Comparative approaches to handling missing data, with particular focus on multiple imputation for both cross-sectional and longitudinal models

A dissertation submitted in fulfilment of the academic requirements for the degree of doctor philosophy in Statistics

By

Ali Satty Ali Hassan

School of Mathematics, Statistics and Computer Science

University of KwaZulu-Natal
South Africa

October, 2012

Preface

The work described in this thesis was carried out in the School of Mathematics, Statistics and Computer Science, University of Kwazulu-Natal, Pietermaritzburg, South Africa, under the supervision and direction of Professor Henry Mwambi (School of Mathematics, Statistics and Computer Science, University of KwaZulu Natal).

I would like to declare that this thesis is my own work. It has not been submitted in any form for any degree or diploma to any other University. It, therefore, represents my original work. Where use has been made of the work of others, it is duly acknowledged within the text or references chapter.

....

Signature (Ali Satty Ali) Date

As the candidate's supervisor, I certify the above statement and have approved this thesis for submission.

....

Signature (Prof. Henry Mwambi) Date

Declaration 1 - Plagiarism

I, Ali Satty Ali, declare that:

- The research reported in this thesis, except where otherwise indicated, is my original research.
- This thesis has not been submitted for any degree or examination at any other university.
- This thesis does not contain other person's data, pictures, graphs, or other information, unless specifically acknowledged as being sourced from other persons.
- This thesis does not contain other person's writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:

  a. Their words have been re-written, but the general information attributed to them has been referenced.

  b. Where their exact words have been used, then their writing has been placed in italics and inside quotation marks, and referenced.
- This thesis does not contain text, graphics, or tables copied and pasted from the internet, unless specifically acknowledged and the source being detailed in the thesis and in the references section.

Declaration 2 - Publications and manuscripts

I, Ali Satty Ali, declare that:

- A. Satty and H. Mwambi (2012). Imputation methods for estimating regression parameters under a monotone missing covariate pattern: A comparative analysis. *South African Statistical Journal*, 46, 327-356.

- A. Satty and H. Mwambi (In review). An analysis for handling dropouts in longitudinal data using multiple imputation and inverse probability weighting. *Statistics and its inference.*

- A. Satty and H. Mwambi (In review). A comparative analysis of likelihood based and multiple imputation methods for incomplete longitudinal data with ignorable missingness. *REVSTAT Statistical Journal* - Presented October 28-30, 2011. Brague: Czech Republic, NC.

- A. Satty, H. Mwambi and G. Molenberghs (In review). Different methods for handling non-Gaussian longitudinal outcome subject to potentially random dropout. *Communications in Statistics - Simulation and Computation.*

- A. Satty, H. Mwambi and M. Kenward (In review). An analysis of incomplete longitudinal data with application to multi-centre trial data: A selection model for non-ignorable missingness. *South African Statistical Journal.*

- A. Satty and H. Mwambi (Accepted). Selection and pattern mixture models for modelling longitudinal data with dropout: An application study. *International Journal of Biostatistics.*

Dedication

To God almighty for his mercy that endures forever. To my mother, Alawia Abd-Elsamad for her constant support and prayers for my success. To my beloved brothers and sisters. To the departed soul of my father, Satty, may Allah forgive him and grant him paradise.

Acknowledgements

I would like to acknowledge many people for helping me during my doctoral work. First and foremost, I thank Allah for granting me power, support and much more to make it possible to complete my PhD thesis. Special thank goes to Prof. Henry Mwambi who was my main supervisor in the field of incomplete data analysis. I cannot adequately put my appreciation and thanks into words but let me just say thank you for your expert scientific guidance, the continuous encouragement, support that you gave to me, the patience that you displayed and the effort you put into helping me to complete my PhD research. I gratefully acknowledge support from Prof. Kenward, G. M. (London School of Hygience and Tropical Medicine, UK) and Prof. Geert Molenberghs (Center for Statistics - CenStat - Hasselt University, Belgium) for their advice, explanations and comments on the implementation of the statistical methodologies used in this thesis. My gratitude goes to Prof. Milliken, G. A. (Kansas State University - USA) for the kind permission to use his data. I apologize to my mother, brothers and sisters for the time not spent with them during the preparation of this thesis and I am grateful for their understanding. I wish to thank the anonymous referees who gave useful comments on the papers during the course of this work. My appreciation extends to the School of Mathematics, Statistics and Computer Science, University of Kwazulu-Natal. There are many others who assisted but I would like to thank Dr. Ahmed Abdelmageed for his assistance in the simulations analysis. Also, I would like to express my deepest gratitude to Dr. Khalid Mansour Yousif for his steadfast support as well as academic and social interactions. Finally, I am thankful to my friend Haythem Hashim for his support and assistance with my family back home in Sudan. Without this image, my study could not have been accomplished.

Abstract

Much data-based research are characterized by the unavoidable problem of incompleteness as a result of missing or erroneous values. This thesis discusses some of the various strategies and basic issues in statistical data analysis to address the missing data problem, and deals with both the problem of missing covariates and missing outcomes. We restrict our attention to consider methodologies which address a specific missing data pattern, namely *monotone* missingness.

The thesis is divided into two parts. The first part placed a particular emphasis on the so called missing at random (MAR) assumption, but focuses the bulk of attention on multiple imputation techniques. The main aim of this part is to investigate various modelling techniques using application studies, and to specify the most appropriate techniques as well as gain insight into the appropriateness of these techniques for handling incomplete data analysis. This thesis first deals with the problem of missing covariate values to estimate regression parameters under a monotone missing covariate pattern. The study is devoted to a comparison of different imputation techniques, namely markov chain monte carlo (MCMC), regression, propensity score (PS) and last observation carried forward (LOCF). The results from the application study revealed that we have universally best methods to deal with missing covariates when the missing data pattern is monotone. Of the methods explored, the MCMC and regression methods of imputation to estimate regression parameters with monotone missingness were preferable to the PS and LOCF methods. This study is also concerned with comparative analysis of the techniques applied to incomplete Gaussian longitudinal outcome or response data due to random dropout. Three different methods are assessed and investigated, namely multiple imputation (MI), inverse probability weighting (IPW) and direct likelihood analysis. The findings in general favoured MI over IPW in the case of continuous outcomes, even when the MAR mechanism holds. The findings further suggest that the use of MI and direct likelihood techniques lead to accurate and equivalent results as both techniques arrive at the same substantive conclusions. The study also compares and contrasts several statistical methods for analyzing incomplete non-Gaussian longitudinal outcomes when the underlying study is subject to ignorable dropout. The methods considered include weighted generalized estimating equations (WGEE), multiple imputation after generalized estimating equations (MI-GEE) and generalized linear mixed model (GLMM). The current study found that the MI-GEE method was considerably robust, doing better than all the other methods in terms of small and large sample sizes, regardless of the dropout rates.

The primary interest of the second part of the thesis falls under the non-ignorable dropout (MNAR) modelling frameworks that rely on sensitivity analysis in modelling incomplete Gaussian longitudinal data. The aim of this part is to deal with non-random dropout by explicitly modelling the assumptions that caused the dropout and incorporated this additional sub-model into the model for the measurement data, and to assess the sensitivity of the modelling assumptions. The study pays attention to the analysis of repeated Gaussian measures subject to potentially non-random dropout in order to study the influence on inference that might be caused in the data by the dropout process. We consider the construction of a particular type of selection model, namely the Diggle-Kenward model as a tool for assessing the sensitivity of a selection model in terms of the modelling assumptions. The major conclusions drawn were that there was evidence in favour of the MAR process rather than an MCAR process in the context of the assumed model. In addition, there was the need to obtain further insight into the data by comparing various sensitivity analysis frameworks. Lastly, two families of models were also compared and contrasted to investigate the potential influence on inference that dropout might have or exert on the dependent measurement data considered, and to deal with incomplete sequences. The models were based on selection and pattern mixture frameworks used for sensitivity analysis to jointly model the distribution of the dropout process and longitudinal measurement process. The results of the sensitivity analysis were in agreement and hence led to similar parameter estimates. Additional confidence in the findings was gained as both models led to similar results for significant effects such as marginal treatment effects.

Table of Contents

List of Figures

List of Tables

Chapter 1

General introduction

1.1 Fundamental concepts

The generic term *missing* or *incomplete data* means that we are missing some types of information about the phenomena of interest in a particular analysis, in the sense that not all planned or desired data are actually made. In other words, there are values in the data matrix that are unknown or uncollected. Day (1999) defined missing data as "data that refer to a data value that should have been recorded but, for some reason, was not". This is especially true when studies are applied on human subjects which is often the case in longitudinal studies, but may as well be on animals, plants or groups of individuals, to mention some examples. Missing data is one of the most common statistical and design problem in many fields of research since they are usually not the focus of any given study but are frequently encountered in empirical studies by statisticians. Missing data, which are typically viewed as a nuisance, cause problems when it comes to data analysis since most standard statistical techniques and softwares are designed assuming complete data for each variable included in the analysis. In principle, the main objective of statistical analysis is to make valid and efficient inferences about a population of interest. This can be achieved with or without missing data, but the missing data problems complicate the process. More consideration in planning a study should therefore be given to minimizing missing data or to avoid this problem and to make the study scientifically sound. However, the existence of missing data is common and cannot be avoided in data-based research even if great effort is put into planning and data collection (Allison, 2002; Carter, 2006; Regoeczi and Riedel, 2003;

Rudas, 2005; Stumpf, 1978).

According to McKnight et al. (2007), the problems of missing data occur under three broad sources affecting either complete subjects or specific items, namely cases, variables and occasions. In the first source, missing cases, the missing data occur when the study participants fail to provide data for a study. In the second source, missing variables, the missing data occur when participants fail to provide data for some but not all variables. In the third source, missing occasions (i.e., follow-up data), the missing data occur when participants are available for some but not all of the data collection periods in a study.

On the other hand, there are various participant and design related reasons for missing data among them: (1) the study participants, where missing data can occur when some participants are offended by specific questions in a questionnaire (participant characteristics); (2) the nature study design, when the nature of the design takes up the participants time leading to withdrawal (design characteristics); and (3) the interaction between the participants and the study design participants who are the sickest are unable to complete the more burdensome aspects of the study (participant and design characteristics). Generally speaking, data may be almost always missing regardless of the field of the study due to accidental or data entry errors.

The effects of missing data can be explained in terms of the amounts, the patterns of missing data and the methods used for handling the missing data which also have implication for the interpretation of the statistical analysis of the study. Such missingness can be associated with three possible difficulties. First, loss of information, efficiency, or power. Second, problems in data handling, computation and analysis due to irregularities in the data patterns and non-applicability of standard software. Third, seriously marked bias if there are systematic differences between the observed and unobserved data (Barnard and Meng, 1999). Particularly, it may mean that there may be insufficient data to draw any useful conclusions from a study. When data are missing it is also hard to specify the impact the data might have presented in the statistical analysis study. The extent of the impact of missing data on study results is dependent on:

- The amount of missing data. This is related to its impact on the study conclusions. Large impact on statistical inference can occur where there is a greater amount of missing data. In other words, the power of the statistical tests can be severely compromised (De Leeuw et al., 2003).

- The mechanisms of missing data. The process causing missing data can affect the validity of the statistical inferences. If the process depends on causal effects factors, the missing data can have dramatic impact on the validity of the results.

- The procedure or method the statistician or data analyst will use to deal with these missing data (Musil et al., 2002; Streiner, 2002).

Missing data are a fact of life in many disciplines of science including medical studies (Piantadosi, 1997; Green et al., 1997; Friedman et al., 1998) and epidemiological studies (Kahn and Sempos, 1989; Clayton and Hills, 1993; Lilienfeld and Stolley, 1994; Selvin, 1996). Missing data in surveys, psychometry and econometrics are discussed in Fowler(1988), Schafer et al. (1993), Rubin (1987) and Rubin et al. (1995), to name but a few literature. Examples are also abound in the context of experimental and observational data in non-human life setting such as environmental, agricultural and biological studies. This thesis focuses on both cross-sectional and longitudinal data examples. There are several earlier studies on the problem of missing data largely concerned with algorithmic and computational solutions to the induced lack of balance or deviations from the intended study design (Molenberghs and Verbeke, 2005). Early work on missing data include, Affifi and Elashoff (1966) and Hartley and Hocking (1971). Shortly afterward, other applications in the literature, such as Expectation Maximization (EM) by Dempster et al. (1977), data imputation and augmentation procedures (see, Rubin, 1987; Tanner and Wong, 1987), combined with strong computing resources of solving the computational difficulties followed. These studies revolutionized the handling of missing data in statistical analysis.

1.1.1 Patterns and categories of missing data

1.1.1.1 Patterns of missing data

Missing data pattern describes and explains the geography of the data set where in the data set the values are observed and where the values are missing. As we will see later, some procedures to deal with missing data can be applied to any missing data pattern whereas other procedures are restricted to specific missing data patterns, and therefore having identified the variables that define the pattern, a suitable analysis procedure can be derived. To investigate the missing data pattern, it is important to identify cases and variables that contribute to the incomplete data (Schafer and Graham, 2002; Allison, 2002). More formally, let $Y = (y_{ij})$ be an $(n \times K)$ rectangular data set containing missing data, with ith row $y_i = (y_{i1}, ..., y_{iK})$, where y_{ij} denotes a value of variable Y_j for subject i. Moreover, in the presence of missing data, we define the missing data indicator matrix, R. We additionally define R as follows: R equals 1 if, y_{ij} is unobserved, and 0 otherwise. The missing data pattern can then be defined by the matrix R whose (i, j)th element is R_{ij}. Figure 1.1 represents some important examples of missing data patterns similar to the illustration in Little and Rubin (2002). Figure 1.1a shows a *univariate pattern*, where missingness is confined only in a single variable. As illustrated in this figure, for K=4, the missing data arise on variable Y_4 but a set of other variables, i.e., $Y_1, ..., Y_3$ is fully observed. In Figure 1.1b, the variables or variable groups $Y_1, ..., Y_p$ can be arranged in such a way, that if Y_j is missing, then $Y_{j+1}, ..., Y_p$ should be missing as well. This type of pattern is a *monotone pattern* (Anderson, 1957; Rubin, 1974; Little and Rubin, 1987). Figure 1.1b illustrates a monotone pattern in the case of p=4. Monotone missingness is common in longitudinal studies, but can also appear in other types of data where the ordering of the variables of interest can be taken into account. This pattern for missing data is common in studies done by the pharmaceutical industry as in protocols for many conditions, data are not collected after a study participant discontinues study treatment. This is highlighted in a recent report on the prevention and treatment of missing data by the National Research Council. A summary of the report was provided by Little et al. (2012). However, even in these studies, there typically is both planned and unplanned missing data. A predominately monotone pattern for missing outcome data is less common in clinical outcome studies and in publically-funded trials which are more of a pragmatic nature (e.g., trials in which the intention-to-treat estimand is the primary objective). Figure 1.1c illustrates another example of missing data pattern, namely an *arbitrary pattern* in which missingness may arise at any set of variables for any unit. This type of pattern also called intermittent missingness, and is more commonly encountered in multivariate based-data. In the case of combining data that can come from two sources, Figure 1.1d shows an extreme version of the *file matching pattern*, with two sets of variables never observed together. This pattern arises under limited situations, for example, when the amounts of missing data are large. In the *file matching pattern*, Y_1 denotes a set of variables that is common to both data sources and observed, Y_2 denotes a set of variables observed for the first data source but not the second, and in contrast to Y_2, the Y_3 indicates a set of variables observed for the second data source but not the first. An excellent discussion of these patterns and others is given in Little and Rubin (2002) and Schafer and Graham (2002).

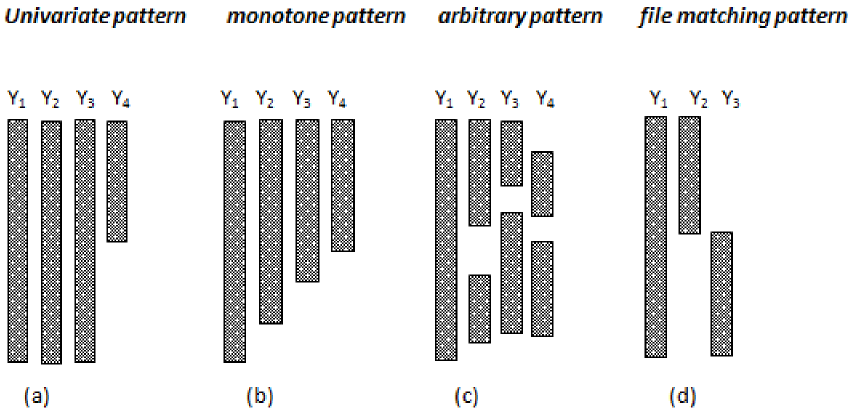


Figure 1.1: *Examples of patterns of non-response. In each cases, rows correspond to observational units and columns correspond to variables*

1.1.1.2 Categories of missing data

When data are missing for reasons unknown or outside the control of the analyst, assumptions about the process that generates the data and its implications for statistical inferences need to be made. A major issue concerns missing data mechanisms which explain *why* data are missing. Missing data mechanisms are the properties of the standard methodologies for incomplete data analysis which depend very much on the nature of the dependencies in these categories (De Leeuw, 2001; Little and Rubin, 2002). These mechanisms however do not imply knowledge about how the missing values came to be unavailable. It is noted that the term *dropout mechanism* can be used when it relates to subjects dropping out of a clinical trial study prematurely, particularly in the context of longitudinal studies. The term "dropout" is misused by many researchers as , in many trials, data are missing not because a participant chooses to dropout but instead because the protocol is written not to follow participants following treatment discontinuation. Discontinuation might be due to adverse effects, lack of efficiency, both of these reasons, or other reasons. As demonstrated by Rubin (1976) and Little and Rubin (1987), the mechanisms that lead to missing data can be classified into three basic categories. Data are considered *Missing Completely at Random* (MCAR) when the mechanism that generates the missing values is a truly random process unrelated to any measured or unmeasured characteristic of the study participants. A second category is *Missing at Random* (MAR) in which the missingness mechanism is random meaning, conditional on the observed measurements characteristics of the study sample, the missingness mechanism is independent of the unobserved measurements. A third category, *Missing Not at Random* (MNAR), is one in which the missingness process depends on unobserved measurements and possibly on the observed measurement characteristics of the study sample. Figure 1.2 illustrates the differences between these three mechanisms. This figure is similar to the one given

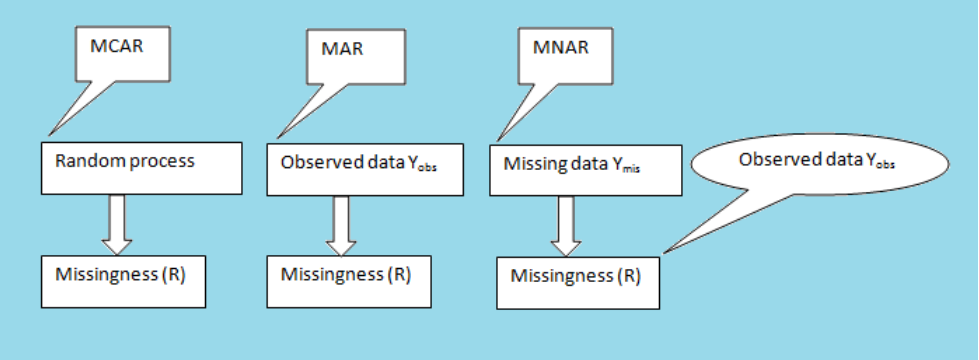


Figure 1.2: *Rubin's (1976) classifications of missing data mechanisms*

in McKnight et al. (2007) for the description of general missing data mechanisms. Each of these mechanisms refer to the probability of missingness, given information about the variable(s) with the missing data, associated variables and a hypothetical mechanism underlying the missing data. Rubin's (1976) classifications are related to the level of bias that missingness may exert on statistical analysis where it is stated that MCAR has negligible potential impact and MNAR has the greatest potential impact. Furthermore, it is impossible to distinguish which underlying categories of missingness are in play, unless one knows the motivation for an individual's dropping out. This problem is discussed further in Molenberghs et al. (2008) who show that the formal-based distinction between MAR and MNAR is not possible. This is because for any MNAR model there exists an MAR model that fits the data equally well, but they differ in the prediction of what is unobserved. Hence, it is broadly agreed that the role of such MNAR model is in sensitivity analysis; that is, if the assumptions are changed, the conclusions from the primary (typically MAR) analysis are also changed. According to Molenberghs and Verbeke (2005) and Molenberghs and Kenward (2007), sensitivity analysis is defined as an analysis in which several statistical models are considered simultaneously under different missing data scenarios. This point will be discussed further in Chapter 7.

1.1.2 Ignorability

Determination of the missing data mechanisms is important. According to Rubin (1976), there are two important broad classes of missing data: missing data that is *ignorable* from the analysis, and missing data that is *non-ignorable*. If one can reasonably assume that missing data occur under either MCAR or MAR conditions, the problem is deemed ignorable, and the missingness process need not be explicitly modelled. Moreover, when data are MCAR or MAR, the likelihood-based and Bayesian frameworks allow to ignore the missingness process since they use only observed data, conditional on the model being correctly specified (Little and Rubin, 2002). In contrast, when data are MNAR, the missingness process cannot be ignored from the analysis (Little and Rubin, 2002). In the application to missing data classifications, ignorability, as it applies to missingness mechanisms, does not mean that investigators can ignore missing values. It refers to the fact that factors that cause missingness are unrelated or weakly related to the estimated intervention effect. In a restricted sense, the term refers to whether missingness mechanisms must be modelled as part of the parameter estimation process or not (Allison, 2002). In addition, the importance of ignorability arises when one needs to evaluate the impact of missing data in the analysis and study conclusions. Because the way they are missing is random, MCAR data should have no systematic effect between complete and missing records on results. However, in the MAR data there is a systematic process underlying the missingness, but this effect can be modelled using the observed data (McKnight et al., 2007).

For simplicity, let x be the auxiliary variable that is observed for the entire sample, y the study variable subject to missingness and r the response missingness status variable. Assume the problem is to find the best prediction model for y in terms of x. In this setup, the prediction model can be used to predict the missing data if the response missing data mechanism is ignorable, which is to say that the relationship between y and x in the respondents also holds for the non-responding part of the sample. Intuitively, the missing data is ignorable if the study variable, y, is independent of the value of the status variable, r, given the auxiliary variable, x. Conversely, the missing data is non-ignorable if the probability of y being missing is dependent on y itself, even after controlling for x. Hence, it follows that the MNAR mechanism implies a violation of the ignorability principle and requires appropriate measures to account for the effects of data that is MNAR which is also referred to as the non-ignorable situation. Further, the effect of mechanisms that are non-ignorable is unknown which means there is not enough information from the data set used in the analysis itself to allow the investigator to model and study the way in which the data are missing. Thus, non-ignorable missing data is substantially more difficult to deal with and must be handled with caution. According to Thijs et al. (2002), it is not feasible to make a satisfactory analysis of the data under non-ignorable missingness. The role of non-ignorable mechanism has been studied in terms of various applications settings, see, for example, Belin et al. (1993), Wachtar (1993), Little and Rubin (2002) and Demirtas and Schafer (2003). According to Verbeke and Molenberghs (2000), to investigate the impact of deviations from the ignorability mechanism on the conclusions, one needs to apply a sensitivity analysis in which models for the non-ignorable process can play a vital role. The ignorability case will be treated in the first four chapters of this thesis. In the last two chapters we switch to the non-ignorability setting that occupies the most prominent role in incomplete Gaussian longitudinal data.

1.2 Techniques for handling missing data

Given the problems that can arise when there are missing data, the following question is forced upon researchers. What methods can be utilized to handle these potential pitfalls? The goal is to use approaches that better avoid the generation of biased results.

