impact of dropout on different models for repeated measurements data. The work focused more on bias than efficiency, and this is an area that could be explored further. Particularly interesting are the scenarios where loss of data increased model efficiency, because extreme values tended to go missing. There is a need for a reliable estimation method for logistic random effects models, as there can be convergence problems with the penalised quasi-likelihood approach. The quadrature method has come under criticism, but the results here that, with an increased number of quadrature points, the approach does have potential. The robustness of the logistic random effects model estimated using adaptive quadrature method should also be investigated.

Appendix A

Demonstration that GEE for Gaussian data are robust to MAR missing data

This is the proof of equation 3.19 to show that GEE for Gaussian data are asymptotically unbiased in the presence of MAR missing data, provided the correlation parameter is estimated without bias. In order to prove 3.19 we need to prove that the following holds:

$$\begin{pmatrix} Corr(Y_1) & Corr^T(Y_1,Y_2) \\ Corr(Y_1,Y_2) & Corr(Y_2) \end{pmatrix}^{-1} \begin{pmatrix} I_{n_1} \\ Corr(Y_1,Y_2)Corr(Y_1)^{-1} \end{pmatrix} = \begin{pmatrix} Corr^{-1}(Y_1) \\ 0_{n_2,n_1} \end{pmatrix} \tag{A.1}$$

For simplicity we introduce the following notation

$$Corr(Y_1) = A$$
$$Corr(Y_1,Y_2) = B$$
$$Corr(Y_2) = C$$

and equation A.1 becomes:

$$\begin{pmatrix} A & B^T \\ B & C \end{pmatrix}^{-1} \begin{pmatrix} I_{n_1} \\ BA^{-1} \end{pmatrix} = \begin{pmatrix} A^{-1} \\ 0_{n_2,n_1} \end{pmatrix} \tag{A.2}$$

Firstly, to find the inverse of $\begin{pmatrix} A & B^T \\ B & C \end{pmatrix}$, the matrix $\begin{pmatrix} D & E^T \\ E & F \end{pmatrix}$ is defined as:

$$\begin{pmatrix} D & E^T \\ E & F \end{pmatrix} \begin{pmatrix} A & B^T \\ B & C \end{pmatrix} = \begin{pmatrix} I_{n_1} & 0_{n_1,n_2} \\ 0_{n_2,n_1} & I_{n_2} \end{pmatrix} \tag{A.3}$$

and from this definition, we obtain the following two equations:

$$DA + E^T B = I_{n_1} \tag{A.4}$$

$$EA + FB = 0_{n_2,n_1} \tag{A.5}$$

Using $\begin{pmatrix} D & E^T \\ E & F \end{pmatrix}$ as the inverse of $\begin{pmatrix} A & B^T \\ B & C \end{pmatrix}$, the left hand side of equation A.1 becomes:

$$\begin{pmatrix} A & B^T \\ B & C \end{pmatrix}^{-1} \begin{pmatrix} I_{n_1} \\ BA^{-1} \end{pmatrix} \tag{A.6}$$

which is equal to:

$$\begin{aligned} \begin{pmatrix} A & B^T \\ B & C \end{pmatrix}^{-1} \begin{pmatrix} I_{n_1} \\ BA^{-1} \end{pmatrix} &= \begin{pmatrix} D + E^T BA^{-1} \\ E + FBA^{-1} \end{pmatrix} && \text{(A.7)} \\ &= \begin{pmatrix} (DA + E^T B)A^{-1} \\ (EA + FB)A^{-1} \end{pmatrix} && \text{(A.8)} \\ &= \begin{pmatrix} A^{-1} \\ 0_{n_2,n_1} \end{pmatrix} && \text{(A.9)} \end{aligned}$$

where A.9 follows from equations A.4 and A.5.

Back in the original notation this proves A.1.

Appendix B

Example prior elicitation questionnaire

Consider first, subjects who drop out because they suffer an acute exacerbation of symptoms. Please assign weights according to your strength of belief about their mean response compared to the mean of the completers (δ_{21} on the plot). Your weights should sum to 100.

Lower than completers by				Same as completers	Higher than completers by				
>60%	40-60%	20-40%	0-20%	–	0-20%	20-50%	40-60%	>60%	TOTAL

Now consider subjects who drop out because they report having stable symptoms. Again, assign weights according to your strength of belief about their mean response compared to the mean of the completers (δ_{22} on the plot). Your weights should sum to 100.

Lower than completers by				Same as completers	Higher than completers by				
>60%	40-60%	20-40%	0-20%	–	0-20%	20-50%	40-60%	>60%	TOTAL

Bibliography

[1] Wood, A. M., White, I. R. and Thompson, S. G. (2004). Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials*, **1**, 368-376.