Different methods have been suggested for dealing with missing data. A comprehensive review of many of these methods is provided by Schafer and Graham (2002) which includes methods that are no longer recommended and more recent methods such as multiple imputation and data augmentation methods. A useful discussion in terms of the technical details of methods for handling missing data is given by Schafer and Olsen (1998). Additionally, Rubin (1987, 2002), Van der Laan and Robins (2003), Molenberghs and Verbeke (2005), Tsiatis (2006) and Molenberghs and Kenward (2007) have all produced rigorous accounts of methods that can be used to deal with missing data. Some techniques or methods use different approaches to addressing missing covariates and missing outcomes. Although the missing data problem is ubiquitous, there is still no firm consensus on what statistical procedures should be used for analysis or on the circumstances under which they should be applied. The literature presents various techniques that can be used to deal with missing data, and these range from simple classical *ad hoc* methods to model-based methods. These methods should be fully understood and appropriately characterized in relation to missing data and should be theoretically proved before they are used practically. Furthermore, each method is based on a specific missingness mechanism, but one needs to realize that at the heart of the missingness problem it is impossible to identify the missing data mechanism.

Prior to 1970s, missing values were solved primarily by editing (Schafer and Graham, 2002). Until the mid 1970's, the principal methods that were used to deal with missing data included case deletion, single imputation and maximum likelihood estimation. From the 1970's to the 1980's, methods like the expectation maximization (EM) algorithm and multiple imputation using maximum likelihood estimation were applied to a broad range of problems. During this period, a framework of inference from missing data was developed by Rubin (1976). Then, the Expectation Maximization (EM) algorithm was formulated by Dempster et al. (1977) as a popular tool for making full use of maximum likelihood (ML) for dealing with incomplete data analysis. The breakthrough idea of multiple imputation was introduced by Rubin (1978) and Little and Rubin (1987), but there were many difficulties concerning the creation of multiple imputations in terms of the computational resources and capabilities available at the time (Schafer and Olsen, 1998). The 1980's did however see the development of many facilities for solving this problem such as fast computer technology and new methods for Bayesian simulation (Schafer, 1997). The drawbacks of case deletion and single imputation techniques have been documented by Little and Rubin (1987). Applications to various problems using maximum likelihood estimation were advanced in the 1990's. It was during this period that the EM algorithm was extended to different forms, namely stochastic EM algorithm (SEM), the expectation conditional maximization algorithm (ECM) and the stochastic expectation conditional maximization algorithm (SECM). Bayes simulation techniques such as Markov Chain Monte Carlo (MCMC) and data augmentation were developed in the same period.

Recently, researchers have shifted their focus to more modern techniques that avoid the specification of a full parametric model for the population (Robins et al., 1994). Since 1995 many techniques for dealing with missing data have been discussed and developed in different applications. Most recently, many techniques have been proposed to assess the sensitivity of the results to the distribution of missingness (Verbeke and Molenberghs, 2000). In the context of non-ignorable missingness setting, the primary focus has been dropout in longitudinal clinical trials data which is where individuals may drop out of the study for reasons closely related to the outcomes being measured (Diggle and Kenward, 1994; Little, 1995; Verbeke and Molenberghs, 2000; Molenberghs and Verbeke, 2005; Molenberghs and Kenward, 2007). We return to this issue in

Chapters 6 and 7.

What follows now is a brief description of the several methods that are commonly used to deal with missing data, including a review of the existing literature in which we examine the effectiveness of these methods in the analysis of incomplete data.

1.2.1 Deletion methods

There are several ways to deal with missing data. One of them is to discard subjects with incomplete sequences, and then analyze only the units with complete data (Nie et al., 1975). Methods that use this approach are called *deletion methods.* These methods do not replace or impute missing values and do not make other adjustments to account for missing values. They share many properties in terms of missing data mechanisms and the inefficiencies inherent in losing data for statistical power, although not all to the same degree. The main advantage of these techniques is their simplicity and the ease with which they can be applied using much of the standard statistical softwares. Brown (1983) states that some of the deletion methods are good options, but only when used under specific circumstances (i.e., when the amount of missing data is small and when data are MCAR, for example, the complete case discussed below and the available case which uses all available cases and discards data only at the level of the variable, not at observation level). However, because such circumstances are rare, McKnight et al. (2007) advise that one should avoid the deletion methods whenever possible. Furthermore, Little and Rubin (2002) do not recommend any of the deletion methods except in specific situations where the amount of missing data is limited.

Next, we briefly discuss the complete case as a deletion method, explaining its use, strengths and weaknesses.

1.2.1.1 Complete case analysis

The simplest deletion approach is the complete case analysis or list-wise deletion analysis in which the analysis uses only those subjects with completely recorded observations. In other words, for all variables under consideration, the complete case confines attention to observations that are available. For example, in longitudinal studies, this method uses only those individuals with observed responses at each time point.

This method has numerous advantages. The first is simplicity, in that the method can be quite effective and may be satisfactorily used with small amounts of missing data. However, it is important to make sure that, even in such situation, the deleted cases are not unduly influential (Schafer and Graham, 2002). The second advantage to complete case analysis is that, it is easy to carry out. It is used by default routines in most statistical software packages, but it has varying details of implementations.

The primary disadvantages of this method are that: (1) it can produce inefficient estimates, in the sense of loss of statistical power specifically when drawing inferences for sub-populations; and (2) when data are not MCAR, then the method can lead to serious biased results. In other words, it is a valid method only when data are MCAR (Little and Rubin, 1987), but even when MCAR holds, it can still be inefficient (Schafer and Graham, 2002). Thus, McKnight et al. (2007) state that one should give careful consideration before the use of this method regardless of its ease of use. Furthermore, it is easy to envisage situations where complete case can be very misleading. Kenward et al. (1994) and Wang-Clow et al. (1995) present examples where the complete case has led to misleading results.

1.2.2 Imputation-based methods

In contrast to the above mentioned techniques, we now discuss methods that do generate possible values for the missing data. These alternative methods are called *imputation* methods, where one “fills-in” (imputes) the missing data to obtain a full data set, and the resultant data are then analyzed by standard statistical methods without concern as if the set represented the true and complete data set (Rubin, 1987; Little and Rubin, 1987). This is the key idea behind commonly used procedures for imputation which include, *simple and multiple imputation* (Little and Rubin, 1987). Multiple imputation fills in more than one value for each missing item to allow for the appropriate evaluation of imputation uncertainty (Rubin, 1987; Little and Rubin, 1987). In contrast to multiple imputation, simple imputation techniques substitute one value for every missing value in the data set (Little and Rubin, 1987, 2002). In this section, we restrict ourselves to outlining several simple imputation methods which are valid under the ignorability assumption (Rubin, 1987; Allison, 2002; Schafer and Graham, 2002).

There are simple imputation methods that include: (1) mean imputation, in which missing observations are replaced with the estimated mean of the data set; (2) last observation carried forward (LOCF), which is revisited briefly in Chapter 2; (3) regression imputation, where the missing data are imputed using the prediction taken from a multiple regression analysis; (4) Hot Deck imputation, in which the missing data can be replaced with the observed data taken from a matched data from the variables that contain non-missing values; and (5) stochastic regression imputation, in which missing values are replaced by a value that is predicted using regression imputation plus a residual that is drawn to reflect uncertainty in the predicted value.

Simple imputation methods are general and flexible for handling missing data, and can be implemented quickly in several statistical softwares (for example, SAS, R, S+ and others). However, with respect to accurately reproducing known population results (parameter estimates and standard errors), each of these single imputation methods have been found to be inadequate (Schafer and Graham, 2002). The problems linked with these techniques include: (1) the performance of these techniques is poor even when the ignorable missing data assumption (MCAR or MAR) holds, a situation that limits their suitability to quite a restricted set of assumptions (Allison, 2002); (2) they produce seriously biased results that may or may not be predictable; (3) when using these techniques, the standard errors and standard deviations tend to be underestimated, and therefore there is a greater likelihood of committing type-I error (see, Schafer and Graham, 2002). The variability of the estimators is also underestimated since imputed data are treated as observed data; and (4) these techniques may present inconsistent point estimates when data are MCAR.

1.2.3 Data augmentation methods

Data augmentation methods avoid many of the inherent shortcomings of deletion methods. Such methods derive parameter estimates from the available data as well as from either the probability model or an underlying distribution. In contrast to some of the single imputation methods, data augmentation does not replace missing values. In estimating parameters, this algorithm takes into account the missing data, the observed data and the relationships between observed data and several underlying assumptions which is to say that parameter estimates from the observed data are augmented by the additional information provided by the proposed probability model or underlying distribution. In the context of missing data, Maximum Likelihood (ML), Expectation Maximization (EM), Markov Chain Monte Carlo (MCMC) and weighting methods are considered to be augmentation methods. However, as argued by McKnight et al. (2007), the classification of several of these methods as augmentation methods is not clear-cut, specifically for the MCMC, ML and EM methods. The MCMC method has been referred to as an augmentation method within the context of multiple imputation (Allison, 2002). The ML and EM methods have been described as model-based methods by Little and Rubin (1987), while these procedures have also been referred to as data augmentation by Schafer (1997). We now limit our focus to just a few of these methods as augmentation methods, namely ML, EM and weighting methods.

1.2.3.1 Maximum likelihood (ML)

ML was not designed specifically to deal with missing data in the same way as do, for example, LOCF or multiple imputation. ML is an estimation procedure for estimating parameters under different models such as structural equation models (SEM) and ordinary least squares in regression. We discuss the ML as a method for handling missing data. Examples for applying ML to missing data problems can be found in Little and Rubin (2002). Furthermore, in a variety of situations, ML has proven to be an excellent technique for dealing with missing data. When missing data are ignorable (MAR or MCAR), ML does well, and it produces unbiased estimates (Arbuckle, 1996; Allison, 2002). Therefore, the ML is fairly easy to describe under this assumption. If the assumption is met, ML estimators for missing data produce estimates that have the following desirable properties: unbiased estimates in large samples, estimates that are asymptotically efficient (small standard errors) and satisfy asymptotic normality which is to say that estimates approximate a normal distribution which can then be used to exploit a normal approximation for statistical inference, such as finding confidence intervals and *p*-values (McKnight et al., 2007). ML can furthermore be implemented in most statistical software including SAS, SPSS, S-Plus and others.

1.2.3.2 Expectation maximization (EM)

The EM algorithm was originally proposed by Dempster et al. (1977). It is the process of calculating and imputing a value for each missing variable based on best prediction models. The EM algorithm is a very general iterative algorithm for ML estimation in missing data problems (Little and Rubin, 2002). This algorithm requires the less restrictive MAR assumption. The key idea behind EM is to deal with the missing data problem and the complications of estimates related to the ML estimation by attempting to solve smaller complete data problems which lead to parameter estimates for the entire data set (missing and complete data). The EM algorithm handles the missing data using the following steps: (1) fill-in the values for missing data by using estimated values generated by ML; (2) estimate parameters based on data in step 1; (3) re-estimate parameters based on the parameter estimates from step 2; and (4) re-estimate parameters based on the re-estimated data from step 3, and so on, iterating the process until the final step converges on a solution that differs by only a little amount from the previous solution. Each iteration of the EM algorithm consists of two steps, namely the expectation step and the maximization step (Little and Rubin, 2002). Each step is completed once within each algorithm cycle which is to say that cycles are repeated until a suitable convergence criterion is satisfied. Further theoretical justification of these steps can be found in Dempster et al. (1977) and Little and Rubin (2002). According to Dempster et al. (1977), the fitted parameters (on convergence) are equal to a local maximum of a likelihood function which is the maximum likelihood estimate in the case of a unique maximum. The EM algorithm has two disadvantages: firstly, it is typically very slow to converge, and secondly, it lacks direct provision of a measure of precision for the maximum likelihood estimates. Several proposals have been made to overcome these drawbacks, and we refer to the techniques as provided by Louis (1982), McLachlan and Krishnan (1997), Rubin (1991) and Baker (1992).

1.2.3.3 Weighting methods

As introduced by Flanders and Greenland (1991) and Zhao and Lipsitz (1992), weighted methods are based on observed values. In this way, after ignoring all the missing values from the analysis, the remaining observed values are weighted in accordance with how their distribution approximates of the full sample or population. The methods employ the weights in order to correct for either standard errors associated with the parameters or the population variability. To derive suitable weights, the predicted probability of each response is estimated from the data for the variable with missing values. Generally speaking, weighting methods are a good option under certain circumstances, for example, when a missing data pattern is monotone or is under univariate analysis.

In the context of survey data, Rubin (1987) discusses several methods for applying and estimating weights. Under a suitable joint model for the outcome and covariates, these weighting methods are, in many instances, expected to produce results similar to those of multiple imputation (Schafer and Graham, 2002). In the field of biostatistics, Rubin et al. (1995) developed a weighting regression model that requires an explicit model for the missingness but relaxes some of the parametric assumptions in the data model. This new weighting method was an extension of the generalized estimating equations (GEE) proposed by Liang and Zeger (1986). The newer method is known as weighted generalized estimating equations (WGEE). The classical GEE method is valid when data are MCAR, however, the WGEE method can accommodate missing data if they are MAR, provided that a model for the missing data with regard to observed outcomes or covariates is correctly specified (Rubin et al., 1995). Further discussion and extension of this method is given in Rubin et al. (1995) as well as in Robins et al. (1998). We discuss this method in more detail in Chapters 3 and 5. Currently, weighting methods can be carried out in most popular packages, such as STATA, SAS and SUDANA. The weighting methods have been studied in a wide variety of applications, see, for example, Schluchter and Jackson (1989), Ibrahim (1990), Lipsitz and Ibrahim (1996, 1998), Horton and Laird (1999), Carpenter et al. (2006) and Seaman and White (2011).

1.2.4 MNAR-based models in longitudinal data

All the above methods do not however provide an optimal solution to the problem of non-ignorable missingness. This kind of missingness poses a major complication, in particular in terms of longitudinal data setting. In longitudinal studies, observations that are repeatedly measured over time are bound to be correlated and some may be lost due to dropout from the study. Missingness occurs as a result of dropout (premature withdrawal) or attrition, which refers to the special situation arising in longitudinal studies where individuals fail to complete the study for whatever reason. Dropout is a special case of monotone missing data pattern (Diggle and Kenward, 1994; Little, 1995; Molenberghs et al., 1997; Michiels et al., 1999). There are several applications in the literature which argue that it might be necessary to accommodate dropouts in the modelling process, see, for example, Diggle and Kenward (1994), Little (1993, 1994, 1995), Verbeke and Molenberghs (2000) and Molenberghs et al. (2004). In other words, it is argued that one must model the measurement process jointly with a model for dropout which can itself be considered to be of a scientific interest. Arguably, in terms of non-ignorable dropout, a wholly satisfactory statistical analysis of the data used in the analysis is not feasible, and therefore more careful consideration is necessary with regard to dealing with the non-ignorable situation.

In the presence of non-ignorable dropout, advanced modelling strategies have been developed by modelling the joint distribution of the dropout indicators pattern and the measurements process (including observed and missing measurements). As summarized by Little (1995), Verbeke and Molenberghs (2000) and Molenberghs and Kenward (2007), there are at least three factorizations possible to model the joint distribution of the measurements and dropout indicators. First of all, there is *outcome-dependence factorization*, in which dropout indicators are conditioned on the measurements. For continuous longitudinal data, this approach was adopted by Diggle and Kenward (1994), and more discussion of this approach will be done in Chapters 6 and 7. Secondly, there is *pattern-dependence factorization*, in which the distribution of the measurements is a mixture of the distribution for individuals of distinct sub-groups as determined by the dropout patterns. We pay considerable attention to this approach in Chapter 7. Thirdly, there is *parameter-dependence factorization*, which is conditional on the group of parameters shared by the two components so that the measurements process and dropout indicators are conditional independent. Correspondingly and based on the above-mentioned factorizations, there are thus three kinds of modelling strategies: selection models, pattern-mixture models and shared parameter models. According to Vach and Blettber (1995), Molenberghs et al. (2001b) and Verbeke et al. (2001), the practical limitation of any of these model factorizations is that they are sensitive to the assumptions made on the measurements model and the dropout mechanisms. Molenberghs et al. (2004) state that different analysis models can have a distinct impact on conclusions drawn from the same study. In Chapter 7 of this thesis, the idea of using sensitivity analysis is adopted where, given a practical data set, various modelling frameworks with different dropout mechanisms are applied to the same data.

1.3 Research objectives

The main aim of this research is to investigate various missing data modelling techniques using application examples and to determine the most appropriate technique or techniques among several as well as to gain insight into the appropriateness of these techniques for handling incomplete data analysis. The study also aims to deal with the non-random dropout problem by explicitly modelling the assumptions that cause dropout and incorporate this additional model into the model for the measurement data and to assess the sensitivity of the modelling assumptions. The specific objectives are:

- To study imputation techniques and compare them with others to estimate regression parameters with missing covariates when the missing data pattern is monotone, with a particular focus on the MAR mechanism.

- To investigate and compare the analysis of likelihood-based, inverse probability weighted (IPW) and multiple imputation (MI) procedures applied to incomplete Gaussian longitudinal outcomes subject to potential dropout, with restriction to the ignorability assumption.

- To explore the performance of different families of modelling frameworks in terms of handling dropout that is MAR in non-Gaussian longitudinal outcomes using weighted generalized estimating equations (WGEE), multiple imputation after generalized estimating equations (MI-GEE) and generalized linear mixed model (GLMM).
- To investigate the influence on inference that might be exerted on considered data by the non-ignorable dropout (MNAR) process using a selection model based on Diggle-Kenward type approach, as well as dealing with incomplete sequences.
- To demonstrate and contrast the application of two families of MNAR-based models, namely selection and pattern mixture models for investigating the potential influence that dropout might exert on the dependent measurement of the considered data as modelling frameworks that could be used for sensitivity analysis.

1.4 Thesis outline

The thesis is organized as a collection of 6 research papers which have been submitted for peer review international journals. Of these 6 papers, 1 paper has been published, 1 paper has been accepted for publication and the others are still in review. Each paper has been written as a stand-alone article that can be read separately from the rest of the thesis but draws separate conclusions that link to the overall research objectives and questions. As a result, a number of overlaps and replications occur in the sections “missing data mechanisms” and “multiple imputation method” in the different chapters. This problem is negligible when one considers the critical peer review process and the fact that the different chapters are papers that can be read separately without losing the overall context. In Chapters 2, 3, 4 and 5 in particular, the MAR mechanism in conjunction with ignorability is the central idea, while in Chapters 6 and 7, the main focus is on MNAR-based models. While most chapters are stand-alone, the study is easier to follow and understand if read in the order that it is presented. Chapters 3 and 4 can be read in any order. However, Chapters 6 and 7 should be read in their sequential order. Lastly, Chapter 8 should be read when all the other chapters have been read as it summarizes the findings. A brief outline follows:

Chapter 1: This chapter serves as an introduction to the study.

Chapter 2: This chapter focuses on missing covariates in regression analysis, with particular focus on monotone missingness pattern. The patterns and mechanisms of missingness for continuous data are presented. The methods of imputation are reviewed. The imputation-based methods that were considered are: Marckov chain monte carlo (MCMC), regression, propensity score (PS) and last observation carried forward (LOCF). For each method a brief literature review is given as well as the description of the multiple imputation for continuous data. An application study including a description of the full data set used in the analysis is presented. The design of the comparative study used in the analysis is discussed in detail. The results from the application of the different methods are contrasted. Finally, the chapter concludes with a discussion of the results.

Chapter 3: In this chapter, the notation and concepts of mechanisms that lead to the dropout process are presented. The two approaches mentioned above, namely multiple imputation (MI) and inverse probability weighting (IPW) are then considered in more detail as principled approaches to be used in the analysis. In addition, the chapter contains the design of the application study and offers a description of the data set used in the analysis in detail. The results for the MI and IPW methods based on the generated dropout of our application study are set out analogous to each other in terms of bias and efficiency criteria. Lastly, the chapter ends up with a discussion of the findings.

Chapter 4: This chapter investigates the performance of MI and direct likelihood methods in handling incomplete Gaussian longitudinal data when there are MAR dropouts. The chapter begins with the data structure and dropout mechanisms including a formal framework for incomplete longitudinal data. An overview of the methods to deal with dropout is given. The chapter also presents an application and a description of the full data set used in the analysis and the study design. The dropout generation schemes are discussed. A data set consisting of the heart rate used as illustration material in this chapter is presented. The results from generated data are compared and contrasted to results based on actual data for detecting inference. Discussion and conclusions are drawn from the comparison.

Chapter 5: This chapter focuses on techniques that are based on both likelihood-based inference and non-likelihood inference for analyzing non-Gaussian longitudinal outcomes subject to potentially random dropout. Data setting and necessary notation in terms of the random dropout assumptions are introduced. An overview of the methods for analyzing incomplete longitudinal non-Gaussian data with the focus on the weighted generalized estimating equations (WGEE), multiple imputation after generalized estimating equations (MI-GEE) and generalized linear mixed model (GLMM) strategies, is given. The simulation study including the design, data generation and evaluation criteria used in the analysis are presented. The results of the simulation from the different methods are compared. The chapter ends with a brief discussion and concluding remarks.

Chapter 6: In this chapter, the data setting and modelling framework with the emphasis on dropout mechanisms are described. A background to the selection model is provided, followed by descriptions of the selection model based on Diggle and Kenward framework and detailed discussion of the linear mixed model and dropout model. An application example including a description of the longitudinal data set in the form of a multi-centre clinical trial data used in the analysis is presented. In the current application, the implementation code in the original model by Diggle and Kenward (1994) has been extended to the case of three treatment arms. The results of the estimation of the selection model are then described. In conclusion, a discussion of the results is given.

Chapter 7: This chapter deals with non-ignorable missingness due to data MNAR requiring the use of sensitivity analysis. This approach is needed because it is not obvious to distinguish between MAR and MNAR as for every MNAR model there corresponds a MAR counterpart. The only difference being that they differ in the manner of prediction of the missing data. The chapter describes necessary notation and concepts regarding modelling incompleteness as well as the general framework of longitudinal data. Two families of models for modelling Gaussian longitudinal data with incomplete measurements are discussed: selection and pattern mixture models. An application study is carried out using a longitudinal clinical trial data set with continuous outcomes. This is followed by a discussion of an application scheme to be used in the analysis. Aggregate results are obtained under the two restriction models. Based on the application results, the chapter ends with a summary of the key points.

Chapter 8: Finally, this chapter gives a synthesis of the study. The findings are summarized and conclusions are derived from the preceding chapters. For future work on the applications of incomplete data analysis, relevant recommendations are made. Special focus is directed towards the operational use of sensitivity analysis in non-ignorable missing data mechanism.

A single reference list is given at the end of the thesis.

Chapter 2

Imputation methods for estimating regression parameters under a monotone missing covariate pattern: A comparative analysis*

2.1 Abstract

The chapter deals with the problem of missing covariate values in a regression model. An attractive approach to avoid this problem is to impute the missing covariate values rather than delete cases with missing covariates. The study is devoted to a comparison of different imputation techniques or methods. The type of missing data pattern in the set of independent variables is the monotone data pattern. We assume that the missing data are missing at random (MAR). Imputation of missing values was achieved using a multivariate normal model. The main objective of this chapter is to study how some imputation approaches compare when the missing data pattern is monotone. The techniques that were considered include the Last Observation Carried Forward (LOCF),

[*] A. Satty and H. Mwambi (2012). Imputation methods for estimating regression parameters under a monotone missing covariate pattern: A comparative analysis. *South African Statistical Journal*, 46, 327-356.

Propensity Score (PS), Markov Chain Monte Carlo (MCMC) and Regression. In order to compare the performance of the proposed methods, we used originally complete data sets (no data are missing), and then we intentionally created missing values to achieve the intended goal. Missingness was imposed on covariate variables. The performance of these four methods is compared on three criteria: bias, efficiency and coverage. Data from a diabetes study are used to illustrate the considered approaches.

Keywords: Missing covariates, Monotone missing data pattern, Multiple imputation, MCMC, Propensity score, LOCF.

2.2 Introduction

The occurrence of missing data in scientific investigations is quite common. Data analysis is seldom performed in the absence of missing data which brings to question efficiency of estimates, validity of analysis and bias due to differences between observed and unobserved data (Little, 1992). To this end, considerable research has recently focused on finding appropriate procedures for handling missing data, see, for example, Zhang (2003), Chen (2004), Horton and Kleinman (2007), Peng and Zhu (2008) and Chen (2002). One approach for handling incomplete data problems that addresses these concerns is imputation which was proposed by Rubin (1978) and described in detail by Little and Rubin (1987), Little and Rubin (2002) and Schafer (1997). Schafer (1999) provides extensive bibliography on missing data imputation approaches.

Regression analysis users are often faced with the problem of dealing with a number of missing values in one or more explanatory variables. The easy solution is often to discard the cases of the missing value variables in the model and to confine attention to complete case analysis. However, this solution, although simple and straightforward, usually leads to adverse statistical consequences and is surely not an efficient strategy. An alternative more efficient approach is to impute new values that replace the missing ones before carrying out the regression analysis. See, for example, Little (1976) for basic considerations and Little (1992) for detailed discussion of regression analysis with imputed values for missing data. The reasons for missing data are commonly called missing data mechanisms. According to Rubin (1976) and Little and Rubin (1987), the missing data mechanisms can be classified as: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). The mechanisms are discussed briefly in later sections.

This chapter focuses on comparison of several imputation techniques applied to an incomplete data set under MAR. We focus on the problem of missing covariate values in a regression model. Imputation of missing values was achieved using a multivariate normal model. The imputation techniques that will be compared include the controversial Last Observation Carried Forward (LOCF) method and the more sound multiple imputation methods implemented in the SAS software PROC MI (Markov Chain Monte Carlo (MCMC), Regression and Propensity Score (PS)). The main objective of this chapter is to study how some imputation approaches compare when the missing data pattern is monotone. In the application, a monotone missing data pattern was created from complete data. The results from complete data were presented and used as a references against which these four techniques were contrasted. Data from a diabetes study is used to illustrate the considered approaches. This data is a cross-sectional study. MATLAB and SAS programs were used to create the missing data and to implement imputation techniques, respectively. A detailed discussion on the computer programmes used for implementing missing value imputations can be found in Horton and Lipsitz (2001) with particular focus on the REG, MI and MIANALYZE procedures in SAS.

The rest of this chapter is organized as follow: In Section 2.3 we consider the missing data patterns and missing data mechanisms. In Section 2.4, we review the methods of imputation to be used: LOCF, Regression, Propensity score and MCMC. For each method a brief literature review is given, as well as the description of the multiple imputation for continuous data. In Section 2.5, we present an application including a description of the full data set used in the analysis. The design of the comparative study used in the analysis is discussed in details. The results from the application of the different methods are presented in Section 2.6. Finally, in Section 2.7, we conclude by a discussion of the results.

2.3 Missing data patterns and classifications

2.3.1 Missing data patterns

An important concept with missing data, specifically where there are multiple variables with missing values, relates to the pattern of missing data. There is a need to identify the pattern of the missing data because some methods apply to missing data in general, while others are restricted to a specific kind of pattern. In this chapter, we focus on the monotone missing pattern as defined by Anderson (1957), Rubin (1974) and Little and Rubin (1987). In their definition, a data matrix is said to have a monotone missing pattern if, whenever an element y_{ij} is missing, the element y_{ik} are also missing for all $k > j$. The notation y_{ij} means the ith observation for variable j, where $i = 1, ..., n$ and $j = 1, ..., p$. For example, given variables Y_1, Y_2 and Y_3, a data set is said to have a monotone missing patterns if the missing information in the variables is ordered in some specific way. For illustration, Table 2.1 represents variables Y_1, Y_2 and Y_3 with observations 1, 2,...,8 assuming n=8, where an X denotes available data and a dot (.) denotes a missing data value. When an element in Y_2 is missing, so is the corresponding element in Y_3, leading to a monotone missing data pattern (Rubin, 1974). In monotone missing pattern the ordering of variables is important.

Table 2.1: *Monotone and non-monotone pattern of missingness*

	Monotone			Non-monotone		
Obs	Y_1	Y_2	Y_3	Y_1	Y_2	Y_3
1	X	X	X	X	.	X
2	X	.	.	X	.	.
3	X	.	.	X	.	.
4	X	.	.	X	.	.
5	X	X	.	X	X	.
6	X	X	.	X	X	.
7	X	X	.	X	X	.
8	X	X	X	X	X	X

Simpler imputation methods can be used if the pattern is monotone, though a monotone pattern is uncommon in most complex investigations. It may be possible however to create a monotone missingness pattern that separates out a small number of observations that are non-monotone. In the non-monotone example in Table 2.1, all but the first observation can be rearranged into a monotone pattern (i.e., 81% of the data set). This type of hybrid pattern is exploited by a number of computer packages. It is important to note that, in many realistic settings, data sets may have missing outcomes as well as missing predictors (Chen, 2004).

2.3.2 Missing data classifications

As mentioned above, in this subsection we discuss three types of missing data mechanisms that are used extensively to refer to missing data in literature. These refer to mechanisms that could have generated the missing data. When a data set has missing values, the difficulty of obtaining valid parameter estimates depends on the mechanism that causes values to be missing. A useful classification for missing value mechanisms was introduced by Little and Rubin (1987). More formally, let Y be an $(n \times p)$ data matrix, $Y = (y_1, y_2, ..., y_n)^T$, where $y_i = (y_{i1}, ..., y_{ip})^T$ is a random sample from a p-dimensional multivariate probability distribution $P(Y \mid \theta)$ governed by parameters θ. We refer to the rows of Y as observations, given by $y_i^T (i = 1, 2, ..., n)$, and the columns of Y as variables, denoted by $Y_j (j = 1, ..., p)$. We now define an $(n \times p)$ missingness indicator matrix $R = r_{ij}$ as follows: r_{ij} equal 1 when Y_{ij} is missing, and equals 0, if not. Defining $P(r_{ij} = 0 \mid y_{ij}) = P(y_{ij} \; observed \mid y_{ij}) = p_{ij}$, then R is subject to a probability distribution $P(R \mid \psi, Y)$ governed by parameter ψ. The joint probability distribution of the response variables and the missing data indicator variables can be expressed as

$$P(Y, R \mid \theta, \psi) = P(Y \mid \theta) P(R \mid \psi, Y), \tag{2.1}$$

where $P(Y \mid \theta)$ and $P(R \mid \psi, Y)$ denote the marginal distribution of the response variables and the conditional distribution of missing data, conditional on the response variables, respectively. Following Little and Rubin (1987) and Rubin (1987), we assume Y_{obs} and Y_{mis} represent the observed and the missing portions of Y, respectively;

that is, $Y_{obs} = (y_{ij} \mid r_{ij} = 0)$ and $Y_{mis} = (y_{ij} \mid r_{ij} = 1)$. We additionally assume $Y_{obs,j}$ and $Y_{mis,j}$ denote the observed and missing portions of variable Y_j, and assume $y_{i(obs)}$ and $y_{i(mis)}$ denote the observed and missing portions of the ith observation. The probability model (2.1) has two sets of parameters (θ and ψ) representing the parameters of interest and the nuisance or missing data parameters, respectively. In model (2.1), the correct inferences on θ in general need to be conducted. Moreover, this inference depends on how the probability model for the missing data is defined. In other words, how the missingness process depends on the full data, Y. Based on the conditional distribution $P(R \mid \psi, Y)$, Rubin (1976, 1987) classified the missing data mechanisms into the following three categories:

- If the missingness process is independent of the responses (observed and missing), i.e., $P(R \mid \psi, (Y_{obs}, Y_{mis})) = P(R \mid \psi)$, then the missing data mechanism is defined as MCAR.

- If the missingness process is independent of the missing responses given the observed values, i.e., $P(R \mid \psi, (Y_{obs}, Y_{mis})) = P(R \mid \psi, Y_{obs})$, then the missing data mechanism is defined as MAR.

- If the missingness process depends on both observed and missing responses, i.e., $P(R \mid \psi, (Y_{obs}, Y_{mis})) \neq P(R \mid \psi, Y_{obs})$, then the missing data mechanism is defined as MNAR.

Note that the parameterization of ψ is not expected to be the same under the three missingness categories. According to Zhang (2003) and based on the standard definitions of Rubin (1987) and Madow et al. (1983), the two sets of parameters in model (2.1) (i.e., θ and ψ) are said to be distinct if from a: 1) Frequentist perspective, the joint parameter space of (θ, ψ) is the Cartesian cross-product of the parameter spaces for θ and ψ. 2) Bayesian perspective, the joint prior distribution of (θ, ψ) can be factored into the independent marginal prior distributions for θ and ψ. In many situations, this assumption is intuitively reasonable as knowing ψ provide no information about θ and vice versa. When θ and ψ are distinct, and under either MCAR or MAR, the missing data mechanism is deemed *ignorable* (Little and Rubin, 1987; Rubin, 1976, 1987), which is to say that the missing data mechanism can be ignored when making likelihood-based or Bayesian statistical inferences on the parameters of interest θ. In contrast, when the missing data mechanism does not satisfy this definition, the process is called *non-ignorable*, and the data is said to be MNAR.

2.4 Methods to handle missing covariates

There are a variety of imputation methods that can be used to deal with missing covariates. The subsections that follow provide a review of the methods that are used in this study.

2.4.1 Last observation carried forward (LOCF)

The LOCF method is a well-known, commonly used imputation technique. It retains all of the originally randomized subjects, eliminates missing data and produces a completed data set. In this approach every missing value is replaced by the last observed value from the same subject or time series, i.e., it is a method that assumes that the outcomes would not have changed from the last observed value. Very strong and unrealistic assumptions have to be made to ensure the validity of this method. LOCF analysis appeals through its simplicity and ease of application, but there are strong grounds for not using it. Specifically, the method may introduce bias in the results, and this bias can, according to circumstance, be in either direction (Molnar et al., 2008). The technique does not itself indicate a particular type of analysis for the resulting data. Molenberghs and Kenward (2007) noted that LOCF can be applied to both patterns of missing data (monotone and non-monotone). We refer to Craig et al. (2003) and Siddiqui and Ali (1998) for more details, and where insightful illustrations of the issues of this method are provided in Molenberghs and Kenward (2007).

2.4.2 Multiple imputation for continuous data (MI)

MI was first proposed by Rubin (1978), and was originally developed as an alternative to earlier single imputation approaches by Madow et al. (1983). It has also been discussed thoroughly elsewhere, see, for example, Schafer (1997), Schafer (1999), Schafer and Graham (2002), Sinharay et al. (2001) and Zhang (2003). MI refers to a procedure of replacing each missing value by a vector of $M \geq 2$ imputed values selected to reflect the uncertainty in the imputation (Rubin, 1987). The technique accounts for uncertainty in sampling from a population by introducing randomness to imputations and creating M imputed data sets, each of which is then subjected to the desired statistical analysis (for example, regression analysis). MI incorporates information from other variables into the imputation process in order to provide more accurate values. The MI method as is used in most common implementations (for example, PROC MI in SAS software) assumes that the method is valid under MAR. Rubin (1987) described multiple imputation as a three step process. First, sets of plausible values for missing observations are created that reflect uncertainty about the stochastic non-response model. Each of these sets of plausible values can be used to fill-in the missing values and create a completed data set. Second, each of these completed data sets are analysed using standard complete-data methods. Finally, the M results are combined using methods that allow for uncertainty regarding the imputation to be taken into account.

For continuous data, multiple imputation requires an assumption that the method is valid under a multivariate normal distribution. Based on this probability model, parameter estimates are obtained using the Bayesian posterior distribution based upon the likelihood function of the proposed model, the observed data and a prior distribution. The Markov Chain Monte Carlo (MCMC) method of data augmentation is used to obtain this posterior distribution from which the imputed values for the missing observation are drawn. This imputation process is repeated M times to create independent data sets, each of which is then subjected to the analysis of interest, such as the regression analysis as in this study. The results of the M separate analyses are then combined into a single value as

$$\bar{Q} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m, \tag{2.2}$$

where $\hat{\theta}_m$ is the parameter estimate of interest from imputation m=1, 2,...,M. The variance for these estimates is composed of two parts: the between imputation variance and within imputation variance. Between imputation variance takes the form

$$B = \frac{1}{M-1} \sum_{m=1}^{M} (\hat{\theta}_m - \bar{Q})^2. \tag{2.3}$$

The within imputation variance, $\bar{U}$, is the mean of estimated variances across the M imputations. The total variance for MI is then calculated as

$$T = \bar{U} + \left(1 + \frac{1}{M}\right) B. \tag{2.4}$$

In this study, we consider only the multiple imputation methods that are implemented in SAS software PROC MI. The following three methods are available in the MI procedure:

2.4.2.1 Markov chain monte carlo (MCMC)

The MCMC method is a Monte Carlo integration method using Markov Chains (Zhang, 2003). It has been applied in many statistical situations, see, Gilks et al. (1996). In the context of multiple imputation techniques, Schafer (1997) applied the MCMC method by utilizing the data augmentation algorithm developed by Tanner and Wong (1987). This method is the most popular multiple imputation method for handling missing data. This is the default method in SAS PROC MI. The technique is valid under the assumption of multivariate normality (Robins et al., 1994) which implies that valid imputations may be generated by linear regression equations. Next, we follow the description provided by Zhang (2003) in formulating the MCMC approach thereby illustrating how MCMC can be conducted to impute the missing data. The MCMC method is based on draws of pseudo random samples from a target probability distribution. In the presence of non-monotone missing data pattern, the target distribution is the joint conditional distribution of Y_{mis} and θ given Y_{obs},

$$P(Y_{mis}, \theta \mid Y_{obs}). \tag{2.5}$$

The MCMC method imputes the missing data as follows: Replace the missing data Y_{mis} by some assumed values, then θ can be simulated from the resulting complete data posterior distribution $P(\theta \mid Y_{obs}, Y_{mis})$. If $\theta^{(t)}$ is the current simulated value of θ

from the complete data posterior distribution, then the next iterative sample of Y_{mis}, $Y_{mis}^{(t+1)}$, can be drawn from the conditional predictive distribution of Y_{mis} given Y_{obs} and $\theta^{(t)}$, i.e.,

$$Y_{mis}^{(t+1)} \sim P(Y_{mis} \mid Y_{obs}, \theta^{(t)}). \tag{2.6}$$

The next simulated value of θ can be drawn from its complete data posterior distribution, conditional on the $Y_{mis}^{(t+1)}$,

$$\theta^{(t+1)} \sim P(\theta \mid Y_{obs}, Y_{mis}^{(t+1)}). \tag{2.7}$$

Repeating the random draws from a starting value of $\theta^{(0)}$ using Equations (2.6) and (2.7) yields a Markov chain $\{(\theta^t, Y_{mis}^{(t)}) : t = 1, 2, ...\}$. Therefore, the stationary distribution of this chain is the joint distribution of θ and Y_{mis} given Y_{obs}, $P(Y_{mis}, \theta \mid Y_{obs})$. Consequently, the marginal stationary distribution of the subsequence $\{\theta^{(t)} : t = 1, 2, ...\}$ is the observed-data posterior distribution $P(\theta \mid Y_{obs})$, and the marginal stationary distribution of the subsequence $\{Y_{mis}^{(t)} : t = 1, 2, ...\}$ is the posterior predictive distribution $P(Y_{mis} \mid Y_{obs})$. As stated by Zhang (2003), when t is sufficiently large, $\theta^{(t)}$ can be viewed as a single simulation from the observed data posterior distribution $P(\theta \mid Y_{obs})$, and $Y_{mis}^{(t)}$ can be viewed as a single imputation from the posterior predictive distribution $P(Y_{mis} \mid Y_{obs})$. The random draw of Equation (2.6) is used to impute the missing data Y_{mis}, and the random draw of Equation (2.7) is used to simulate the unknown parameter θ. Therefore, according to Tanner and Wong (1987), Equations (2.6) and (2.7) denote the Imputation or I-step and the Posterior predictive or P-step, respectively. This algorithm was first used by Li (1988) who presented an argument for convergence and used it to impute the missing data Y_{mis}.

2.4.2.2 Regression method

A regression model is fitted for each variable with missing data, using the remaining variables as covariates (Yuan, 2000). Based on the fitted regression coefficients, a new regression model is formulated from a Bayesian predictive distribution of the parameters (regression parameter estimates and associated covariance matrix) which is then used to impute the missing values for each variable (Rubin, 1987). Let Y_j be a continuous variable with missing values, satisfying the model

$$E(Y_j) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k, \tag{2.8}$$

which is fitted using observations with observed values for the variable Y_j and its covariates $X_1, X_2, X_3, ...X_k$, where k is the remaining numbers of variables, and $0 < k < p$. According to Rubin (1987), this method assumes multivariate normality, therefore, the above model imply a Normal error model with mean 0 and variance σ_j^2. The fitted model provides the regression parameter estimates $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k)'$ and the associated covariance matrix $\hat{\sigma}_j^2 V_j$, where V_j is the usual $(X'X)^{-1}$ matrix derived from the intercept and covariates $X_1, X_2, ..., X_k$.

Steps for generating imputed values for each variable:

Step 1: New parameters $\beta_* = (\beta_{*0}, \beta_{*1}, ..., \beta_{*(k)})$ and σ_{*j}^2 are drawn from the posterior predictive distribution of the parameters. That is, they are simulated from $(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k)$, σ_j^2 and V_j. The variance is drawn as

$$\sigma_{*j}^2 = \hat{\sigma}_j^2 (n_j - k - 1)/g, \tag{2.9}$$

where g is a $X_{n_j-k-1}^2$ random variate and n_j is the number of non-missing observations for Y_j. The regression coefficients are drawn as

$$\beta_* = \hat{\beta} + \sigma_{*j} V'_{h_j} Z, \tag{2.10}$$

where V_{hj} is the upper triangular matrix in the cholesky decomposition $V_j = V'_{hj} V_{hj}$ and Z is a vector of $k+1$ independent random standard normal variates.

Step 2: The missing values are then replaced by

$\beta_{*0} + \beta_{*1} x_1 + \beta_{*2} x_2 + ... + \beta_{*(k)} x_k + z_j \sigma_{*j}$,

where $x_1, x_2, ..., x_k$ are the values of the covariates and z_j is a simulated standard normal deviate.

2.4.2.3 Propensity score method (PS)

Another imputation method available for continuous variables is the propensity score (PS) method. The PS method for multiple imputations was proposed by Lavori et al. (1995). In fact, the PS method follows a nonparametric approach in which the missing data are imputed by resampling of the observed data (Zhang, 2003). The method is also valid under the assumption that the data set has a monotone missing pattern. In medical research it can be defined as the conditional probability of assignment to a particular treatment given a vector of observed covariates (Rosenbaum and Rubin, 1983). For each variable with missing values, propensity scores are generated for all observations to estimate the probabilities that each observation is missing. For each group Bayesian imputation technique is applied.

To discuss the PS method in detail, we will follow the approach provided by Zhang (2003). As we defined the missingness indicator variables r_{ij}, the missingness process can be applied by a linear logistic regression model. Therefore, the conditional probability of observing y_{ij}, conditional on the previous history $y_{i1}, ..., y_{i,j-1}$, can be called a propensity score, s_{ij}. According to Rosenbaum and Rubin (1983), definition of the propensity score can be expressed as follows

$$s_{ij} = Pr(r_{ij} = 0 \mid y_{i1}, ..., y_{i,j-1}). \tag{2.11}$$

In the presence of monotone missing data pattern, the propensity score can be modelled by

$$log\left(\frac{s_{ij}}{1 - s_{ij}}\right) = \beta_0 + \beta_1 y_{i1} + ... + \beta_{j-1} y_{i,j-1}, \tag{2.12}$$

where β_0, β_1,...,β_{j-1} are the regression coefficients. Now, based on Equation (2.12) and after the regression coefficients are estimated from the observed r_{ij} for the response variable Y_j and the complete data for the covariates $Y_1, ..., Y_{j-1}$, each observation can be assigned an estimated propensity score,

$$\hat{s}_{ij} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 y_{i1} + ... + \hat{\beta}_{j-1} y_{i,j-1})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 y_{i1} + ... + \hat{\beta}_{j-1} y_{i,j-1})}, \tag{2.13}$$

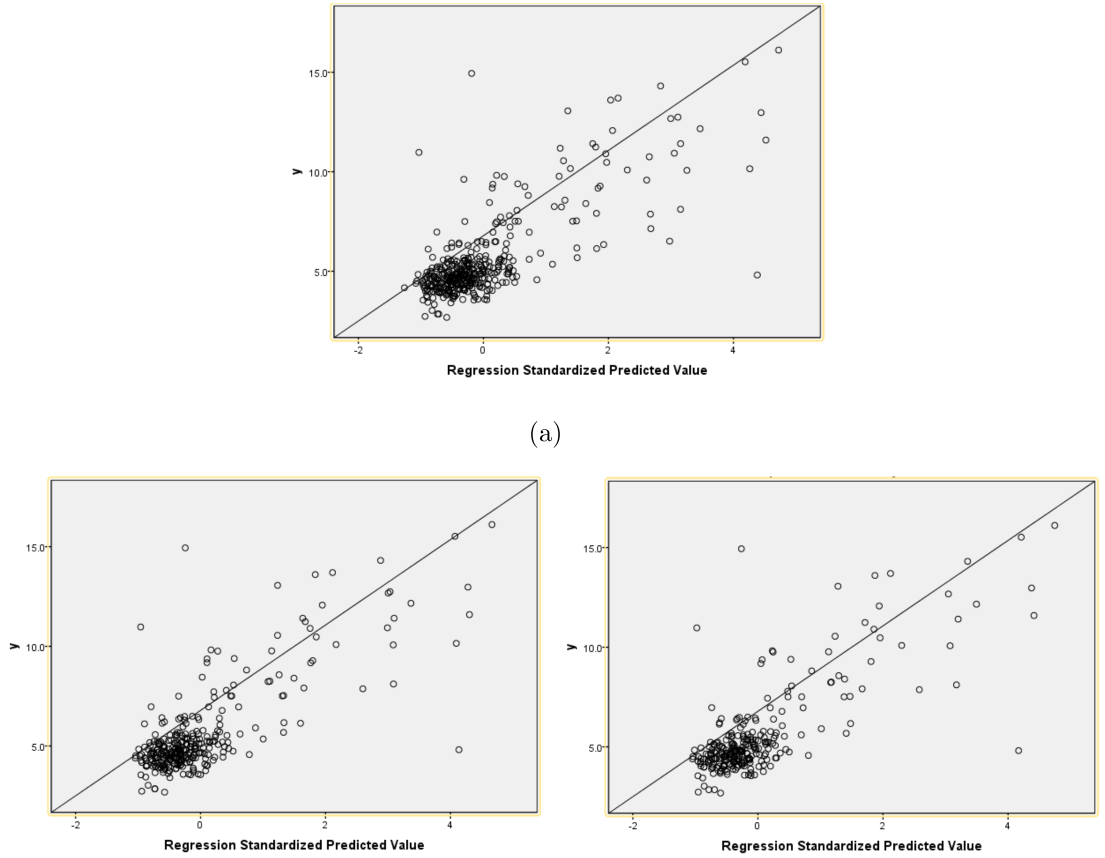
and then all observations are stratified into q strata based on the quantiles of estimated propensity scores. Within each stratum, a “donor pool” is created by applying the approximate Bayesian bootstrap (Rubin and Schafer, 1986). That is, a random sample is created by random draws with replacement from the observed values of Y_j within each stratum. The approximate Bayesian bootstrap method is applied in order to reflect additional uncertainty about the posterior distribution of the underlying parameters, given the observed values of Y_j within each stratum. As noted by Zhang (2003), the

MCMC method is roughly equivalent to choosing the values of the parameters θ for the conditional posterior predictive distribution $P(Y_{mis} \mid Y_{obs}, \theta)$ from the observed-data posterior distribution $P(\theta \mid Y_{obs})$. After the donor pools are created, each missing value of Y_j is then imputed by a single random draw from its donor pool. M sets of multiple imputations are obtained by creating M conditionally independent donor pools for each individual missing value and then taking a random draw from each donor pool. Note that imputing a missing value by a random draw from its stratum rather than from its donor pool would result in an improper multiple imputation in the sense that the between imputation variance would be underestimated because the uncertainty due to selecting the imputation model is not incorporated into the imputation.