[2] Park, T. (1993). A comparison of the generalized estimating equation approach with the maximum likelihood approach for repeated measurements. *Statistics in Medicine*, **12**, 1723-1732.

[3] Touloumi, G., Babiker, A. G., Pocock, S. J. and 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**, 3715-3728.

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

[5] Feng, Z., McLerran, D. and Grizzle, J. (1996). A comparison of statistical methods for clustered data analysis with gaussian error. *Statistics in Medicine*, **15**, 1793-1806.

[6] Burton, P., Gurrin, L. and Sly, P. (1998). Extending the simple linear regression model to account for correlated responses: An introduction to generalized estimating equations and multi-level mixed modelling. *Statistics in Medicine*, **17**, 1261-1291.

[7] Goldstein, H. (1995). *Multilevel statistical models*, 2nd Ed. John Wiley and Sons, New York.

[8] Albert, P. S. (1999). Longitudinal Data Analysis (Repeated Measures) in Clinical Trials. *Statistics in Medicine*,**18**,1707-1732.

[9] Diggle, P. J., Liang, K. Y. and Zeger, S. L. (1993). *Analysis of longitudinal data*, Oxford University Press, Oxford.

[10] Hardin, J.W. and Hilbe, J.M. (2003). *Generalized Estimating Equations*. Chapman and Hall.

[11] Matthews, J.N.S. et. al. (1990). Analysis of serial measurements in medical research. *BMJ*, **300**, 230-235.

[12] Matthews, J.N.S. (1993). A refinement to the analysis of serial data using summary measures. *Statistics in Medicine*, **12**, 27-37.

[13] Agresti, A. (1999). Modelling ordered categorical data: recent advances and future challenges. *Statistics in Medicine*, **18**, 2191-2207.

[14] Zeger, S. L. and Liang, K. Y. (1992). An overview of the methods for longitudinal data. *Statistics in medicine*, **11**, 1825-1839.

[15] Zhou, X. H., Perkins, A. J. and Hui, S. L. (1999). Comparisons of software packages for generalized linear multilevel models. *American statistical association*, **53**, 282-290.

[16] Zeger, S. L., Liang, K.Y. and Albert, P.S. (1988). Models for longitudinal data: A generalised estimating equation approach. *Biometrics*, **44**, 1049-1060.

[17] Lindsey, J. K. and Lambert, P. (1998). On the appropriateness of marginal models for repeated measures in clinical trials. *Statistics in Medicine*, **17**, 447-469.

[18] Lee, Y., and Nelder, J. A. (2004). Conditional and Marginal Models: Another View. *Statistical Science*, **19**, 219-238.

[19] Neuhaus, J. M., Kalbfleisch, J. D. and Hauck, W. W. (1991). A comparison of cluster-specific and population-averaged approaches for analysing correlated binary data. *International Statistical Review*, **59**, 25-35.

[20] Begg, M.D., and Parides, M.K. (2003) Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data. *Stats in Med*, **22**, 2591-2602.

[21] Goldstein, H. and Rasbash, J. (2002). Tutorial in biostatistics: Multilevel modelling of medical data. *Statistics in medicine*, **21**, 3291-3315.

[22] Zorn, C. J. W. (2001). Generalized estimating equations models for correlated data: A review with applications. *American Journal of Political Science*, **45**, 470-490.

[23] Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data*, 2nd Ed., John Wiley, New Jersey.

[24] Little, R. J. A. (1995). Modelling the dropout mechanism in repeated-measures studies. *Journal of the American Statistical Association*, **90**, 1112-1121.

[25] Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1995). Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data. *Journal of the American Statistical Association*, **90**, 106-129.

[26] Heyting, A., Tolboom, J. and Essers, J. (1992). 'Statistical handling of dropouts in longitudinal clinical trials. *Statistics in Medicine*, **11**, 2043-2061.

[27] Carpenter, J., Pockock, S. and Lamm, C. J. (2002). Coping with missing data in clinical trials: A model-based approach applied to asthma trials. *Statistics in medicine*, **21**, 1043-1066.

[28] Hogan, J.W., Roy, J. and Korkontzelou, C. (2004). Handling dropout in longitudinal studies. *Statistics in Medicine*, **23**, 1455-1497.

[29] Scharfstein, D.O., Daniels, M.J. and Robins, J.M. (2003). Incorporating prior beliefs about selection bias into the analysis of randomized trials with missing outcomes. *Biostatistics*, **4**(4), 495-512.

[30] White, I.R., Carpenter, J., Evans, E. and Schroter, S. (2004). Eliciting and using prior opinions about dropout bias in randomised controlled trials. Submitted to *Clinical Trials.*

[31] Rubin, D. B. (1987). *Multiple Imputation for Non-response in Surveys* John Wiley, New York.