2.5 Application

2.5.1 Example: Diabetes data

We consider a response variable Y that is fully observed, while all the explanatory variables $X_1, X_2, ..., X_p$ contain missing values. Let X be the design or model matrix of dimension $(n \times p)$, while Y represents the vector of response values of length n. The vectors $(Y_i, X_{i1}, ..., X_{ip})$ for i =1,...,n are independent. Due to missing observations, the design matrix can be split into two parts, $X = (X_{obs}, X_{mis})$; X_{obs} represents the part of the design matrix X with covariates that are completely observed, and X_{mis} is the subset of X with explanatory variables which have at least one value that is not observed. We assume that X_{mis} follows a monotone pattern. To illustrate this application we describe a medical data set as reported by Willems et al. (1997). The original (diabetes data set example) database consists of 19 variables on 403 subjects out of 1040 subjects who were originally interviewed in a study to understand the prevalence of obesity, diabetes and other cardiovascular risk factors in Central Virginia for African Americans. The 403 subjects were the ones who were actually screened for diabetes. At this point we have to make it clear that our present study is essentially a cross-sectional study. We compare the various imputation methods using generated missing values in a complete subset of the whole database. So, our initial strategy was



(a)

(b)

(c)

Figure 2.1: *(a) Scatter plot - complete data set. (b) Scatter plot - 20% missing covariates. (c) Scatter plot - 30% missing covariates*

to find a data set with no missing values on certain important continuous variables and then to use it for our study. We thus ended up with a complete data set with 403 subjects. The variables which we selected are: Y - glycosolated hemoglobin, X_1 - age, X_2 - cholestrol, X_3 - high density lipoprotin, X_4 - cholestrol/HDL ratio and X_5 - stabilized glucose. We use a linear regression model for the response variable, Y, as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon, \tag{2.14}$$

where $\epsilon \sim$ *Normal* $(0,\, \sigma^2)$, therefore,

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + +\beta_4 X_4 + +\beta_5 X_5, \tag{2.15}$$

where $E(Y)$ denotes the expectation of Y for given $X_1, ..., X_5$. The model assumptions also imply

$$\epsilon = Y - E(Y \mid X_1, X_2, X_3, X_4, X_5) \sim \mathit{Normal}\,(0, \sigma^2), \tag{2.16}$$

where glycosolated hemoglobin (Y) was used as a surrogate outcome for diabetes because of its known high predictive capacity with respect to type two diabetes. In this application, all covariates ($X_1 - X_5$) were subjected to missing observations while Y was left completely observed for all cases. In order to create a monotone missing data pattern, we assumed that if a datum is missing, all subsequent data on that observation is also missing. Thus, the incomplete data set is constructed from the complete data set and the process yields a monotone missing pattern in variables X_1-X_5. Figure 2.1 represents the relationships between the outcome Y and regression standardized predicted values of Y before and after creating missing observations.

2.5.2 Study design

To investigate the performance of the proposed methods, five steps were planned and executed as follows:

Step 1: We fit a linear regression model using the complete data set with the Y as the dependent variable and $X_1 - X_5$ as explanatory variables to obtain the regression analysis coefficients.

Step 2: From the complete data set, we draw 1000 random samples of $N = 400$. We create a monotone missing data pattern in $X_1 - X_5$. The monotone pattern of missingness was created by randomly selecting values from the data matrix, then the pattern was created by discarding values that lie next to the selected ones. Missing data for $X_1 - X_5$ are generated assuming MAR mechanism. To achieve MAR mechanism, we assume the missing data is related to the observed data (i.e., past responses). Missing data were created in X_1 by randomly deleting 20% and 30% of observations greater than 55, resulting in monotone missingness pattern in incomplete data. A monotone missing pattern was used which is to say that for X_1, if a X_1's observation was deleted, the subsequent observations in the other variables for that subject were also deleted. To generate the missing data, we used a function written in the MATLAB software package.

Step 3: We fill in the missing values in $X_1 - X_5$ to generate completed data sets under a multivariate normal model using imputation methods described above. For multiple imputation techniques, we used the SAS software. The MI procedure options were set so that the program used up to 1000 iterations of the EM method to find the posterior mode. We used the defaults of the SAS MI algorithm: a Jeffreys prior and initial sampler values from the EM posterior mode. Then, the method uses Markov-chain Monte Carlo (MCMC) to generate the actual imputations. PROC MI was applied to generate $M = 5$ complete data sets (conventional wisdom suggests taking $M = 5$, see, for example, Schafer, 1999). PROC REG was used to set up effect parameterizations for the variables and the BY statement is also used to allow the analysis to be repeated for each data set. The simpler LOCF technique replaced the missing data by the last available observed data, and once the data set has been completed in this way, it is analysed as if it were fully observed.

Step 4: We re-estimate the regression model using each imputed data set separately and we record the results of interest. For the MI techniques, we combine the results of the 5 imputations using formulas that account for variation within and between the imputed data sets using Equations (2.2), (2.3) and (2.4). We use PROC MIANALYZE to combine the estimates from the 5 completed data sets to generate valid statistical inferences about the regression model.

Step 5: Finally, we compare the results of imputation techniques from the fourth step with the results of the true complete data set from the first step to assess the performance of the proposed methods.

2.5.3 Criteria of performance

We assess the performance of the methods largely using the criteria recommended in Schafer and Graham (2002) and Peng and Zhu (2008), namely bias, efficiency and coverage. In this chapter, we defined bias as the difference between the average of the 1000 coefficient estimates and the corresponding true coefficient. Thus, a better approach that does on the average presents a population value with less bias. We defined the efficiency as the variability of the estimates around the true population coefficient. It was measured in this study by the average width of the 95% confidence interval. Thus, a wider interval implies a less efficient technique. Coverage was defined as the percentage of 95% confidence intervals estimates across 1000 replicates. If a method is working well, the actual coverage should be close to the nominal rate (95%).

2.6 Results

Table 2.2 shows the overall performance of the imputation techniques considered in this study when the missing data rate was 20%. For all covariates the greatest bias, also the worst, is highlighted. Examining these results we find the following. When compared with the results based on the PS and LOCF methods in Table 2.2, the MCMC and Regression methods offered better performance than the PS and LOCF methods. Covariates (X_1-X_5), were more accurately estimated by MCMC and Regression methods as compared to those of PS and LOCF. In particular, the regression coefficient for the covariate X_3 was unbiased for MCMC and Regression methods. Furthermore, the MCMC and Regression methods yielded the same estimates for all covariates, except for X_4. Their respective estimates did not differ significantly from those of the true actual data. Differences were never more than 0.003, with one exception - the estimate of X_4 which was slightly different than that from true complete data. Both, PS and LOCF contained more biased estimates as compared to MCMC and Regression methods of imputation used in this study. Covariates which showed most bias were X_5 under the PS method and (X_1, X_2, X_3 and X_4) under the LOCF method. Since a wider interval implied a less efficient, thus the widest also implies the worst, 95% is highlighted. The LOCF was less efficient most frequently. This should not be a surprise because intuitively the LOCF's weakness is that it tends to create inflated artificial values than truly expected. Between the two LOCF and PS methods, the LOCF was less efficient than the PS. The exception was with the estimate of X_1 for the PS. For X_1-X_5, MCMC and regression methods uniformly the best approaches in terms of efficiency. Therefore, both MCMC and Regression methods were more efficient than PS and LOCF. In terms of coverage, according to Schafer and Graham (2002), the

Table 2.2: *Bias, Efficiency and Coverage of MCMC, Regression, PS and LOCF, under 20% missing covariates*

					coverage
Rate	Method	Parameter	Bias	Efficiency	true level=95%
	MCMC				
		X_1	0.002	0.020	0.972
		X_2	-0.003	0.011	0.978
		X_3	0.000	-0.001	0.991
		X_4	0.042	0.120	0.967
		X_5	0.001	0.007	0.991
	Regression				
		X_1	0.002	0.021	0.973
		X_2	-0.003	0.012	0.978
		X_3	0.000	-0.004	0.992
		X_4	0.061	0.120	0.961
		X_5	0.001	0.007	0.991
	PS				
		X_1	-0.005	**0.030**	0.954
		X_2	-0.003	0.012	0.968
		X_3	0.008	-0.014	0.951
		X_4	-0.139	0.121	0.942
		X_5	0.004	0.013	0.966
	LOCF				
		X_1	**-0.011**	**0.030**	0.954
		X_2	**0.016**	**0.035**	0.941
		X_3	**-0.043**	**-0.113**	**0.897**
		X_4	**-0.630**	**0.159**	**0.784**
		X_5	**-0.678**	**0.211**	**0.677**

Note: The largest bias and efficiency for each given estimate appear in bold. PS=propensity score; LOCF=last observation carried forward; MCMC=markov chain monte carlo.

performance of a method can be regarded to be poor if its coverage drops below 90% and hence leads to substantially increased Type-I error rate. By this rationale, the MCMC, Regression and PS methods yielded equally acceptable performance across all covariates. The LOCF's coverage at 95% was consistently lower than 90%, except for X_1 and X_2, thus, this coverage was indicative a seriously low level of coverage because 90% corresponds to a doubling of the nominal rate of error (0.05).

With 30% missingness, the results from the performance of the four methods are listed in Table 2.3. For all four methods, the MCMC and Regression methods yielded equally good performance and outperformed the PS and LOCF methods. In terms of bias condition, the benefits of MCMC and Regression methods over a PS and LOCF are clearly evident. In particular, the estimates from PS and LOCF for X_1 and (X_2, X_3, X_4 and X_5), respectively, contain seriously biased estimates. Both MCMC and Regression methods for X_1, X_2 and X_5 yielded the same estimates of regression coefficients. Thus, while there appeared to have been little or no bias, they were more efficient than those obtained from the PS and LOCF. An examination of the efficiency suggested that the estimates from MCMC and Regression methods were typically lower than those from the PS and LOCF methods, thus such estimates were more efficient than were those based on the PS and LOCF methods. Efficiency by PS and LOCF approaches appeared to be independent of the missingness rate. The worst performance on efficiency occured with estimates of covariates X_1 and X_2 for PS, and with estimates of covariates X_3, X_4 and X_5 for LOCF when compared with those based on MCMC and Regression methods. Consequently, both methods, being asymptotically less efficient. In comparison with results for PS, LOCF's efficiency was worst than that of PS. With respect to coverage, similar to the findings obtained under 20% missing data, the MCMC, Regression and PS method produced uniformly acceptable coverage; none was less than 90%. On the other hand, coverage rates obtained by the LOCF method in most cases were unsatisfactory with two exceptions - coverage rates for X_1 and X_2. It is appeared that its low coverage rates can also be attributed to its large biases.

Table 2.3: *Bias, Efficiency and Coverage of MCMC, Regression, PS and LOCF, under 30% missing covariates*

					coverage
Rate	Method	Parameter	Bias	Efficiency	true level=95%
	MCMC				
		X$_1$	0.006	0.021	0.982
		X$_2$	-0.001	0.008	0.984
		X$_3$	0.001	0.003	0.977
		X$_4$	-0.059	0.361	0.955
		X$_5$	0.001	0.008	0.987
	Regression				
		X$_1$	0.006	0.022	0.981
		X$_2$	-0.001	0.008	0.984
		X$_3$	0.000	0.003	0.994
		X$_4$	-0.051	0.345	0.952
		X$_5$	0.001	0.007	0.984
	PS				
		X$_1$	**0.013**	**0.033**	0.950
		X$_2$	-0.002	**0.010**	0.969
		X$_3$	-0.005	-0.011	0.965
		X$_4$	0.001	0.206	0.947
		X$_5$	0.010	0.006	0.959
	LOCF				
		X$_1$	-0.012	0.030	0.952
		X$_2$	**-0.021**	-0.046	0.933
		X$_3$	**-0.046**	**0.126**	**0.877**
		X$_4$	**-0.738**	**0.240**	**0.653**
		X$_5$	**-0.556**	**0.216**	**0.755**

Note: The largest bias and efficiency for each given estimate appear in bold. PS=propensity score; LOCF=last observation carried forward; MCMC=markov chain monte carlo.

2.7 Discussion and conclusion

In this chapter, we have studied the comparison of four imputation techniques applied to user-made incomplete data sets with missing covariates. The missingness pattern considered is the monotone pattern. The imputation techniques compared included the last observation carried forward (LOCF) method and three imputation techniques, namely markov chain monte carlo (MCMC), regression and propensity score (PS). In order to compare the performance of the methods, we used the originally complete data set (actual data), and then we artificially created missing values to achieve the intended goal. Missingness was imposed on covariates. The results of the regression analysis of the imputed models were compared under three criteria: bias, efficiency and coverage. Data from a diabetes study is used to investigate the performance of the considered approaches.

Generally, comparing the analysis based on 20% and 30% missing data rate, among the imputation techniques examined here, the PS and LOCF techniques were notable for consistently producing more biased estimates versus those in the MCMC and Regression methods, regardless of the missing data rate. This agrees with known theoretical findings that PS can give biased estimates of regression coefficients when data on predictor variables are missing (Allison, 2000). Schafer (1999) found that the PS technique is not appropriate for analyses involving relationships among variables, such as a regression analysis. It would appear that Schafer's (1999) recommendation to not use PS for regression analyses with missing values is strongly supported by the results presented here. On the other hand, in this particular data set, the LOCF technique gave worse results and thus a less efficient technique, in terms of bias, efficiency and coverage than other imputation techniques considered in the current study. LOCF can be justified from the fact that the method is used primarily for analyses of longitudinal studies that have experienced attrition (Little and Rubin, 1987). Also, Schafer (2000) refers to the LOCF as inferior to regression problems because it ignores regression to the mean. In addition, Molenberghs and Kenward (2007) argued against the method and noted that the technique makes the strong assumption that there is no change in the subject response between the observed time points and the missing time period which can lead to biased estimates. However, the LOCF may be appropriate for the particular data set when the pattern of missingness is monotone.

Findings in general favored the MCMC and Regression methods over the PS and LOCF methods. As expected, the MCMC and Regression methods yielded better results, regardless of the missingness rate. Both, MCMC and Regression methods yielded acceptable performance of parameter estimates, and obviously have an advantage over PS and LOCF methods. Furthermore, MCMC and Regression methods yielded estimates closer to each other for bias, efficiency and coverage, and in some cases they yielded the same estimates. This confirmed the theoretical findings from Rubin (1987) and Schafer (1997) who concluded that to handle missing values for a continuous variables in data sets with monotone missingness pattern, we should use methods that assume multivariate normality as MCMC and Regression.

In conclusion, the results show that we have universally best methods to deal with missing covariates under monotone missing data pattern. From the results, it appears that either MCMC or Regression methods of imputation for estimating regression models with monotone missingness are preferable to PS method (which is a non-parametric technique) and LOCF method (which is a single imputation technique).

Chapter 3

An analysis for handling dropout in longitudinal data using multiple imputation and inverse probability weighting*

3.1 Abstract

Missing data is a pervasive problem in longitudinal studies, and it is the result mainly of non-responses due to individuals who leave the study and are therefore lost to follow-up. This chapter deals with incomplete longitudinal data when there are dropouts. Statistical methods that ignore the mechanism for dropouts are susceptible to biased inference. This study focuses on dropouts missing at random (MAR). We demonstrate application and the performance of inverse probability weighting (IPW) and multiple imputation (MI) in handling dropouts in longitudinal data with continuous response. The main objective of this study is to compare the performance of the above approaches for handling missing outcomes in longitudinal data under different dropout rates. Data from a study with individual heart rate as the outcome is used

* A. Satty and H. Mwambi (In review). An analysis for handling dropouts in longitudinal data using multiple imputation and inverse probability weighting. *Statistics and its inference.*

to investigate the performance of the two approaches considered in the current study. Based on this longitudinal data, results from the IPW approach will be compared with those obtained from the MI approach. The performance of these two approaches are compared in terms of bias and efficiency.

Keywords: Incomplete longitudinal data, Dropout, Inverse probability weighted (IPW), Continuous outcomes.

3.2 Introduction

Despite the fact that longitudinal studies are frequently designed to collect data on every individual within a sample at each assessment or measurement occasion, incompleteness or missingness often arises. Incompleteness for longitudinal data often occurs as dropout which is when individuals fail to complete a study for some reasons. A common problem with the analysis of longitudinal data is that, subjects may dropout of the study before the end of the follow-up period resulting in a monotone missingness pattern. This chapter only pays attention to the monotone missing data pattern, in the sense that if a subject drops out from the study prematurely, then on that subject no subsequent repeated measurements of the outcome are obtained. The potential impact of missing data is best understood by considering the process (i.e., the mechanisms) leading to the incompleteness. Rubin (1976, 1987) classified these mechanisms into three basic categories: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).

Two general methods are used for handling incomplete longitudinal data with dropout under MAR. One is the so-called multiple imputation (MI) which imputes values for the incomplete data and carries out the analysis as if the imputed values were observed data. MI was first proposed by Rubin (1978), and a comprehensive treatment is given in Little and Rubin (1987). The method has been recognized as an attractive, important and influential technique for handling missing data, and has become very practical for dealing with dropout in longitudinal settings (Schafer, 1999). A useful summary of MI is presented by Schafer (1997). For more references, see, for example, Rubin and Schenker (1986), Little and Rubin (2002), Rubin (1986) and Horton and Lipsitz (2001). The second approach is the inverse probability weighted (IPW) estimating equations in which complete cases are weighted by the inverse of their probabilities of being observed in order to adjust for dropouts. IPW was first described by Robins et al. (1995) who noted that it deals with incomplete longitudinal data arising from a MAR mechanism. The roots of this approach in survey analysis have been presented by Horvitz and Thompson (1952). IPW has been recognized as an attractive approach because it does not require complete specification of the joint distribution of the longitudinal responses, but rather is based only on specification of the first two moments (Grace and Wenqing, 2009). Several methodological research papers in the literature (Robins et al., 1995; Robins and Rotnitzky, 1995; Scharfstein et al., 1999) have proposed improved IPW estimates that are theoretically more efficient, these are estimates where the MAR may be assumed. The IPW method is discussed in more detail in Fitzmaurice et al. (1995), Yi and Cook (2002a, 2002b), Carpenter et al. (2006) and Seaman and White (2011).

MI and IPW are two approaches to use for handling missing data that provide unbiased estimates under MAR mechanism. These approaches can often give similar results when the mechanism causing the missing data is MAR. The main difference between IPW and MI is that IPW needs a model for the missingness mechanism, whereas MI needs the analyst to specify the variables to be used as regressors in the imputation model. For more detail, see Seaman and White (2011). In addition, unless a monotone missing data pattern is used, the missingness model for IPW can only use complete variables. Both methods can be used for all types of outcomes, but a great deal of work has been devoted to binary response data. However, comparisons between IPW and MI are not common because they come from two opposing schools of thought. Therefore, it is not surprising that essentially little has been done in terms of comparing them for continuous responses data. A recent comparison of these methods in a cross sectional setting found the performance of these methods to be similar, with MI only slightly more efficient than the IPW (Carpenter et al., 2006). In the context of survey data, Seaman and White (2011) compare the performance of MI with IPW. In their paper, they illustrate why, despite MI generally being more efficient, IPW may sometimes be preferred. Using marginal structural models, a comparison of these approaches found that MI was slightly less biased and considerably less variable than IPW (Moodie et al., 2008).

In this chapter, our focus will be on the comparison of MI and IPW approaches to analyze longitudinal data with dropouts, and application will be confined to the continuous outcome case. Our strategy to achieve this goal will be to compare the performance of these two methods in handling incomplete Gaussian longitudinal data, under three different dropout rates. In both methods the dropout mechanism is assumed to be MAR. In order to compare the performance of these methods, the data set we used was originally complete (no missing values) and the dropouts were created by generating missing data at random. The comparison was based on a heart rate trial data which gives heart rate observations for individuals exposed to three different treatments. The performance of these two approaches are assessed on two criteria namely bias and efficiency. Because the likelihood-based inference is valid under the MAR mechanism (Verbeke and Molenberghs, 2000; Mallinckrodt et al., 2003a, 2003b), its results were presented and used as references against which these two approaches were contrasted. In Section 3.3, we present the notation and concepts of possible mechanisms that can lead to the dropout process. In Section 3.4, the two approaches mentioned above are then considered in more detail as the principle approaches to be used in the analysis. Section 3.5, contains the design of the application study and offers a description of the data set used in the analysis in detail. The results based on the generated dropout data in our application study are set out in Section 3.6. We conclude with a discussion of the results in Section 3.7.

3.3 Notation and dropout mechanisms

To describe the different dropout mechanisms and definitions, we follow the notations that is commonly used in the missing data literature. So, suppose that N individuals are to be observed at n occasions. Then for the ith individual ($i = 1, 2, ...N$) we can form a $(n \times 1)$ vector Y_i=$(Y_{i1}, ..., Y_{in})'$, where Y_{ij} is the jth outcome for individual i, which can be continuous or discrete depending on the study problem. Each individual also has a $(n \times p)$ covariate matrix X_i. The covariates may be both time stationary and time varying. In longitudinal studies, individuals can be unobserved at all n occasions on account of some stochastic missing data mechanism. Now, suppose R_i is a $(n \times 1)$ random vector for the ith individual, whose jth component R_{ij} equals 1 when Y_{ij} is fully observed, and equals 0, if not. The purpose of the random vector R_i is to aid in modelling the missingness process. Thus, the full data information for the ith individual are given jointly by Y_i and R_i, with a joint distribution that can be expressed as

$$f_{y,r} = (y_i, r_i \mid X_i, \theta, \gamma) = f_r(r_i \mid y_i, X_i, \gamma) f_y(y_i \mid X_i, \theta), \tag{3.1}$$

where θ and γ are vectors that parameterize the joint distribution. The "missing data mechanism", $f_r(r_i \mid y_i, X_i, \gamma)$ is parameterized by γ. In general, the mechanism of missing data can depend on the full vector of responses, Y_i (including possibly unobserved component of Y_i) and the matrix of covariates X_i. We denote the observed and unobserved components of Y_i by Y_i^o and Y_i^m, respectively. Rubin (1976, 1987) specified three distinct missing data mechanisms. First, we have data that is missing completely at random (MCAR), meaning that the missingness process does not depend on Y_i, i.e., $P(R_i \mid Y_i^o, Y_i^m, X_i, \gamma) = P(R_i \mid X_i, \gamma)$. Second, the missing data is said to be missing at random (MAR) if, the missingness process depends on the observed responses and probably on measured covariates but not on the unobserved responses, i.e., $P(R_i \mid Y_i^o, Y_i^m, X_i, \gamma) = P(R_i \mid Y_i^o, X_i, \gamma)$. The third missing data mechanism is that which allows the missingness process to depend on the unobserved responses, and here such a process is called missing not at random (MNAR), or in probability terms, $P(R_i \mid Y_i^o, Y_i^m, X_i, \gamma) = P(R_i \mid Y_i^m, X_i, \gamma)$. In terms of likelihood based inference for parameters of the complete data vector Y_i, Rubin (1987) showed that the contribution to the likelihood attributable to the missingness mechanism can be ignored under MAR assumption. In the context of likelihood based analysis, an MCAR is a special case of MAR, and these two mechanisms are referred to as being "ignorable". In contrast, an MNAR mechanism is often referred to as a "non-ignorable" mechanism.

The focus of this chapter is on missing data due to subject dropouts, in the sense that all components of Y_i will be missing and all components of R_i will be 0 starting from the dropout time. The dropout time for the ith individual can be defined by introducing a quantitative variable

$$D_i = 1 + \sum_{t=1}^{n} R_{it}, \tag{3.2}$$

and hence the model for missing data or dropout process can be rewritten as

$$\iota_{id_i} = f(r_i \mid y_i, X_i, \gamma) = Pr(D_i = d_i \mid y_i, X_i, \gamma). \tag{3.3}$$

where d_i is a realization of the variable D_i. In Equation (3.2), it is assumed that all subjects are observed on the first occasion so that D_i takes values between 2 and $n+1$. The maximum value $(n+1)$ corresponds to a complete measurement sequence. Using Equation (3.3), a dropout missing completely at random (MCAR) model reduces to $P(D_i = d_i \mid Y_i, X_i, \gamma) = P(D_i = d_i \mid X_i, \gamma)$, while the dropout missing at random (MAR) model is given by: $P(D_i = d_i \mid Y_i, X_i, \gamma) = P(D_i = d_i \mid Y_i^o, X_i, \gamma)$, where dependence on Y_i is only through Y_i^o.

3.4 Methods for handling dropouts

3.4.1 Multiple imputation (MI)

This method is a simulation-based approach that imputes missing values multiple times (Little and Rubin, 1987). The method is valid under the assumption that the data are MAR (Little and Rubin, 1987). The key idea of this approach is to fill in the missing values multiple times in order to construct multiple complete data sets. MI involves three distinct phases or using Little and Rubin's (1987) terminology: In the first step, the missing values are filled in M times to generate M complete data sets. In the process of filling in missing values, a joint distribution for the complete data set (including observed and unobserved data) and a prior distribution of parameters are assumed for the data augmentation algorithm to simulate random draws from the missing data distribution. Under the MAR mechanism, M independent random numbers can, given the observed values, be generated from the stationary conditional distribution of the missing values, as in the Bayesian estimation technique. After the first step (the imputation step), M complete data sets are obtained. In the second step, each of the M complete data sets are analyzed using standard procedures, such as ordinary least squares regression analysis, linear mixed model, generalized linear model, generalized linear mixed models, etc, depending on the types of response and assumptions used for the model. Finally, in the third step, the estimates from the M analyses are then combined to produce a single estimate that incorporates the usual sampling variability as well as the variability due to the missing data.

There is an important question to be solved when applying MI approach; that is, what variables should be included in the imputation model. The MI inference assumes that the model that is used to analyze the multiply imputed data (the analysis model) is the same as the model used to impute missing values in MI (the imputation model). However, practically, the two models might not be the same (Schafer, 1997). The quality of the imputation model will influence the quality of the analysis model results, so it is important to carefully consider the design of the imputation model. Therefore, to obtain high-quality imputations for a particular variable, the imputation model should include variables that are potentially related to the imputed variable and variables that are potentially related to the missingness of the imputed variable (Schafer, 1997). Van Buuren et al. (1999) recommended to include the following covariates in the imputation model: variables in the analysis model, variables associated with missingness of the imputed variable and variables correlated with the imputed variable. However, one can include auxiliary variables which may or may not have missing values. Generally, including variables that do not have missing values are required in the imputation model. For more details of the imputation model, we recommend Schafer and Graham (2002), Schafer (1999) and Van Buuren et al. (1999). We now consider the theoretical justifications of the MI method provided by Verbeke and Molenberghs (2000) in describing the MI processes for data analysis. Recall that the observed data are Y_i^o and the complete data are Y_i. MI uses Y_i^o to fill in Y_i^m, leading to the complete data $Y_i = (Y_i^o, Y_i^m)$. If we knew the distribution of Y_i^m, with parameter vector γ, then we could impute Y_i^m by drawing from the conditional distribution $f(Y_i^m \mid Y_i^o, \gamma)$. Since γ is unknown, we estimate it from the data, yielding $\hat{\gamma}$, and use the distribution $f(Y_i^m \mid Y_i^o, \hat{\gamma})$. Because $\hat{\gamma}$ is a random variable, we must also take its variability into account in drawing imputations. In Bayesian terms, γ is a random variable of which the distribution depends on the data. So, we first obtain the posterior distribution of γ from the data, a distribution which is a function of $\hat{\gamma}$. After formulating the posterior distribution of γ, the following imputation algorithm can be used: (1) Draw γ^* from the posterior distribution of γ, $f(\gamma \mid X_i, Y_i^o)$. We approximate this posterior distribution by the normal distribution. (2) Draw Y_i^m from $f(Y_i^m \mid X_i, Y_i^o, \gamma^*)$. (3) Use the complete data Y_i and the model to estimate the parameter of interest (β^*) and its variance ($\Sigma(\beta^*)$) called the within-imputation variance. The steps described earlier are repeated independently M times, resulting in β_k^*, $\Sigma(\beta^*)$, $k = 1, ..., M$. Steps 1 and 2 are referred to as the imputation task, and step 3 is the estimation task. Finally, we combine the estimates obtained after M imputations. The overall estimated parameter vector is the average of all individual estimates:

$$\beta^* = \frac{1}{M}\sum_{k=1}^{M}\beta_m^*. \tag{3.4}$$

We obtain the variance as a weighted sum of the within-imputation variance and the between-imputations variability:

$$\Sigma^* = W + \left(\frac{M+1}{M}\right)B, \tag{3.5}$$

where

$$W = \frac{1}{M}\sum_{k=1}^{M}\Sigma(\beta_k^*), \tag{3.6}$$

is the average of the within-imputation variances, and

$$B = \frac{1}{M-1}\sum_{k=1}^{M}(\beta_k^* - \beta^*)(\beta_k^* - \beta^*)', \tag{3.7}$$

is the between-imputations variance (Little and Rubin, 1987). According to Verbeke and Molenberghs (2000), γ is an easily estimated set of parameters characterizing the distribution of Y_i, and in this situations MI is most useful. In contrast, β is complicated to estimate in the presence of non-response and obtaining a correct estimate for the variability is nontrivial.

3.4.2 Inverse probability weighting (IPW)

IPW is a standard method used for handling dropouts. This method is valid under MAR assumption (Robins et al., 1995), but requires specification of a dropout model in terms of observed outcomes and/or covariates. IPW is more generally used in marginal models for discrete outcomes than for continuous outcomes, however in this study, IPW is adopted for dealing with continuous outcomes in order to correct the bias that is caused by dropout under MAR assumption. The primary idea behind IPW is that, if individual i has a probability of being observed at occasion t of λ_{it}, then, this individual should be given weight, ω_{ij}, so as to minimize the bias caused by dropouts in the analysis. The weight ω_{ij} for the ith individual at time j is assigned as inverse of the cumulative product of fitted probabilities, $\widehat{\omega_{ij}}(\hat{\alpha})$=$(\hat{\lambda}_{i1}(\hat{\alpha}) \times \hat{\lambda}_{i2}(\hat{\alpha}) \times ... \times \hat{\lambda}_{ij}(\hat{\alpha}))^{-1}$, where α is a $(q \times 1)$ vector of unknown parameters. In order to discuss the idea of what these weights are, we follow illustration provided by Carpenter et al. (2006). Suppose that we have the following data, then the average response is 3. However, if we have missing values as shown below, then the average response is 19/6 which is biased.

Group:	A	B	C
Response:	222	333	444

Group:	A	B	C
Response:	2??	333	?44

In order to correct this bias, we calculate the probabilities of being observed in each group corresponding to 1/3 in group A, 1 in group B and 2/3 in group C. We thereafter calculate a weighted average where each observation is weighted by 1/[Probability of observed response]. In this case the weighted average is given by

$$\frac{2 \times \frac{3}{1} + (3+3+3) \times 1 + (4+4) \times \frac{3}{2}}{\frac{3}{1} + 1 + 1 + 1 + \frac{3}{2} + \frac{3}{2}} = 3, \tag{3.8}$$

which now corrects the bias. The conclusion to be drawn from this simple illustration is that IPW has eliminated the bias by reconstructing the full population by up-weighting the data from individuals who have small chance of being observed. Generally, it may give biased but consistent parameter estimates (Carpenter et al., 2006). To discuss the above mentioned idea of IPW in longitudinal data setting, we now describe the IPW approach thereby illustrating how IPW can be incorporated into the conventional generalized estimating equations (GEE) by Liang and Zeger (1986) as based on the article by Robins et al. (1995). The primary idea behind GEE methodology is to generalize the usual univariate likelihood equations by introducing the covariance matrix of the vector of response, Y_i. The GEE methodology is used to model the marginal expectation of responses as a function of a set of covariates. We briefly introduce the classical form of GEE (Liang and Zeger, 1986). Let $X_i = (x_{i1}, ..., x_{in})'$ denote an $(n \times p)$ covariates matrix, where its tth row is given $x'_{it} = (x_{it1}, ..., x_{itp})'$ based on p predictor variables or covariates, Y_{it} denote the response variable and hence $y_i = (y_{i1}, ..., y_{in})'$ the $(1 \times n)$ observed response vector and $\mu_{it} = E(y_{it})$, $i = 1, ..., N$ and $t = 1, ..., n$. Now, assume the marginal regression model is given as

$$g(\mu_{it}) = x'_{it}\beta, \tag{3.9}$$

where β is the $(p \times 1)$ regression parameters of interest and $g(.)$ is a link function, a function of the mean response. Assume the $(n \times n)$ covariance matrix for Y_i is: $\Sigma(\varphi) = \phi A_i^{1/2} R(\rho) A_i^{1/2}$, where A_i is a diagonal matrix of variance functions, $R(\rho)$ is the working correlation matrix of Y_i as a function of ρ the correlation parameter and ϕ is the dispersion parameter. The collection of parameters in the covariance matrix are assumed to contained in the parameter vector φ. Then, the GEE estimators for regression parameters are the solutions of

$$\sum_{i=1}^{n} D_i \Sigma(\varphi)^{-1}(Y_i - \mu_i) = 0, \tag{3.10}$$

where $D_i = \frac{\partial \mu_i}{\partial \beta}$ is the derivative matrix of the mean vector μ_i with respect to β. The GEE methodology is very popular especially for analysis of marginal models for discrete responses than for continuous responses. However, in this study, we restrict our attention to the continuous responses. Consequently, the following assumptions can be made for the marginal models with the continuous response, Y_{it}

- The mean of Y_{it} is related to the covariates by an identity link function: $\mu_{it} = \eta_{it} = x^{'}_{it}\beta$. The link function $g(.)$ generally relates the expected values, μ_i of the response vector, Y_i to the covariate matrix, X_i. It takes the general form $g(\mu_i) = \eta_i = X_i\beta$, where η_i denotes the linear predictor vector whose tth row is $g(\mu_{it}) = \beta_1 x_{it1} + \beta_2 x_{it2} + \cdots + \beta_p x_{itp}$. This function, i.e., $g(.)$, should be monotone and differentiable. Thus, in the case of monotonicity, we can define the inverse function $g(.)^{-1}$ by the relation $g^{-1}(g(\mu_i)) = \mu_i$. Here, we note that the choice of link function depends on the distributional assumptions on the data. Therefore, for a continuous response with normal assumption, as in our case, the link function is an identity link: $g(\mu_i) = \mu_i$ and the inverse simply $\mu_i = g(\mu_i)$. Under this identity link, the expected value of the response is simply a linear function of the covariates multiplied by their regression coefficients.

- The variance of each Y_{it}, conditional on the effects of the covariates, is ϕ and does not depend on the mean response. Namely, $\nu(\mu_{ij}) = 1$ is a known "variance function", therefore $Var(Y_{it}) = \nu(\mu_{it})\phi = \phi$. Here ϕ denotes the variance of the conditional normal distribution of the response, given the covariates. The assumption that the variance is constant over time may be unrealistic and to relax it, a separate scale parameter, ϕ_t could be estimated at the tth occasion if the longitudinal design is balanced on time.

- The within-individual correlation among repeated responses is modelled by assuming, for example, a first-order autoregressive AR(1) covariance structure, $Corr(Y_{ij}, Y_{ik}) = \rho^{|k-j|}$, which stand for the pairwise correlation between observations, for all j and k and $0 \leq \rho \leq 1$. The AR(1) specifies homogeneous variances. In addition, it specifies that the correlations between observations on the same subject are not equal, but decrease toward zero with increasing length of the time interval between observations.

For marginal models with an identity link function, the generalized least square means (GLS) of β can be considered as a special case of the GEE. Therefore, the estimates of parameters in marginal model for continuous response with an identity link are

$$\hat{\beta} = \{\sum_{i=1}^{n} X_i' \hat{\Sigma}_i^{-1} X_i\}^{-1} \sum_{i=1}^{n} (X_i' \hat{\Sigma}_i^{-1} Y_i), \tag{3.11}$$

where $\hat{\Sigma}_i$ is the REML (Restricted Maximum Likelihood Estimation that can be used to find the best unbiased estimates (Verbeke and Molenberghs, 2000)) estimate of Σ_i and

$$Cov(\hat{\beta}) = \{\sum_{i=1}^{n} X_i'\hat{\Sigma}_i^{-1}X_i\}^{-1} \sum_{i=1}^{n}(X_i'\hat{\Sigma}_i^{-1}\hat{V}_i\hat{\Sigma}_i^{-1}X_i)\{\sum_{i=1}^{n} X_i'\hat{\Sigma}_i^{-1}X_i\}^{-1}, \tag{3.12}$$

where $\hat{V_i} = (Y_i - X_i\beta)(Y_i - X_i\beta)$ is an estimate of $Var(Y_i)$ which yields a robust estimator of $Cov(\hat{\beta})$ when substituted in Equation (3.12). With incomplete data that are MAR, the GEE method provides inconsistent estimates of the model parameters (Liang and Zeger, 1986). In weighted generalized estimating equations (WGEE), an individual's contribution to the usual GEE is weighted by the inverse probability of dropout at particular time point, given the individual did not leave or dropout in any of the periods occasions (Robins et al., 1995). Therefore, setting all assumptions that are made in this section together, we will get valid parameter estimates in longitudinal studies with MAR dropout by solving the weighted estimating equations

$$\sum_{i=1}^{n}(Y_i - X_i\beta)'\Sigma_i^{-1}W_i(\hat{\alpha})(Y_i - X_i\beta) = 0, \tag{3.13}$$

where $W_i(\hat{\alpha})$ is a diagonal matrix which contains inverse probability weights for ith patient, $W_i(\hat{\alpha}) = diag\{\hat{w_{i1}}(\hat{\alpha}), ..., \hat{w_{in_i}}(\hat{\alpha})\}$ for $j = 2, ..., n_i$, $\hat{w_{i1}} = 1$, and $\Sigma_i = A_i^{1/2}R(\rho)A_i^{1/2}$ is a $(n \times n)$ working covariance matrix for Y_i and $R(\rho)$ is a $(n \times n)$ working correlation matrix which are assumed known. The missingness is taken into account through specification of a $(n \times n)$ diagonal weighting matrix of $W_i(\hat{\alpha})$, $W_i(\hat{\alpha}) = diag(R_{i1}\hat{w_{i1}}(\hat{\alpha}), ..., R_{in}\hat{w_{in}}(\hat{\alpha}))$ and $R_{it} = 1$ if the ith subject is observed at time t, and 0 for the unobserved time. The weight, w_{ij} is the inverse of the probability that the ith subject is observed at the jth time which is often unknown and needs to be estimated. It requires modelling the missingness process in order to obtain the weights w_{ij}. We denote $\lambda_{ij}(\alpha) = P(R_{ij} = 1 \mid R_{i(j-1)} = 1, X_i, Y_i, \alpha)$ as the probability of a response being observed at time j for the ith subject given the subject is observed at the time $j - 1$. If the missingness is assumed to be MAR, we have

$$\lambda_{ij}(\alpha) = P(R_{ij} = 1 \mid R_{i(j-1)} = 1, X_i, Y_{i1}, ..., Y_{i(j-1)}, \alpha), \tag{3.14}$$

where the missingness mechanism only depends on observed data and may be specified up to a $(q \times 1)$ vector of unknown parameters, α. Here, λ_{ij} can be modelled as a logistic regression model of Z_{ij}, a vector of predictors which may include missingness indicator variables, covariates and previous responses:

$$logit\lambda_{ij}(\alpha) = Z'_{ij}\alpha, \tag{3.15}$$

or by inverting the logit function we have:

$$\lambda_{ij}(\alpha) = \frac{e^{Z'_{ij}}\alpha}{1 + e^{Z'_{ij}}\alpha}. \tag{3.16}$$

The log partial likelihood for ith subject takes the form

$$\ell(\alpha) = \sum_{j=1}^{n}\sum_{i=2}^{j_i} R_{i(j-1)} log\{\lambda_{ij}(\alpha)^{R_{ij}}[1 - \lambda_{ij}(\alpha)]^{1-R_{ij}}\}. \tag{3.17}$$

Differentiation of (3.17) in terms of α gives the estimating equations

$$S_i(\alpha) = \{\sum_{i=1}^{N}\sum_{j=2}^{j_i} R_{i(j-1)}[R_{ij} - \lambda_{ij}(\alpha)]\}. \tag{3.18}$$

Setting (3.18) equal to zero yields $\hat{\alpha}$, therefore we can obtain estimate of $\lambda_{ij}(\alpha)$ which is $\hat{\lambda}_{ij}(\hat{\alpha})$. According to Hogan et al. (2004), in addition to MAR dropout, two assumptions must be fulfilled to provide consistent estimates of parameters β in weighted method. *First assumption* (Non-zero probability of remaining in study): Conditionally on past history of observed responses and covariates, the probability that individual i is still in the study at time j is bounded away from zero or formally, $p[R_{ij} = 1 \mid R_{i(j-1)} = 1, X_i, Y_i] > \delta > 0$. *Second assumption* (Correct specification of dropout model): The probability of dropout model must be correctly specified, i.e., $\bar{\lambda}_{ij}(\alpha) = p[R_{ij} = 0 \mid R_{i(j-1)} = 1, X_i, Y_{i(j-1)}]$. Thus, it should be noted that $\lambda_{ij}(\alpha) = 1 - \bar{\lambda}_{ij}(\alpha)$. When MAR and monotone missingness assumptions hold, the probabilities of remaining in the study,

$$\pi_{ij}(\alpha) = p[R_{ij} = 1 \mid R_{i(j-1)} = 1, X_i, Y_{i1}, ..., Y_{i(j-1)}, \alpha] = \prod_{k=1}^{j}\{1 - \bar{\lambda}_{ik}(\alpha)\}. \tag{3.19}$$

Thus, the weight $\widehat{w_{ij}}(\hat{\alpha})$, the inverse of the unconditional probability of being observed at time j, can be calculated as,

$$\hat{w}_{ij}(\hat{\alpha}) = \frac{1}{1 \times (1 - \hat{\bar{\lambda}}_{i2}(\hat{\alpha})) \times ... \times (1 - \hat{\bar{\lambda}}_{ij}(\hat{\alpha}))}, \qquad i = 2, ..., J, \tag{3.20}$$

and $\hat{w}_{ij}(\hat{\alpha}) = 1$ for $j = 1$. Therefore, if the above two assumptions due to Hogan et al. (2004) hold, and if dropout occurs according to the MAR mechanism, then the estimators of the parameters $\hat{\beta}$ in the weighted marginal model for a continuous response with an identity link will be of the form

$$\hat{\beta} = \{\sum_{i=1}^{n} X_i'\hat{\Sigma}_i^{-1}W_i(\hat{\alpha})X_i\}^{-1}\sum_{i=1}^{n}(X_i'\hat{\Sigma}_i^{-1}W_i(\hat{\alpha})Y_i), \tag{3.21}$$

and

$$Cov(\hat{\beta}) = \{\sum_{i=1}^{n} X_i'\Sigma_i^{-1}W_i(\hat{\alpha})X_i\}^{-1}(\sum_{i=1}^{n} X_i'\Sigma_i^{-1}W_i(\hat{\alpha})W_i(\hat{\alpha})'X_i)\{\sum_{i=1}^{n} X_i'\Sigma_i^{-1}W_i(\hat{\alpha})X_i\}^{-1}, \tag{3.22}$$

where $\hat{\beta}$ is consistent for β and $\hat{\alpha}$ is a consistent estimator of α under a correctly specified model, $\lambda_{ij}(\alpha)$.

3.5 Application study

3.5.1 Description of the data

In this section, we describe the application of the aforementioned methods for handling dropouts in longitudinal data. The methods are applied to data from heart rate experiment for which there were no dropouts. Our current study is an application study rather than a case study, that we tested the performance of the two approaches by generating dropouts from a complete data. So, our main interest was to generate a random sample of the whole data set and then to use it for the analysis. The data set to be analyzed in this study originates from the clinical trial to study the effect of three treatments on heart rate of humans. Full details of this experiment are given in Millikin and Johnson (2009). It is an experiment involving three drugs (AX23, BWW9 and CTRL) and where each subject was measured repeatedly at four different time points ($j = 1, 2, 3, 4$). After the drug was administered, each patient's heart rate was measured every five minutes for a total of four times. To be precise, each patient's heart rate was measured 5, 10, 15 and 20 minutes after administering the treatment. This experiment illustrates the layout for a simple repeated measures experiment. The large size of experiment units is the subject and the smaller size experiment unit is the time interval when using the split-plot in time notation. At the start of the study n female human subjects were randomly assigned to each drug. Figure 3.1 shows the

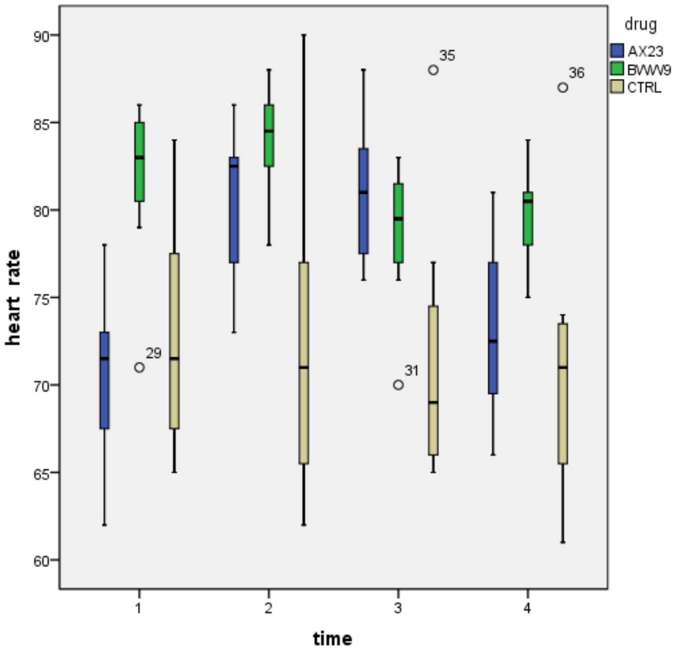


Figure 3.1: *Box-plot for the distribution of heart rate across all four time points for all three drug groups.*

distribution of measurements in terms of box-plots at all four time points by all three drug groups. The objective of this experiment was to investigate the drug-response effects; that is, if the drugs have an effect on heart rate, compare drug groups with each other including time effects and to find the least-square means. In this chapter, we consider the significance of drug main effects, time main effects and the interaction of time and drug effects, and we are also interested in investigating the differences between the drug and time effects in least-square means.

3.5.2 Model formulation

In the proceeding, we analyze the data from the clinical trial introduced above by formulating a model based on the data with heart rates. According to the study design, we include the fixed categorical effects of drug, time and drug-by-time interaction. Therefore, the continuous outcome for the analysis reported here was heart rate, or as we will denote it in the remainder of this study, HR. Let HR_{ijk} denote the heart rate of patient i at time j on drug k, where i=1,...,8, j=1,...,4, and k=1, 2, 3. In the following, we consider the linear model for HR$_{ijk}$, where the response of the subject i at time j:

$$HR_{ijk} = \beta_0 + \beta_1 Time_j + \beta_2 Drug_k + \beta_3 (Time * Drug)_{jk} + \varepsilon_{ijk}, \tag{3.23}$$

where $(Time * Drug)$ denotes the drug-by-time interaction and ε_{ijk} are unknown independent and identically distributed normal random error, with mean 0 and variance σ_ε^2. The fitted model 3.23 is not a simple multiple linear regression model but a GEE model as the two are not the same under a GEE model requires a correlation structure to be specified i.e. ε_{ijk} is not $\sim N(0, \sigma^2)$. As mentioned above, in this data set, there are no actual dropouts. This provides us with an opportunity to generate dropouts missing at random in order to compare the performance of MI and IPW methods to deal with dropouts.

3.5.3 Generating dropouts and the MAR mechanism

We used the full data set to artificially generate missing values by mimicking the MAR mechanism. From the complete data set described above, 1000 random samples of n=96 were drawn. The dropouts in HR were created according to the MAR assumption, assuming the missingness in HR is related to observed values, in the sense that patients with higher HR at one measurement occasion tend to dropout out of the experiment at the next occasion. The implication of the MAR assumption in our case is that, patients who are observed to be weaker (deduced by way of their previous observed outcome) are more likely to dropout when they reach a certain value of the HR as long as their probability of dropout does not further depend upon their missing responses. The other predictor variables other than HR were however kept intact. For the MAR mechanism, three dropout rates were implemented. The dropout rates were set at 10%, 20% and 30%. Dropouts were created in HR by randomly deleting 10%, 20% and 30% of all observations greater than 75 as a threshold indicating high heart rate. The observations that triggered the missing data were kept but all other subsequent observations were deleted. This scenario was generated or replicated 1,000 times. Each generated samples was analyzed using MI, IPW and direct likelihood analysis to derive parameters of interest. A monotone missing pattern was assumed which is to say that for each patient, if a HR's observation was deleted for a third time point, the subsequent observation in the fourth time point for that patient was also deleted.