[32] Carpenter, J., Kenward, M. and Vansteelandt, S. (2005). A comparison of multiple imputation and inverse probability weighting for analyses with missing data. *Journal of the Royal Statistical Society, Series A*, in press.

[33] Simons, F.E.R. and the Canadian Beclomethasone Dipropionate – Salmeterol Xinafoate Study Group (1997). A comparison of Beclomethasone, Salmeterol and placebo in children with asthma. *New England Journal of Medicine*, **337**, 1659-1665.

[34] Fitzmaurice et. al. (2001). Bias in Estimating Association Parameters for Longitudinal Binary Responses with Drop-Outs. *Biometrics* **57**, 15-21.

[35] Oman, S.D. and Zucker, D.M. (2001). Modelling and generating correlated binary variables. *Biometrika*, **88**, 287-290.

[36] *Stata Cross-sectional Time Series Reference Manual Release 8*, Texas, Stata Press Publication (2003).

[37] Liu, Q. and Pierce, D.A. (1994). A note on Gauss-Hermite quadrature. *Biometrika*, **81**, 624-629.

[38] Carey, V., Zeger, S.L. and Diggle, P.J. (1993). Modelling multivariate binary data with alternating logistic regression. *Biometrics* **80**, 517-526.

[39] Lipsitz, S.R. and Fitzmaurice, G.M. (1996). Estimating equations for measures of association between repeated binary responses. *Biometrics* **52**, 9003-912.

[40] Lipsitz, S.R., Laird, N.M. and Harrington, D.P. (1992). A three-stage estimator for studies with repeated and possibly missing binary outcomes. *Applied Statistics* **41**, 203-213 (1992).

[41] Little, R. J. and Raghunathan, T. (1999). On summary measures analysis of the linear mixed effects model for repeated measures when data are missing completely at random. *Statistics in Medicine*, **18**, 2465-2478.

[42] Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, **83**, 1198-1202.

[43] Park, T. and Davis, C.S. (1993). A test of the missing data mechanism for repeated categorical data. *Biometrics*, **49**, 631-638.

[44] Park, T. (1997). A test of missing completely at random for longitudinal data with missing observations. *Statistics in medicine*, **16**, 1859-1871.

[45] Diggle, P. J. (1989). Testing for random dropouts in repeated measurement data. *Biometrics*, **45**, 1255-1258.

[46] Ridout, M. S. (1991). Testing for random dropouts in repeated measurement data. *Biometrics*, **47**, 1617-1621.

[47] Listing, J. and Schlittgen, R. (1998). Tests if dropouts are missed at random. *Biometrical Journal*, **40**, 929-935.

[48] Listing, J. and Schlittgen, R. (2003). A nonparametric test for random dropouts. *Biometrical Journal*, **45**, 113-127.

[49] Chen, H. Y. and Little, R. (1999). A test of missing completely at random for generalised estimating equations with missing data. *Biometrika*, **86**, 1-13.

[50] Qu, A. and Song, P. X.-K. (2002). Testing ignorable missingness in estimating equation approaches for longitudinal data. *Biometrika*, **89**, 841-850.

[51] Qu, A., Lindsey, B. G. and Li, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika*, **87**, 823-36.

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

[53] Spiegelhalter, D.J. et. al. (1994). Bayesian Approaches to Randomized Trials. *Journal of the Royal Statistical Society A*, **157** (3) 357-416.

[54] Gould, A. L., (1980). A new approach to the analysis of clinical drug trials with withdrawals. *Biometrics*, **36**, 721-727.

[55] Carpenter, J., Kenward, M., Evans, S. and White, I. (2004). Letter to the editor: Last observation carry forward and last observation analysis by J. Shao and B. Zhong, Statistics in Medicine, 22, 2429-2441 (2003). *Statistics in Medicine*, **23**, 3241-3244.

[56] Lachin, J. M. (1999). Worst-rank score analysis with informatively missing observations in clinical trials. *Controlled Clinical Trials*, **20**, 408-422.

[57] Rubin, D.B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, **72**, 538-543.

[58] Dufouil, C., Brayne, C. and Clayton, D. (2004). Analysis of longitudinal studies with death and drop-out: a case study. *Statistics in Medicine*, **23**, 2215-2226.

[59] Garthwaite, P.H. et, al. (2005). Statistical Methods for Eliciting Probability Distributions. *Journal of the American Statistical Association*, **100**, 680-701.

[60] Gelman, A., Carlin, J. B., Stern, H.S. and Rubin, D. B. (2004). *Bayesian Data Analysis, second edition*, 517-540. Chapman and Hall.

[61] Spiegelhalter, D., Thomas, A., Best, N. and Lunn, N., 'WinBUGS User Manual', MRC Biostatistics Unit, Cambridge, England. www.mrc-bsu.cam.ac.uk/bugs/winbugs/manual14.pdf