3.5.4 Implementation of MI

The MI approach was carried out using SAS PROC MI to fill in all the missing values for each generated sample. Under the MAR setting, MI requires the analyst to specify which variables are to be used as regressors in the imputation model. The imputation model was fitted using the above data set of all patients who had both time and drug data available. The imputation model is based on model (3.23) that assumes multivariate normality of the variables. In the imputation model, the variables that are included in Y_i^o should be those that make the HR (the response variable) missing at random so that $P(Y_i^m$ is observed $\mid Y_i^o)$ does not depend on Y_i^m. Thus, to increase the plausibility of the MAR assumption as well as to improve the accuracy and efficiency of the imputation, we used all the available data, including the HR variable, to predict the missing values since they are potentially related to the imputed variable as well as to the missingness of the imputed variable. This in line with the recommendation provided by Van Buuren et al. (1999) to include the following covariates in the imputation model: variables in the analysis model, variables associated with missingness of the imputed variable and variables correlated with the imputed variable. In the analysis, PROC MI was then applied to generate $M = 5$ complete data sets. Note that the choice of M=5 was considered adequate and the efficiency of the parameter estimate based on imputation given by $(1 + \xi/M)^{-1}$, where ξ is the rate of missing data (Rubin, 1987). This formula shows that the relative efficiency of the MI inference is related to the missingness rate (ξ) in combination with the number of imputations (M). Rubin's (1987) simulation indicates that the number of imputations can generally be constrained to fewer than 10. Many statistical practices tend to support Rubin's heuristics of 3 to 10 imputations. In general however Schafer and Olsen (1998) and Peng et al. (2006) recommended the use of M=5 before the results are combined. In this study, for 10%, 20% and 30% rates of missing data and estimates based on

M=5 implies we will have at least 98%, 96% and 94% efficiency, respectively. Each data set that was produced thus consists of 96 outcomes (HR) with complete data at all four times. This MI method used Markov Chain Monte Carlo (MCMC) sampling to draw imputations. We used a burn in of period of 200 iterations, 100 iterations between each step and five imputations. The defaults of the SAS MI algorithm are used for the MCMC computation, namely a Jeffreys prior and initial sampler values from the EM posterior mode was also used. The linear mixed model was then fitted to each imputed data set, and therefore the results combined for final inference. We used PROC MIXED to generate the fixed-effect parameter estimates and covariance matrix for each imputed data set. Additionally, we used the ODS statement to create an output data sets that matches PROC MIANALYZE for combining the results from the 5 completed data sets and generating valid statistical inferences about the parameters. However, to obtain the effect means associated with the drug and contrasts between drug groups from PROC MIXED, PROC MIANALYZE cannot directly pool the least square means and their differences. Therefore, the LSMEANS table (the LSMEANS statement computes least squares means of fixed effects) has been sorted differently so that PROC MIANALYZE can then be invoked using the BY statement in order to derive the pooled least square means for each effect.

3.5.5 Implementation of IPW

The IPW approach was applied to each generated sample using the SAS macros provided by Molenberghs and Verbeke (2005). These macros presented *DROPOUT* and *DROPWGT* macros to construct the variables “dropout” and “previous measurements” and to pass the weights (predicted probabilities) to be used for WGEE. In contrast to the MI approach, the IPW approach requires a model for the missing data mechanism. Thus in the MAR setting, we assume the IPW models the missingness mechanism via logistic regression model, introduced in model (3.24) which requires the data to be MCAR or MAR. The IPW was used according to the following three steps:
Step 1: The dropout model was fitted within logistic regression using *DROPOUT* macro. The outcome variable “dropout” indicator for HR was generated, and it was

binary taking the value 1 when the HR is observed, 0 if not, thereby indicating whether or not dropout occurred at a given time from the start of the measurement until the end of the study period (Molenberghs and Verbeke, 2005). In the dropout model, predictor variables were the outcomes at previous occasions ($y_{i,j-1}$), supplemented with genuine covariate information. To estimate the dropout probabilities, we used the following logistic regression of dropout indicators

$$logit[P(D_i = j \mid D_i \geq j)] = \psi_0 + \psi_1 y_{i,j-1} + \gamma drug_j, \tag{3.24}$$

where $y_{i,j-1}$ is the binary indicator at the previous occasion.

Step 2: Using data and fitted probabilities from step (1), a weighted regression of the response variable in model (3.13) was fitted based on the inverse of the “probability of a patient dropping out at a given time and was not missing in all the previous times” as weights. This was done by using the *DROPWGT* macro in SAS. These weights were defined at the individual measurement level and were equal to the product of the probabilities of not dropping out up to the measurement occasion (Molenberghs and Verbeke, 2005). The last factor was the probability of either dropping out at that time or continuing with the study.

Step 3: Once the selected model (3.24) is fitted and the weight distribution checked, we formulate the full-data regression model using inverse probability weighting. The weighted regression model is formulated by re-defining the response as $Y_{ij}^* = \hat{w}_{ij}(\alpha)Y_{ij}$ and covariate as $x_{ij}^* = \hat{w}_{ij}(\alpha)x_{ij}$. Now, let HR_{ij} denote the heart rate from patient i at time j for $j = 1, 2, 3, 4$. Further, let x_{ij}^* be a vector of covariates with length p, where $p = 1, 2, 3$. Then, the mean response model can be expressed as follows

$$E(HR_{ij}^* \mid x_{ij}^*) = \mu_{ij}^* = \beta_0 + \beta_1 x_{ij1}^* + \beta_2 x_{ij2}^* + \beta_3 x_{ij3}^*, \tag{3.25}$$

where the covariate x_{ij1}^* denotes the time, the covariate x_{ij2}^* denotes the drug group, the covariate x_{ij3}^* denotes the drug by time interaction, β_0 is the population average intercept and β_1, β_2, β_3 is the average rate of change due to the time, drug main effects and their interaction. Since the response variable of interest at each occasion was the HR which is continuous, we used the identity link function and the scale parameter, ϕ. In addition to the marginal model in (3.25), the covariance structure of the correlated

HR weights on a given patients should be modelled. In the application of IPW only first order-autoregressive AR(1) and compound symmetry covariances can be implemented. The other structures such as unstructured covariance, toeplitz and heterogeneous (AR), may easily present computational problems. Therefore, we used the AR(1) covariance structure since it is the most reasonable in longitudinal data analysis problems. Using model (3.25), the parameter estimates can be calculated as the root of the weighted estimating equations

$$\sum_{i=1}^{n} = (Y_i - X_i\beta)'\Sigma_i^{-1}W_i(\hat{\alpha})(Y_i - X_i\beta) = 0, \tag{3.26}$$

where Y_i and X_i are vectors of HR and covariates, respectively, for ith patient and $W_i(\hat{\alpha})$ is a diagonal matrix consisting of inverse probability weights for the ith patient. Model (3.25) was fitted using SAS procedure GENMOD with a WEIGHT statement.

3.5.6 Assessment criteria

In this study, the performance of MI and IPW was assessed on two criteria, namely bias and efficiency. Schafer and Graham (2002) used these criteria to study the performance of list-wise deletion, single imputation, maximum likelihood and MI. With small multivariate data sets, Graham and Schafer (1999) used these criteria to evaluate the the performance of MI. In this chapter, we defined these criteria as follows: bias refers to the differences between the average of the 1,000 coefficient estimates and the corresponding true coefficient obtained from a mixed model analysis of the original complete data. Thus, a better technique is that which does on average approach the population value is one which has less bias, i.e., there is a small difference between the true population value and estimated value. Efficiency was defined as the variability of the estimates around the true population coefficient. It was calculated by the average width of the 95% confidence interval. The 95% confidence interval width is approximately four times the magnitude of the standard error. Thus, a wider interval implies a less efficient method.

Table 3.1: *Bias and efficiency of MI, LMM and IPW approaches, under different dropout rates: MIXED least squares means - (interaction terms are not shown)*

		Bias			Efficiency		
dropout rate	parameters	MI	LMM	IPW	MI	LMM	IPW
10%	AX23	0.27	0.28	**-1.20**	0.83	0.89	**1.15**
	BWW9	-0.18	-0.18	**-1.75**	0.83	0.90	**1.13**
	CTRL	0.27	0.29	**1.25**	0.83	0.89	**1.10**
	time$_1$	**0.50**	**0.50**	0.48	0.97	0.96	**1.61**
	time$_2$	0.50	0.50	**1.45**	0.97	0.96	**1.13**
	time$_3$	-0.01	-0.08	**1.22**	0.97	1.09	**1.26**
	time$_4$	-0.39	**-0.49**	0.46	0.97	1.07	**1.12**
20%	AX23	0.25	0.41	**1.40**	0.84	0.93	**1.14**
	BWW9	0.50	0.38	**1.40**	0.84	0.94	**1.07**
	CTRL	0.62	0.64	**1.90**	0.84	0.94	**1.04**
	time$_1$	0.08	0.08	**1.90**	0.98	0.96	**1.37**
	time$_2$	0.48	0.48	**2.33**	0.98	**0.99**	0.78
	time$_3$	1.22	1.10	**1.37**	0.98	1.27	**1.54**
	time$_4$	0.06	**0.24**	0.14	0.98	1.27	**1.34**
30%	AX23	0.57	1.24	**1.46**	0.86	1.08	**1.20**
	BWW9	0.89	**1.14**	1.01	0.86	1.08	**1.08**
	CTRL	0.73	1.13	**1.20**	0.86	1.09	**1.20**
	time$_1$	0.56	0.56	**1.41**	1.01	0.97	**1.16**
	time$_2$	0.71	0.71	**1.07**	**1.01**	0.98	0.74
	time$_3$	1.16	1.05	**2.27**	1.01	1.55	**1.68**
	time$_4$	1.20	**1.64**	-0.83	1.01	1.58	**1.66**

Note: The largest bias and efficiency for each given estimate presented in bold. MI=multiple imputation; LMM=linear mixed model; IPW=inverse probability weighting.

3.6 Results

The results of the bias and efficiency of the MI, LMM and IPW approaches under different dropout rates are presented in Table 3.1. Note that we do not show full output as the results of interactions terms are excluded. In terms of the biasedness of the estimates, the performance of MI was unsurprisingly, better than those of LMM and IPW. Among the three approaches examined here, IPW was notable for consistently producing the most biased estimates vis-a-vis those estimates in the MI and LMM. This advantage for MI and LMM is well documented in terms of continuous outcomes (Verbeke and Molenberghs, 2000; Mallinckrodt et al., 2003a, 2003b). Consequently, it appears that the greatest bias of the estimates by the IPW approach was independent of the dropout rates. The results based on MI and LMM were generally similar for the 10% and 20% dropout rates, and in some cases they produced the same estimates. We refer here to estimates of BWW9, time_1 and time_2 for the 10% dropout, and estimates of time_1 and time_2 for the 20% dropout. There appears to be little differences between the MI and LMM methods for 10% and 20% dropout rates in terms of the bias of the estimates.

As we have noted earlier, a wider interval implies a less efficient approach, thus the widest and hence the worst, 95% confidence intervals are highlighted. In terms of efficiency, across all three dropout rates, the IPW was uniformly the worst approach, regardless of the dropout rates, except for estimates of time_2 for the 20% and 30% rates. When compared with the IPW approach, MI and LMM yielded acceptable performance for all dropout rates. Both approaches being asymptotically more efficient, except in estimating time_2 under 20% and 30% rates for LMM and MI, respectively. Relatively, LMM yielded wider intervals than did MI. The degree of difference in the width of the intervals between the two methods increased with increasing dropout rate. Thus, MI was more efficient than LMM. Therefore, the MI method is more robust against loss of efficient due to increased dropout rate compared to the LMM method and therefore between all the three methods considered.

We now compare and discuss the different approaches by looking at Table 3.2 which shows the results of bias and efficiency for pairwise comparisons among drug and time

Table 3.2: *Bias and efficiency of MI, LMM and IPW approaches, under different dropout rates: Pairwise comparisons among drug main effect means and time main effect means*: ***Differences of least squares means***

						Bias			Efficiency		
dropout rate	effect	drug	time	drug	time	MI	LMM	IPW	MI	LMM	IPW
10%	drug	AX23		BWW9		0.46	0.47	**0.55**	1.18	1.27	**1.61**
	drug	AX23		CTRL		0.01	0.02	**0.04**	1.18	1.27	**1.60**
	drug	BWW9		CTRL		-0.45	-0.47	**-0.50**	1.18	1.27	**1.58**
	time		1		2	0.00	0.00	**0.97**	1.36	1.37	**1.89**
	time		1		3	0.50	0.58	**0.74**	1.36	1.44	**2.03**
	time		1		4	0.89	**0.99**	-0.02	1.36	1.44	**1.97**
	time		2		3	0.50	0.58	**-2.23**	1.36	1.42	**1.75**
	time		2		4	0.98	0.99	**-2.99**	1.36	1.46	**1.60**
	time		3		4	0.48	0.49	**-0.76**	1.36	1.52	**1.69**
20%	drug	AX23		BWW9		0.24	0.33	**0.73**	1.20	1.33	**1.56**
	drug	AX23		CTRL		0.21	0.22	**0.39**	1.20	1.33	**1.54**
	drug	BWW9		CTRL		-0.14	**-0.26**	0.21	1.20	1.33	**1.49**
	time		1		2	0.39	0.39	**0.43**	1.38	1.38	**1.54**
	time		1		3	0.68	**1.28**	1.02	1.38	1.57	**2.02**
	time		1		4	-0.14	-0.15	**2.04**	1.38	1.59	**1.93**
	time		2		3	0.89	0.62	**1.10**	1.38	1.67	**1.84**
	time		2		4	-0.54	-0.24	**2.46**	1.38	**1.62**	1.54
	time		3		4	**-1.43**	-0.86	1.35	1.38	1.83	**2.09**
30%	drug	AX23		BWW9		0.23	0.**32**	**-0.32**	1.22	1.53	**1.62**
	drug	AX23		CTRL		**0.25**	0.12	-0.15	1.22	1.54	**1.71**
	drug	BWW9		CTRL		-0.06	-0.12	**0.16**	1.22	1.53	**1.63**
	time		1		2	0.15	0.15	**1.33**	**1.41**	1.35	1.36
	time		1		3	0.49	0.60	**0.85**	1.41	1.82	**2.03**
	time		1		4	1.64	2.08	**2.25**	1.41	1.84	**2.01**
	time		2		3	0.33	0.45	**1.19**	1.41	1.90	**1.94**
	time		2		4	0.91	**1.93**	1.48	1.41	**1.94**	1.83
	time		3		4	1.15	1.47	**2.11**	1.41	**2.62**	2.37

Note: The largest bias and efficiency for each given estimate presented in bold. MI=multiple imputation; LMM=linear mixed model; IPW=inverse probability weighting.

main effect means, under the three dropout rates. Overall, the bias for the MI approach is negligible regardless of the dropout rate, except for the estimate of the comparison involving the time pair (time$_3$, time$_4$) for the 20% dropout rate as well as the estimate of the comparison involving drug pair (AX23, CTRL) for 30% dropout rate which are asymptotically more biased. This, additionally to the validity of MI approach under MAR dropouts and hence its divergence from IPW for dealing with incomplete longitudinal continuous outcomes, provides a strong justification for the MI approach. As was the case with the results in Table 3.1, the bias associated with IPW's estimates were typically more than with MI and LMM, though this pattern did not hold in all the estimates. However, MI outperformed both LMM and IPW in most cases. When the dropout rate decreased to 10%, the results from MI became nearly indistinguishable from those of LMM as they yielded similar estimates, but in one case more bias was evident regarding the time pairwise comparison (time$_1$, time$_4$) for LMM. However, the LMM estimates appeared to have resulted in fairly minor bias compared to the IPW estimates. Both MI and LMM gave exact estimates with reference to the complete data for the time pairwise comparison (tim$_1$, tim$_2$). This is not surprising since there were no missing observations at this time.

In terms of efficiency condition investigated, the results displayed in Table 3.2 reveal that, as expected, the results were very comparable to what was found from the Table 3.1. Once again narrow intervals are desirable provided that their coverage is near to 95%. On the basis of these results, we make the following general observations. On the one hand, MI and LMM gave more efficient estimates for most cases under all three dropout rates than does IPW. On the other hand, the confidence intervals calculated for the MI, were slightly narrower than does LMM, except for (time$_1$, time$_2$) under 30%, thus the MI method is more efficient than the LMM method for all dropout rates. Furthermore, LMM was uniformly the least efficient approach in estimating pairwise comparison among (time$_2$, time$_4$) for 20% and (time$_2$, time$_4$ and time$_3$, time$_4$) for 30%. In nearly all cases, the performance of estimated intervals by IPW was worst for all dropout rates. The only exceptions to this rule occur for estimates of (time$_2$, time$_4$) for 20% and (time$_1$, time$_2$), (time$_2$, time$_4$) and (time$_3$, time$_4$) for 30%, but for 10% it has seriously less efficient estimates. Thus, overall, IPW was a less efficient approach.

3.7 Discussion and conclusion

In this chapter, we have discussed the performance of using MI and IPW approaches for handling continuous outcomes when there are MAR dropouts in longitudinal data. Both of the approaches were selected for their solid foundations on the dropout under MAR mechanism, and both methods can be used for continuous outcomes, but that work applying the IPW has been devoted to binary responses data. However, little comparison has been done between IPW and MI as they come from two opposing schools of thought. Our main objective was to compare the performances of MI and IPW for handling incomplete longitudinal data based on a continuous outcomes under three different dropout rates. From the complete data set, we generated the MAR dropouts under three dropout rates (10%, 20% and 30%). The comparison between the two methods was based on a heart rate trial data and the estimates corresponding to the MI method were then compared to those obtained from the IPW method. The performances of these two approaches was assessed in terms of bias and efficiency. The results were also compared with those obtained by direct-likelihood analysis. Since direct likelihood analysis is valid under the MAR mechanism (Verbeke and Molenberghs, 2000; Mallinckrodt et al, 2003a; Liang and Zeger, 1986; Molenberghs and Kenward, 2007), its results were presented and used as references against which these two approaches were contrasted.

Findings in general favoured MI over IPW. MI consistently outperformed IPW in terms of bias and efficiency. By considering both criteria simultaneously, MI and LMM approaches performed best under all three dropout rates when compared with the IPW approach. This is to be expected as both approaches are, respectively, Bayesian and likelihood model based which are valid under the assumption of MAR (Molenberghs and Kenward, 2007). Schafer and Graham (2002) stated that because MI relies on Bayesian arguments, its performance is similar to that of a likelihood method that makes similar assumptions. The MI approach was less biased and considerably less variable than the IPW approach. The lower variability achieved by the MI approach makes it desirable in most statistical analyses. This agrees with the theoretical results in that IPW can be less efficient and less powerful than Bayesian estimators under a well specified parametric model, see, Seaman and White (2011) and Schafer and Graham (2002). Given these results, it appears that either the MI or the LMM approaches for MAR dropouts with continuous outcomes are preferable to the IPW approach. The latter approach was irrespective of the type of parameter of interest, associated with greater estimation bias as well as less efficiency.

Our results further suggested that even though the mechanism of dropout was MAR, the performance of the IPW approach was unsatisfactory. This demonstrates that the IPW approach has shortcomings as shown clearly in current analysis results. This situation can be justified by some previous studies which show that IPW is more widely used in marginal models for discrete outcomes than for continuous outcomes, see, for example, Robins et al. (1995) and Fitzmaurice et al. (1995). Despite these shortcomings, the IPW approach has been the longitudinal binary approach of choice for the primary analysis for handling MAR dropout because of its simplicity as well as the ease with which it can be implemented. Here we refer to statistical softwares such as SPSS, STATA and SAS. Thus, the IPW approach might become attractive in specific circumstances. In particular, with regard to marginal structural models with discrete outcomes. We here believe that MI approach can be recommended over IPW as the default analysis for longitudinal data with continuous outcomes when MAR dropouts are valid.

In conclusion, we note that the use of the MI and the IPW approaches must be undertaken with care when the longitudinal data analyses have dropouts in continuous outcome. In addition, it is clear from our findings that the dropout has a substantial impact on the type of outcome. Thus, to sufficiently address dropout using MI and the IPW approaches, the effect must be thoroughly investigated by way of carefully designed simulation studies as well as a theoretical investigation.

Chapter 4

A comparative analysis of likelihood based and multiple imputation methods for incomplete longitudinal data with ignorable missingness*

4.1 Abstract

In this chapter, we carry out an application for analyzing incomplete longitudinal data with missing outcomes. Explanatory variables of interest are assumed to be completely observed. The chapter focuses on two methods, namely direct likelihood and multiple imputation, both based on the MAR assumption. We implement these techniques to data for a study with a continuous outcome. The analysis is done using PROC MIXED and PROCs (MI and MIANALYZE) in SAS. In order to explore the performance of the suggested methods, analysis is first performed on the complete

[*] A. Satty and H. Mwambi (In review). A comparative analysis of likelihood based and multiple imputation methods for incomplete longitudinal data with ignorable missingness. *REVSTAT Statistical Journal* - Presented October 28-30, 2011. Brague: Czech Republic, NC.

data (no data are missing). Next, we create missing values in the form of dropout then re-analyze the data in the presence of missing values to compare with the analysis with no missing values. The chapter is motivated by the need to assess the strength of the multiple imputation approach in comparison to direct likelihood approach. The advantages of the direct likelihood and multiple imputation approaches are discussed. The results show that both direct likelihood and multiple imputation methods offer high efficiency under ignorable non-response mechanism.

Keywords: Ignorable missingness, Likelihood-based, Direct likelihood, Dropout missing at random (MAR), Missing outcomes.

4.2 Introduction

Longitudinal studies represent one of the main design strategies employed in medical and social research. In longitudinal studies the response of interest is scheduled or planned to be measured repeatedly over time for each subject, experimental or observation unit. Observations that are repeatedly measured over time are bound to be correlated and some may be missing. Dropouts arise in longitudinal data whenever one or more of the measurements scheduled for participating subjects are not taken or not available due to reasons known or unknown to the researcher. Understanding the pattern of missing data is important before any analysis of data with missing values is attempted. The pattern of missing data is said to be monotone if subjects leave the study prematurely, i.e., any missing value is never followed by an observed value (Diggle, 1989; Heyting et al., 1992; Little, 1995). In the context of longitudinal studies, missingness predominantly occurs in the form of dropout in which subjects fail to complete the study for one reason or another. Dropout will be the form of missing data assumed in the current chapter.

Analyses of missing data depends very much on the assumptions made about the process that creates the missing values. The classification of missing values or nonresponse process proposed by Rubin (1976) gives three possible different mechanisms. A nonresponse process is defined as missing completely at random (MCAR) if the probability of the missingness does not depend on the measurements, observed or unobserved. A nonresponse process is defined as missing at random (MAR) if the probability of missingness does not depend on the unobserved measurements. A nonresponse process is defined as non-random if the probability of a missing value depends on unobserved measurements and may be on the observed measurements. Specific names for these mechanisms for the case of longitudinal data were coined by Diggle and Kenward (1994). In the following section, these classifications of missing data will be discussed briefly.

Proposed approaches for handling missing data in longitudinal studies use methods that are valid under the MAR assumption (Little and Rubin, 2002; Verbeke and Molenberghs, 2000; Molenberghs and Verbeke, 2005; Molenberghs and Kenward, 2007; William, 2000). Imputation strategies for longitudinal continuous responses have been an active area of research. Examples of this can be found in Schafer (1997), Schafer and Yucel (2002), Liu et al. (2000), Demirtas and Schafer (2003) and Van Burren and Van Rijckevorsel (1992). This chapter focuses on two approaches proposed by Mallinckrodt et al. (2003a, 2003b) and Lavori et al. (1995) to deal with incomplete longitudinal data. One of these methods is direct likelihood analysis. In this method, the observed data are used without deletion nor imputation. The strength of this method relies on the accurate formulation of the likelihood of the data as it is. To do so, under valid MAR mechanism, suitable adjustments can be made to parameters sometimes when data are prone to incompleteness due to the within-patient correlation. For incomplete longitudinal data, a mixed model only needs missing data be MAR (Mallinckrodt et al., 2003a, 2003b; Verbeke and Molenberghs, 2000). These mixed effects models permit the inclusion of subjects with missing values at some time points, including both dropout and intermittent missingness patterns (Verbeke and Molenberghs, 2000). For continuous outcome data, this leads to the general linear mixed model (Verbeke and Molenberghs, 2000).

A MAR-based method that has seen a number of applications recently is multiple imputation (Rubin, 1978; Rubin, 1987; Rubin and Schenker, 1986; Little and Rubin, 2002; Schafer, 1999). This method involves constructing a fixed number of complete data sets from an incomplete one by drawing from the conditional distribution of the unobserved outcomes, given the observed ones. These complete data sets are then analysed and the results combined to produce reliable inferences. The method is discussed in the context of continuous longitudinal data in Verbeke and Molenberghs (2000), while Molenberghs and Verbeke (2005) and Yuan (2000) illustrated how the SAS procedures (MI and MIANALYZE) can be used in this context. Multiple imputation is valid under the MAR condition as for direct likelihood, and therefore does not suffer from the problems encountered in most single imputation methods.

In this chapter, we carry out an application for analyzing incomplete longitudinal data with missing outcomes. The models considered here assume that the missing data are confined to the repeated measures outcome, and covariates information is fully observed. We assume that the dropouts are MAR. The chapter is concerned with the comparison of two techniques applied to incomplete longitudinal data set with missing outcome. The techniques discussed are: direct likelihood and multiple imputation methods. In order to investigate the performance of the methods, analysis is first performed on the full data (no data are missing). Then the results from the analysis of incomplete generated data using the two proposed methods are compared, and also with reference to results based on the complete data. The incomplete data was created by generating new data sets with missing observations which have a similar distributional properties as the original data set. This is done by generating dropout, assuming the dropout occurred at random. The data used is based on a clinical trial to compare three drugs on the heart rate of an individual as the outcome.

The outline of this chapter is as follows: The data structure and the missingness mechanisms are introduced in Section 4.3, as well as a formal framework for incomplete longitudinal data. In Section 4.4, an overview of methods for analyzing incomplete longitudinal data are given. In Section 4.5, we present an application including a description of the full data set used in the analysis and the study design. The dropout generation scheme is also discussed. Section 4.6, contains the results based on incomplete data and those based on actual-data. Finally, the chapter ends with a discussion and conclusion in Section 4.7.

4.3 Data structure and notation

Some notation is necessary to describe methods for analyzing incomplete longitudinal data with dropout. We will follow the terminology based on the standard framework of Rubin (1976) and Little and Rubin (1987) in formulating definitions for data structure and missing data mechanisms. Let Y_i=$(Y_{i1}, ..., Y_{in_i})' = (Y_i^o, Y_i^m)'$ be the outcome vector of n_i measurements for subject i, i=1,...,n, where Y_i^o represents the observed data part and Y_i^m denotes the missing data part. Let $R_i = (R_{i1}, ..., R_{in_i})'$ be the corresponding missing data indicator vector of the same dimension as Y_i, defined as

$$R_{ij} = \begin{cases} 1 & \text{if } Y_{ij} \text{ is observed,} \\ 0 & \qquad \text{otherwise.} \end{cases} \tag{4.1}$$

Complete data refers to the vector Y_i of planned measurements. This is the outcome vector that would have been recorded if no data had been missing. The vector R_i and the process that generates it are referred to as the missingness process. The R_i can be designed to represent participant dropout, and so it has a monotone pattern (Verbeke and Molenberghs, 2000; Xu and Blozis, 2011). The full data for the ith subject can be represented as (Y_i, R_i) and the joint probability for the data and missingness can be expressed as: $f(y_i, r_i \mid X_i, W_i, \theta, \xi) = f(y_i \mid X_i, \theta) f(r_i \mid y_i, W_i, \xi)$, where X_i and W_i are design matrices for the measurements and dropout mechanism, respectively, θ is the parameter vector associated with the measurement process and ξ is the parameter vector for the missingness process. According to the dependence of the missing data process on the response process, Little and Rubin (1987) and Rubin (1976) classified missing data mechanisms as: missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR). The missingness process is defined as MCAR if the probability of non-response is independent of the response; that is, $f(r_i \mid y_i, W_i, \xi) = f(r_i \mid W_i, \xi)$ and the missingness process is defined as MAR when the probability of non-response is depends on the observed values of the response; that is, $f(r_i \mid y_i, W_i, \xi) = f(r_i \mid y_i^o, W_i, \xi)$. Finally, the missingness process is defined as MNAR if neither the MCAR nor the MAR assumptions hold. That is, the probability of non-response depends on the missing outcomes and possibly on the observed outcomes. In this study, we assume a MAR mechanism for the dropout process.

When missingness is restricted to dropout or attrition, we can replace the vector R_i by a scalar variable DR_i, the dropout indicator, commonly defined as

$$DR_i = 1 + \sum_{j=1}^{n} R_{ij}. \tag{4.2}$$

For an incomplete dropout sequence, DR_i denotes the occasion at which dropout occurs. In the formulation described above, it is assumed that all subjects are observed on the first occasion so that DR_i takes values between 2 and $n + 1$. The maximum value $n + 1$ corresponds to a complete measurement sequence. According to Equation (4.2), an MCAR dropout mechanism occurs when $f(DR_i = dr_i \mid y_i, W_i, \xi) = f(DR_i = dr_i \mid W_i, \xi)$, MAR dropout mechanism, when $f(DR_i = dr_i \mid y_i, W_i, \xi) = f(DR_i = dr_i \mid Y_i^o, W_i, \xi)$ and MNAR dropout mechanism, when $f(DR_i = dr_i \mid y_i, W_i, \xi) = f(DR_i = dr_i \mid Y_i^m, Y_i^o, W_i, \xi)$.

4.4 Methodology for incomplete longitudinal data

Much of the literature involving missing data (or dropout) in longitudinal data pertain to the various techniques developed to handle the problem. This section is devoted to providing an overview of the various strategies in handling missing data in longitudinal studies. The techniques used in this chapter for dealing with missing data and for comparisons are: multiple imputation and direct likelihood analysis.

4.4.1 Multiple imputation

Multiple imputation was introduced by Rubin (1978). It has been discussed in some details in Rubin (1987), Rubin and Schenker (1986), Tanner and Wong (1987) and Little and Rubin (1987), they give excellent description of the technique. The key idea behind multiple imputation is to replace each missing value with a set of M plausible values (Rubin, 1996; Schafer, 1997). The resulting incomplete data sets obtained through imputation are then analyzed by using standard procedures for complete data and combining the results from these analyses. The technique require the assumption that the missingness mechanism be MAR. Thus, the multiple imputation procedure is accomplished through three distinct phases:

Step 1: Imputation - create M data sets from M imputations of missing data drawn from a different distribution for each missing variable.
Step 2: Analysis - analyze each of the M imputed data sets using standard statistical analysis.
Step 3: Data pooling - combine the results of the M analyses to provide one final conclusion or inference.

To discuss these steps in detail, we will follow the approach provided by Verbeke and Molenberghs (2000). Recall that we partitioned complete data (Y_i) into Y_i^o and Y_i^m to indicate observed and unobserved data, respectively. Multiple imputation fills in the missing data Y_i^m using the observed data Y_i^o several times, and then the completed data are used to estimate ξ. If we know the distribution of $Y = (Y_i^o, Y_i^m)$, depends on the parameter vector ξ, then we could impute Y_i^m by drawing a value of Y_i^m from the conditional distribution $f(y_i^m \mid y_i^o, \xi)$. Because $\hat{\xi}$ is a random variable, we must also take its variability into account in drawing imputations. In Bayesian terms, $\hat{\xi}$ is a random variable of which the distribution depends on the data. So we first obtain the posterior distribution of ξ from the data, a distribution which is a function of $\hat{\xi}$. Given this posterior distribution, imputation algorithm can be used to draw a random ξ^* from the distribution of ξ, and to put this ξ^* in to draw a random Y_i^m from $f(y_i^m \mid y_i^o, \xi^*)$, using the following steps: (1) Draw ξ^* from the distribution of ξ, (2) Draw Y_i^{m*} from $f(y_i^m \mid y_i^o, \xi^*)$, and (3) Use the complete data (Y^o, Y^{m*}) and the model to estimate β, and its estimated variance, using the complete data, (Y^o, Y^{m*}):

$$\hat{\beta_m} = \hat{\beta}(Y) = \hat{\beta}(Y^o, Y^{m*}), \tag{4.3}$$

where the within-imputation variance is $U_m = \hat{Var}(\hat{\beta})$. The steps described earlier are repeated independently M times, resulting in $\hat{\beta_m}$ and U_m, for $m = 1, ..., M$. Steps 1 and 2 are referred to as the imputation task, and step 3 is the estimation task (Verbeke and Molenberghs, 2005). Finally, the results are combined using the following steps also repeated for pooling the estimates obtained after M imputations (Rubin, 1987; Verbeke and Molenberghs, 2000). With no missing data, suppose the inference about the parameter β is made by $(\beta - \hat{\beta}) \sim N(0, U)$. The overall estimated parameter vector is the average of all individual estimates:

$$\hat{\beta}* = \frac{\sum_{m=1}^{M} \hat{\beta_m}}{M}, \tag{4.4}$$

with the normal-based inferences for β based upon $(\beta - \hat{\beta}*) \sim N(0, V)$ (Verbeke and Molenberghs, 2000). We obtain the variance as a weighted sum of the within-imputation variance and the between-imputations variability:

$$V = W + \left(\frac{M+1}{M}\right) B, \tag{4.5}$$

where

$$W = \frac{\sum_{m=1}^{M} U_m}{M} \tag{4.6}$$

defined to be the average within-imputation variance, and

$$B = \frac{\sum_{m=1}^{M} (\hat{\beta_m} - \hat{\beta}*)(\hat{\beta_m} - \hat{\beta}*)'}{M-1} \tag{4.7}$$

defined to be the between-imputation variance (Rubin, 1987).

4.4.2 Direct likelihood

An alternative method for handling missing data in longitudinal studies is the likelihood-based approach of using available data instead of imputation. Likelihood-based mixed effects models which are valid under the MAR assumption were proposed by Laird and Ware (1982) for continuous outcomes. This likelihood-based MAR analysis is also termed likelihood-based ignorable analysis, or direct likelihood analysis (Verbeke and Molenberghs, 2005). In contrast to the MI approach, direct likelihood analysis uses the observed data without the need of neither deletion nor imputation. In other words, no additional data manipulation is necessary when a direct likelihood analysis is envisaged, provided the software tool used for analysis is able to handle measurement sequences of unequal length (Molenberghs and Kenward, 2007). To do so, under valid MAR assumption, suitable adjustments can be made to parameters at times when data are prone to incompleteness due to the within-subject correlation. Thus, even when interest lies in a comparison between two treatment groups at the last measurement time, such a likelihood analysis can be conducted without problems since the fitted model can be used as the basis for inference. When a MAR mechanism is valid, a direct likelihood analysis can be obtained with no need for modelling the missingness process. It is increasingly preferred over ad hoc methods, particularly when tools like the generalized linear mixed mixed effect models (Molenberghs and Verbeke, 2005) are assumed. The major advantage of this method is its simplicity, it can also be fitted in standard statistical software without involving additional programming, using such tools as SAS software, PROCs MIXED, GLIMMIX and NLMIXED. The use of these procedures have been illustrated by Verbeke and Molenberghs (2000) and Molenberghs and Verbeke (2005). A useful summary for these procedures is presented by Molenberghs and Kenward (2007). Despite the flexibility and ease of implementation of direct likelihood method, there are fundamental issues when selecting a model and assessing its fit to the observed data which do not occur with complete data. The method is sensible under linear mixed models in combination with the assumption of ignorability. Such an approach, tailored to the needs of clinical trials, has been proposed by Mallinckrodt et al. (2001a, 2001b). For incomplete longitudinal data context, a mixed model only needs missing data to be MAR. According to Verbeke and Molenberghs (2000), these mixed-effect models permit the inclusion of subjects with missing values at some time points for both missing data patterns, namely monotone and intermittent. Since direct likelihood ideas can be used with a variety of likelihoods, in this study we consider the general linear mixed-effects model (Laird and Ware, 1982) as a key modelling framework which can be combined with the ignorability assumption. For Y_i the vector of observation from individual i, the model can be written as follow

$$Y_i = X_i\beta + Z_ib_i + \varepsilon_i, \tag{4.8}$$

where $b_i \sim N(0, D)$, $\varepsilon_i \sim N(0, \Sigma_i)$ and $b_1, ..., b_N, \varepsilon_1, ..., \varepsilon_N$ are independent. The meaning of each form in Equation (4.8) are described as follows. The outcome Y_i is the n_i dimensional response vector for subject i, containing the outcomes at n_i various measurement occasions, $1 \leq i \leq N$, N is the number of subjects, X_i and Z_i are $(n_i \times p)$ and $(n_i \times q)$ dimensional matrices of known covariates, β is the p-dimensional vector containing the fixed effects, b_i is the q-dimensional vector containing the random effects and ε_i is a n_i dimensional vector of residual components, combining measurement error and serial correlation. Finally, D is a general $(q \times q)$ covariance matrix whose (i, j)th element is $d_{ij} = d_{ji}$ and Σ_i is a $(n_i \times n_i)$ covariance matrix which generally depends on i only through its dimension n_i, i.e., the set of unknown parameters in Σ_i will not depend upon i. This means marginally

$$Y_i \sim N(X_i\beta, Z_iDZ_i' + \Sigma_i). \tag{4.9}$$

Thus, if we define $V_i = Z_iDZ_i' + \Sigma_i$ as the general covariance matrix of Y_i, then,

$$f(y_{ij}, \beta, V_i) = (2\Pi)^{\frac{-n}{2}}|V_i|^{\frac{-1}{2}} \exp\{-(y_i - X_i\beta)'V_i^{-1}(y_i - X_i\beta)/2\}, \tag{4.10}$$

from which a marginal likelihood can be constructed to estimate β. In the likelihood context, Little and Rubin (1987) and Rubin (1976) stated that when the MAR assumption and mild regularity conditions hold, parameters θ and ξ are independent, and that likelihood based inference is valid when the missing data mechanism is ignored. In practice, the likelihood of interest is then based on the factor $f(y_i^o \mid \xi)$ (Verbeke and Molenberghs, 2000). This is referred as ignorability.

4.5 Application example

Next, we consider the application of the methods described earlier for handling missing data in longitudinal studies with continuous outcomes. We present the details of the study design, data set and the generation of missing values used in this chapter. The methods are applied to data from a published clinical trial data on the effect of the three drugs on individuals heart rate. This data comes from Milliken and Johnson (2009).

4.5.1 The design of the study

To examine the performance of direct likelihood and multiple imputation methods, four steps were planned. The steps were as follow: First, a model was fitted to the full data (no data are missing), thus producing what we refer as true full data estimates. Second, we generated incomplete data at 10%, 15% and 20% dropout rates in the outcome (selected at random) variable using defined rules to achieve the required mechanism under MAR assumption. Third, the resulting incomplete data was analyzed using the two different methods using multiple imputation and direct likelihood. Fourth, results from the complete and incomplete data analysis were compared. The true full data results were presented and used as references. The study aimed to investigate how direct likelihood and multiple imputation compare to each other and to the true analysis. Note that we adopt the use of "true full data" terminology to differentiate this from full data from multiple imputed data sets. These are clearly different complete data sets.

4.5.2 Data set - heart rates trial

This example is not intended to be an extensive simulation study, but rather simple scenarios for comparison of the methods. The results are presented to demonstrate the potential advantages or disadvantages of the two strategies for handling missing data. Like designed experiments using split-plot designs, experiments utilizing repeated measures designs have structures that involve more than one size of experimental unit. A subject may be measured over time where time is one of the factors in the treatment structure of the experiment. By measuring the subject at several different times, the subject is essentially being split into parts (time intervals) and the response is measured on each part. However, the subtle distinction with the split plot design is that for repeated measures the subplot treatments (time intervals) can not be randomly applied to subplot units. The larger experimental unit is the subject or the collection of time intervals. The smaller unit is the interval of time during which the subject is exposed to a treatment or an interval just between time measurements. As a typical longitudinal example, we analyze the effect of three drugs on heart rate, while taking account of within subject correlation. The study was reported in detail in Milliken and Johnson (2009). Repeated measures designs can be applied in clinical trials involving two or more treatments.

This example investigates the effects of three treatments involving two active treatments and a control (AX23, BWW9 and CTRL) on heart rates, where each treatment was randomized to female individuals and each patient observed over four time periods. Specifically, each patient's heart rate was measured 5, 10, 15 and 20 minutes after administering the treatment. The only constraint is that the time intervals are not randomly distributed within an individual. The example was used in Milliken and Johnson (2009) to demonstrate analyses of repeated measures designs and to show how to determine estimates of effects of scientific interest and provide methods to study contrasts of interest. In our case, we use the data to achieve a comparative analysis of two methods to deal with missing data. A model which is used to describe the data is similar to a split-plot in a completely randomized design. The model is

$$H_{ijk} = \mu + Time_j + \delta_{ik} + Drug_k + (Time * Drug)_{jk} + \varepsilon_{ijk}, \qquad (4.11)$$

where H_{ijk} is the heart rate of individual i at time j on drug k, $i = 1, ..., 8$, $j = 1, 2, 3, 4$ and $k = 1, 2, 3$. The model has two error terms: δ_{ik} represents a subject within drug random effect, and ε_{ijk} represents a time error component. The ideal conditions for a split-plot in time analysis is that: (1) the δ_{ik} are independently and identically $N(0, \sigma_\delta^2)$, (2) the ε_{ijk} are independently and identically $N(0, \sigma_\varepsilon^2)$, and (3) the δ_{ik} and ε_{ijk} are all independent of one another.

The main purpose of this example is to investigate the effects of the three drugs. Thus, the type III tests of fixed effects and the differences between effects were the quantities of interest in the study. The primary null hypothesis (the difference between the drug main effects) will be tested. The null hypothesis is that of no difference among drugs. The significance of differences in least-square means is based on Type III tests. These tests examine the significance of each partial effect; that is, the significance of an effect with all the other effects in the model. In the analysis results we present the significance of drug main effects, time main effects and the interaction of time and drug effects.

4.5.3 Analysis of missing values

Since there are no missing values in the example data set described above, it provides us a convenient platform to design a comparative study to compare the two methods to deal with missing data using the results from the true complete data analysis as the reference. We carry out an application study to generate the data set with dropouts.

In this application, we distinguish between two stages: (1) The dropout generation stage. (2) The analysis stage.

4.5.3.1 Generating missing data

In the first stage, we use the full data set to artificially generate missing values by mimicking the dropout at random. From the complete data, we draw 1000 random samples of N=96. The incomplete data was generated with 10%, 15% and 20% dropout rate. A dropout rate was implemented via a missing data generation process, assuming the dropout depends only on the observed data. Furthermore, a monotone dropout pattern was imposed in the heart rate (outcome of interest); that is, if H_{ij} is missing, then H_{is} is missing for $s \geq j$. The explanatory variables drug, time and interaction between drug and time are assumed to be fully observed. In addition, in order to create our dropout model, we assume that dropout can occur only after the first two time points. Namely, dropout is based on values of H, assuming the H is fully observed in the first two time (time = 1, 2), while for the later times (time = 3, 4) some dropouts may occur. We assume an MAR mechanism for the dropout process and the dropout mechanism depends on individual previously observed values of one of the endpoints. For MAR mechanism, H was made missing if its measurements exceeded 75 (the baseline mean for heart rate) at the previous occasion time period, beginning with the second post baseline observation. Thus in the generation, the missingness at time = 3, 4 was dependent on the most recently observed values. This was done to achieve the required mechanism under MAR assumption.

4.5.3.2 Computations and handling missing data

After implementing the missing data mechanism and thus generating the data set with dropout, the next step was to deal with the dropout. Handling dropout was carried out using direct likelihood analysis and multiple imputation methods with functions available in the SAS software package. Ultimately, likelihood, multiple imputation and analysis results from the true full data set can be compared in terms of their impact on various linear mixed model aspects (fixed effects and least squares means). The proposed methods dealt with the dropout according to the following steps:

- Imputing dropouts using multiple imputation techniques. This was achieved using PROCs MI, MIXED and MIANALYZE with an LSMEANS option. The imputation model is based on model (4.11) which assumes normality of the outcomes. For the dropout under MAR, the imputation model should be specified (Rubin, 1987). Thus, in the imputation model, we included all the available data (including the outcome, H) to predict the dropouts since they were potentially related to the imputed variable as well as to the missingness of the imputed variable. This means we used variables in the analysis model, variables associated with missingness of the imputed variable and variables correlated with the imputed variable. This was done to increase the plausibility of the MAR assumption, as well as to improve the accuracy and efficiency of the imputation. Once the multiple imputation model is chosen, the number of imputations must be decided. PROC MI was applied to generate M=5 complete data sets. We fixed the number of multiple imputations at M=5, since the 5 imputed data sets are sufficient (see, Schafer and Olsen, 1998; Schafer, 1999). PROC MIXED was used to set up effect parameterizations for the class variables and we used ODS statement output to create output data sets that match PROC MIANALYZE for combining the effect means estimates from the 5 imputed data sets. While PROC MIANALYZE cannot directly combine the least square means and their differences to obtain the effect means of drug and contrasts between drug groups from PROC MIXED, the LSMEANS table was sorted differently so that we used the BY statement in PROC MIANALYZE to read it in.

- For comparison, the data was analyzed as they are, consistent with ignorability for direct likelihood analysis implemented with PROC MIXED with LSMEANS option. The REPEATED statement was used, and obtained the effect means estimates from the generated data set. Parameters were estimated using Restricted Maximum Likelihood with the Newton-Raphson algorithm.

4.6 Results

As mentioned earlier, the purpose of the chapter is to present a comparative study of two strategies dealing with incomplete longitudinal data with missing outcomes. This section illustrates results of our example in the tables below. Overall, we find that the performance of the direct likelihood and multiple imputation methods shows sufficiently strong similarities and in many cases yielded similar results to those based on the true complete data set, with few exceptions. Note that due to the similarities in the findings under three dropout rates, the results for the 10% and 20% dropout rates are not presented.

Table 4.1: *Statistical test for drug, time and drug × time effects of complete data, direct likelihood and multiple imputation, under 15% dropout rate*

		Type III tests of fixed effects			
	Effect	Num *df*	Den *df*	F-value	$Pr > F$
Actual-data					
	drug	2	21	5.99	0.0088
	time	3	63	12.96	< 0.0001
	drug x time	6	63	11.80	< 0.0001
Direct likelihood					
	drug	2	17.1	7.78	0.0040
	time	3	15.8	18.13	< 0.0001
	drug x time	6	15.8	25.74	< 0.0001
Multiple imputation					
	drug	2	21	7.14	0.0043
	time	3	447	84.15	< 0.0001
	drug x time	6	447	76.00	< 0.0001

The results which show the significance of the effects using direct likelihood and multiple imputation to handle dropout are presented in Table 4.1. Compared with the results based on the complete data set, we see that type III tests of fixed effects show that both direct likelihood and multiple imputation methods yielded statistically similar results. The analysis shows that the drug effect has significant p-values of

< 0.0040 and < 0.0043 for direct likelihood and multiple imputation, respectively, indicating a rejection of the null hypothesis of equal drug means. The p-value of the drug effect under multiple imputation (0.0043) was slightly higher in comparison to that of the direct likelihood analysis (0.0040), but both methods indicate strong evidence of significance compared to the p-value of 0.0088 for the original complete data set. Evidently, there are no extreme differences between the direct likelihood and multiple imputation methods. However, the p-value for the drug effect was significantly reduced by about 50% compared to the actual data p-value. This indicates a real problem with dropout, both multiple imputation and direct likelihood may lead to rejection of the null hypothesis with a higher probability than would be the case if the data were complete. The test of significance for the time effect in type III tests of fixed effects produced significant p-values of < 0.0001 in both methods. The test for the interaction between drug and time effects gave a p-value of < 0.0001 in both methods, indicating a strong evidence of time dependence on the drug effects. Generally, the proposed methods presented acceptable performance with respect to estimates of p-values in all cases when compared to that based on actual data. In two cases, namely p-values of time effect and interaction drug × time, the methods yielded the same results as those for complete data.

The drug main effect means, the time main effect means and the drug time two-way means are given in Tables 4.2, 4.3 and 4.4 for actual-data, direct likelihood and multiple imputation analysis, respectively. When compared with the main effect means based on the actual data set, treating the data with direct likelihood and multiple imputation methods appeared to have resulted either in the same main effects means (with regard to p-values) or closer to each other (with regard to parameter estimates and standard errors). The p-values of < 0.0001 for drug, time and drug × time main effects means were the same for both methods. Both methods yielded significant p-values for all cases. Furthermore, the parameter estimates for direct likelihood and multiple imputation analysis were generally (though not always) closer to each other. In some cases (estimates of time 1, time 2 and interactions between all drugs and time 1 and time 2), the methods offered the same estimates as compared to the estimates from complete data. Such results are expected considering the fact that the first and

Table 4.2: *Drug main effect means, time main effect means and drug × time main effect means of the complete data*

MIXED Least Squares Means

Effect	Drug	Time	Estimate	Standard Error	$Pr > \|t\|$
drug	AX23		76.2813	1.8652	< 0.0001
drug	BWW9		81.0312	1.8652	< 0.0001
drug	CTRL		71.9062	1.8652	< 0.0001
time		1	75.0000	1.1800	< 0.0001
time		2	78.9583	1.1800	< 0.0001
time		3	77.0417	1.1800	< 0.0001
time		4	74.6250	1.1800	< 0.0001
drug x time	AX23	1	70.5000	2.0438	< 0.0001
drug x time	AX23	2	80.5000	2.0438	< 0.0001
drug x time	AX23	3	81.0000	2.0438	< 0.0001
drug x time	AX23	4	73.1250	2.0438	< 0.0001
drug x time	BWW9	1	81.7500	2.0438	< 0.0001
drug x time	BWW9	2	84.0000	2.0438	< 0.0001
drug x time	BWW9	3	78.6250	2.0438	< 0.0001
drug x time	BWW9	4	79.7500	2.0438	< 0.0001
drug x time	CTRL	1	72.7500	2.0438	< 0.0001
drug x time	CTRL	2	72.3750	2.0438	< 0.0001
drug x time	CTRL	3	71.5000	2.0438	< 0.0001
drug x time	CTRL	4	71.0000	2.0438	< 0.0001

second time points contained observed data for all patients considered in the analysis. The standard errors associated with both methods were much closer to those from the complete data set, and only slightly different from each other.

As discussed above, we find significant drug effects in the data, thus one would ideally need to compare various means with each other. If there is no drug × time interaction, then we will often need to make comparisons between the drug main effect means and the time main effect means. Since the interaction effect means is significant (as shown in our results), we need to compare the drugs with one another at each

Table 4.3: *Drug main effect means, time main effect means and drug × time main effect means of the direct likelihood analysis, for 15% dropout rate*

MIXED Least Squares Means

Effect	Drug	Time	Estimate	Standard Error	$Pr > \|t\|$
drug	AX23		76.2666	1.8341	< 0.0001
drug	BWW9		81.5009	1.8447	< 0.0001
drug	CTRL		71.1112	1.8816	< 0.0001
time		1	75.0000	1.1277	< 0.0001
time		2	78.9583	1.1785	< 0.0001
time		3	76.8235	1.1868	< 0.0001
time		4	74.3898	1.2297	< 0.0001
drug x time	AX23	1	70.5000	2.9533	< 0.0001
drug x time	AX23	2	80.5000	2.9533	< 0.0001
drug x time	AX23	3	81.0873	2.8143	< 0.0001
drug x time	AX23	4	73.9793	2.0075	< 0.0001
drug x time	BWW9	1	81.7500	2.9533	< 0.0001
drug x time	BWW9	2	84.0000	2.2145	< 0.0001
drug x time	BWW9	3	78.7494	2.8518	< 0.0001
drug x time	BWW9	4	80.5040	2.0751	< 0.0001
drug x time	CTRL	1	72.7500	2.9533	< 0.0001
drug x time	CTRL	2	72.3750	2.2145	< 0.0001
drug x time	CTRL	3	70.6338	2.9775	< 0.0001
drug x time	CTRL	4	70.6860	2.2962	< 0.0001

time point and/or times to one another for each drug. Comparisons of the time means within a drug are given in Figure 4.1. Since the levels of time are quantitative and equally spaced, orthogonal polynomials can be used to check for linear and quadratic trends over time for each drug. The linear and quadratic trends in time for all drugs reveal that drug BWW9 shows a negative linear trend, and drug AX23 shows a strong quadratic trend in all methods. Evidently, the differences occurred with drug CTRL in graphs (b) and (c) for direct likelihood and multiple imputation, respectively. Both methods yielded slightly different linear trends as compared to that from actual data.

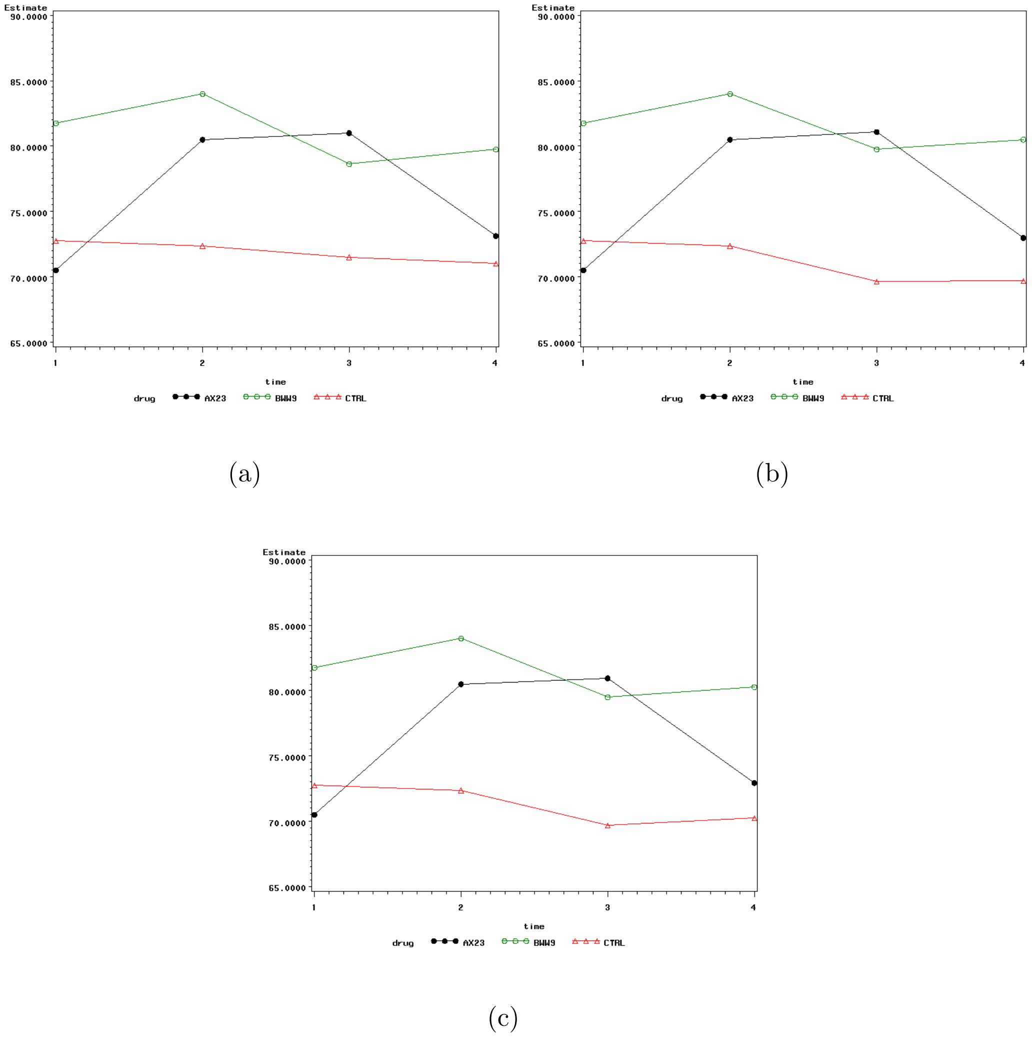
Table 4.4: *Drug main effect means, time main effect means and drug × time main effect means of multiple imputation, for 15% dropout rate*

MIXED Least Squares Means

Effect	Drug	Time	Estimate	Standard Error	$Pr > \|t\|$
drug	AX23		76.2202	1.8528	< 0.0001
drug	BWW9		81.3778	1.8528	< 0.0001
drug	CTRL		71.2641	1.8528	< 0.0001
time		1	75.0000	1.1793	< 0.0001
time		2	78.9583	1.1793	< 0.0001
time		3	77.7089	1.1793	< 0.0001
time		4	74.4822	1.1793	< 0.0001
drug x time	AX23	1	70.5000	2.0213	< 0.0001
drug x time	AX23	2	80.5000	2.0213	< 0.0001
drug x time	AX23	3	81.9468	2.0213	< 0.0001
drug x time	AX23	4	73.9339	2.0213	< 0.0001
drug x time	BWW9	1	81.7500	2.0213	< 0.0001
drug x time	BWW9	2	84.0000	2.0213	< 0.0001
drug x time	BWW9	3	78.4915	2.0213	< 0.0001
drug x time	BWW9	4	80.2697	2.0213	< 0.0001
drug x time	CTRL	1	72.7500	2.0213	< 0.0001
drug x time	CTRL	2	72.3750	2.0213	< 0.0001
drug x time	CTRL	3	70.6885	2.0213	< 0.0001
drug x time	CTRL	4	71.2429	2.0213	< 0.0001

The graph in Figure 4.1 displays these relationships.

We continued to explore the pairwise comparisons among drug main effect means and time main effect means by comparing the estimates, standard errors and *p*-values under the both methods in Table 4.5. We obtained a non-significant effect means from all our different models with respect to pairwise comparisons between drugs (AX23, BWW9) and (AX23, CTRL), indicating that there is no overall drug effect means difference among (AX23, BWW9) and (AX23, CTRL). However, assessment of the effect means among time 1 and time 4 leads to *p*-value=0.6357, *p*-value=0.5354 and



(a)

(b)

(c)

Figure 4.1: *(a) Actual data - Means over time for each drug for the heart rate data. (b) Direct likelihood - Means over time for each drug for the heart rate data. (c) Multiple imputation - Means over time for each drug for the heart rate data*

p-value=0.5969 for actual data, direct likelihood and multiple imputation, respectively, where all three strategies yielded non-significant effect means. Both approaches yielded very similar estimates and standard errors of pairwise comparisons among drug main effect means and time main effect means compared to those from the complete data set. Nevertheless, the estimates of pairwise comparisons among effect means of (time 1, time 4), (time 2, time 3) and (drug AX23, drug CTRL) were somewhat slightly different from those of the complete data. Whereas, the other estimates from both methods were very close to the actual estimates, and in one case (pairwise comparison among effect mean of time 1 and time 2), their estimates were the same as that from the actual data. This is to be expected as there were no unobserved data in the first and second time points for all patients.

Table 4.5: *Pairwise comparisons among drug main effect means and time main effect means of complete data and generated data using direct likelihood and multiple imputation methods, for 15% dropout rate*: ***Differences of least squares means***

Method	Effect	Drug	Time	Drug	Time	Estimate	Standard Error	$Pr > \|t\|$
Actual data								
	drug	AX23		BWW9		-5.7500	2.6378	0.0861
	drug	AX23		CTRL		4.0050	2.6378	0.1121
	drug	BWW9		CTRL		9.1250	2.6378	0.0023
	time		1		2	-3.9583	0.7878	< 0.0001
	time		1		3	-1.0417	0.7878	0.0119
	time		1		4	0.3750	0.7878	0.6357
	time		2		3	1.9167	0.7878	0.0178
	time		2		4	4.3333	0.7878	< 0.0001
	time		3		4	2.4167	0.7878	0.0032
Direct likelihood								
	drug	AX23		BWW9		-5.2342	2.6013	0.0607
	drug	AX23		CTRL		5.1554	2.6276	0.1661
	drug	BWW9		CTRL		9.2896	2.6350	0.0010
	time		1		2	-3.9583	0.7201	< 0.0001
	time		1		3	-1.0235	0.7601	0.0139
	time		1		4	0.5102	0.7650	0.5354
	time		2		3	2.1349	0.7417	0.0252
	time		2		4	4.3686	0.7583	< 0.0005
	time		3		4	2.4337	0.7501	0.0055
Multiple imputation								
	drug	AX23		BWW9		-5.7576	2.6768	0.0776
	drug	AX23		CTRL		4.9961	2.6768	0.1282
	drug	BWW9		CTRL		9.1137	2.6768	0.0011
	time		1		2	-3.9583	0.7713	< 0.0001
	time		1		3	-1.0389	0.7713	< 0.0001
	time		1		4	0.4178	0.7713	0.5969
	time		2		3	2.0494	0.7713	< 0.0001
	time		2		4	4.4762	0.7713	< 0.0001
	time		3		4	2.4267	0.7713	< 0.0001

4.7 Discussion and conclusion

In this chapter, we have discussed a common problem encountered in longitudinal studies namely that of missing data when some variables have missing values in some units. In particular, we consider the problem of longitudinal missing data when the outcome has missing values due to dropout. The chapter is concerned with comparison of two techniques applied to an incomplete longitudinal data set with continuous outcomes. The techniques we compared were direct likelihood and multiple imputation methods. In order to investigate the performance of the proposed methods, analysis was first performed on the complete data and then the results from the two methods to handle missing data were compared to those based on complete data. We generated a new data set with missing outcome values which had similar distributional properties to the original data set. This was done by generating a dropout process, assuming the dropout was at random. The model considered assumed that dropouts are confined to the repeated measures outcome and covariates information is fully observed. We used data to study the effect of three treatments on heart rates.

The findings of our application in general noted that both direct likelihood and multiple imputation performed best under all three dropout rates, and they are more broadly similar in results. This is to be expected as both approaches are likelihood based and Bayesian analysis, respectively, and therefore valid under the assumption of MAR (Molenberghs and Kenward, 2007). The findings of direct likelihood are in line with the findings that likelihood-based analyses are appropriate under ignorability situation (Verbeke and Molenberghs, 2000; Molenberghs and Verbeke, 2005; Mallinckrodt et al., 2001a, 2001b). Because of simplicity, and ease of implementation using many statistical tools such as the SAS software procedures MIXED, NLMIXED and GLIMMIX, direct likelihood might be adequate to deal with dropout data when the MAR mechanism holds, provided appropriate distributional assumptions for a likelihood formulation of the data also hold. Moreover, a method such as multiple imputation can be conducted without problems using statistical software such as SAS procedures MI and MIANALYZE, and if done correctly, is a versatile, powerful and reliable technique to deal with dropouts that are MAR in longitudinal data with continuous outcomes.

It would appear that the recommendation of Mallinckrodt et al. (2003a, 2003b) to use direct likelihood and multiple imputation for dealing with incomplete longitudinal data with continuous outcomes is strongly supported by the current analysis.

At this point, we have to make it clear that the scope of this chapter is limited to direct likelihood and multiple imputation strategies. We note that there are several other strategies available to deal with incomplete longitudinal data with continuous outcome under the ignorability assumption however these methods are not covered in this study. For instance, the EM algorithm by Dempster et al. (1977) which is an alternative method to use for handling incomplete longitudinal data, as well as the Mixed-effects Model Repeated-Measures analysis (MMRM) by Mallinckrodt et al. (2001a, 2001b) which is a particular form of a linear mixed model fitted within direct likelihood analyses, since they are valid under the MAR assumption.

Chapter 5

Different methods for handling non-Gaussian longitudinal outcome subject to potentially random dropout*

5.1 Abstract

The present chapter compares and contrasts several statistical methods for analyzing incomplete non-Gaussian longitudinal outcomes when the underlying study is subject to dropout. We focus here on binary outcomes. The methods that are considered include weighted generalized estimating equations (WGEE), multiple imputation after generalized estimating equations (MI- GEE) and generalized linear mixed models (GLMM). The chapter aims to explore the performance of the above methods in terms of handling dropouts that are missing at random (MAR). The methods are compared on simulated data. The correlated binary variables are generated from a random effects model. Dropouts are generated under several different dropout rates and sample sizes.

* A. Satty, H. Mwambi and G. Molenberghs (In review). Different methods for handling non-Gaussian longitudinal outcome subject to potentially random dropout. *Communications in Statistics - Simulation and Computation.*

The comparison will be made through the evaluation of bias, accuracy and mean square error. MI-GEE was considerably robust, doing better than all the other methods in terms of the small and large sample sizes, regardless of the dropout rates.

Keywords: Multiple imputation GEE, Weighted GEE, Generalized linear mixed model (GLMM), Incomplete non-Gaussian longitudinal outcome, Random dropout.

5.2 Introduction

Longitudinal non-Gaussian studies repeatedly measure the outcome and covariates over a series of time. However, data arising from such studies often show inevitable incompleteness due to dropouts or lack of follow-up. To be precise, an individual's outcome can be missing at one follow-up time and be measured at the next follow-up time. This leads to a large class of missing data patterns. The current chapter, however, focuses on the monotone missing data pattern that results from attrition, in the sense that when an individual drops out from the study, no more measurements are obtained on that individual. Where there are dropouts, the choice of statistical methods for handling incompleteness has important implications on the estimation of the results since several statistical methods are appropriate only for certain missing data mechanisms. Thus, it is important to address the mechanisms that govern dropout. Based on definitions given by Rubin (1976) and Little and Rubin (1987), dropout mechanisms can be classified as missing completely at random (MCAR) which means the dropout process is independent of both unobserved and observed data, missing at random (MAR) if, given the observed outcomes, the dropout process is independent of the unobserved outcomes, i.e., depends only on the observed outcomes and possibly on covariates, and missing not random (MNAR) when the dropout process is dependent on the unobserved data and possibly on the observed data.

Molenberghs and Verbeke (2005) distinguished between various families of models to model longitudinal non-Gaussian data, namely marginal, random effects (or subject-specific) and conditional models. In this article, we consider the generalized linear mixed model (GLMM) (Breslow and Clayton, 1993) as a random effects model that is typically estimated through maximum likelihood (Jansen et al., 2006). An early instance of a random-effects model is the beta-binomial model (Skellam, 1948). Thorough discussions on GLMM can be found in Fitzmaurice et al. (2004), Molenberghs and Verbeke (2005) and Jansen et al. (2006). In the GLMM, the measurement model and the dropout model are both specified, and the inference is based on maximizing the likelihood function, conditional on the observed data as well as the dropout process. Such models give valid inferences under the restrictive assumption of MAR, where the specification of a dropout model is not necessary, and inference is based on the likelihood function conditional on the observed data alone (Molenberghs and Kenward, 2007). In other words, when data are MAR, parameters of the measurement process are not involved in the dropout process which is to say that a likelihood based analysis provides valid inferences, with no need to impute, delete, or weight.

In the case of non-likelihood marginal models, the semi-parametric method of generalized estimating equations (GEE) by Liang and Zeger (1986) has been widely applied for handling dropouts (Liang and Zeger, 1986). However, GEE requires the stronger MCAR mechanism to hold (Laird, 1988; Liang and Zeger, 1986). This can be seen by the fact that GEE no longer has zero expectation when a MAR mechanism holds. So, GEE requires the strong MCAR assumption for the missing data mechanism to be ignorable. Two subsequent modifications of the GEE method were proposed to make it valid under the more general MAR condition: weighted generalized estimating equations (WGEE) and multiple imputation after generalized estimating equations (MI-GEE). Robins et al. (1995) devised WGEE extending GEE and which requires MAR rather than the much stronger MCAR mechanism, but needs the specification of a dropout model with regard to observed outcomes or covariates, in view of specifying the weights. WGEE involves weighting response measurements by their inverse probability of being observed, estimated from some assumed dropout model (Robins et al., 1995). The idea of WGEE was first discussed in Cochran (1977) where estimation is based on the observed responses after weighting them to account for the probability (propensity) of dropout. Early account of WGEE can be found in Robins et al. (1995) and Fitzmaurice et al. (1995).

An alternative approach that is valid under the weaker MAR assumption is multiple imputation after generalized estimating equations, or, as we will term it in the remainder of this article, a (MI-GEE). MI-GEE denotes a method based on a combination of MI and GEE model analysis. The primary idea of the combination of MI and GEE comes from Schafer (2003). He proposed an alternative mode of analysis based on the following steps:

Step 1: Impute the missing outcomes multiple times using a full-parametric model, such as a random effects type model.

Step 2: After drawing the imputations, analyze the so-completed data sets using a conventional marginal model, such as the GEE method.

Step 3: Finally, perform MI inference on the so-analyzed sets of data.

As pointed out by Beunckens et al. (2008), MI-GEE comes down to first using the predictive distribution of the unobserved outcomes, conditional on the observed ones and covariates. Thereafter, when MAR is valid, the missing data can be ignored in the analysis. In terms of the dropout mechanism, in the MI-GEE method, the imputation model needs to be specified. This specification can be done by an imputation model that imputes the missing values with a given set of plausible values (Beunckens et al., 2008). Details of this method can be found in Molenberghs and Kenward (2007), Beunckens et al. (2008), Yoo (2009) and Birhanu et al. (2011).

In closely related studies, Beunckens et al. (2008) studied the comparison between the two GEE versions (WGEE and MI-GEE), and Birhanu et al. (2011) compared the efficiency and robustness of WGEE, MI-GEE and doubly robust GEE (DR-GEE). In this chapter, however, we restrict attention to study how the two types of GEE (WGEE and MI-GEE) compare to the likelihood-based GLMM for analyzing non-Gaussian longitudinal outcomes with dropout. The primary objective of the present study is to investigate the performance of WGEE, MI-GEE and GLMM for handling incomplete non-Gaussian longitudinal data. The dropout mechanism is assumed to be MAR. The methods are compared using simulated data sets under several different dropout rates and sample sizes. A comparison will be made through the evaluation of bias, efficiency and mean square error. Note that the parameters in a marginal model, such as GEE, and a hierarchical model, such as GLMM, do not have the same interpretation. Indeed, the fixed effects in the latter are to be interpreted conditional upon the random effect.

While there is no difference between the two in the linear mixed model, this is not the case for non-Gaussian outcomes, in particular for binary data. Fortunately, as stated in Molenberghs and Verbeke (2005) and references therein, the GLMM parameters can be approximately transformed to their marginal counterpart. In particular, when the random-effects structure is confined to a random intercept b_i, normally distributed with mean 0 and variance σ, then the ratio between the marginal and random effects parameter is approximately equal to $\sqrt{1+c^2\sigma}$, where $c = 16\sqrt{3}/(15\pi)$. This ratio will be used in our simulation study to make the parameters comparable. We focus here on binary outcomes. Similar results will likely apply to other data types as well but should, ideally, be the subject of additional research. The outline of this chapter is as follows. In Section 5.3, the data setting and necessary notation in terms of the dropout mechanism are introduced. In Section 5.4, an overview of methods for analyzing incomplete longitudinal non-Gaussian data is given with the focus on WGEE, MI-GEE and GLMM. Section 5.5 presents the simulation study scheme including the study design, data generation and the evaluation criteria used in the analysis. The results of the simulation are presented in Section 5.6. Finally, a brief discussion and concluding remarks are provided in Section 5.7.

5.3 Data setting and notation

Let Y_{ij} be the response measurement of individual i at time j, where $i = 1, 2, ...N$ and $j = 1, 2, ...n_i$, which can be observed or missing. Let R_{ij} be an indicator variable, where $R_{ij} = 1$ if Y_{ij} is observed and $R_{ij} = 0$ if Y_{ij} is missing. Therefore, corresponding to the ith individual's set of measurements, denoted by $Y_i = (Y_{i1}, Y_{i2}, ..., Y_{in_i})$, there is a $(1 \times n_i)$ vector for the dropout indicators, $R_i = (R_{i1}, R_{i2}, ..., R_{in_i})$. Considering the missingness due to dropout, we split Y_i as $Y_i = (Y_i^o, Y_i^m)$, representing observed and unobserved measurements, respectively. We additionally define D_i to be the dropout indicator for each individual i, where $D_i = 1 + \Sigma_{j=1}^{n_i} R_{ij}$ which measures the occasion when the dropout occurs. In principle, one often needs to consider the density of full data $f(y_i, r_i \mid X_i, \theta, \gamma)$. So, we use the parameter vectors θ and γ to indicate the measurement and dropout process, respectively. Thus, the full data for the ith individual are given by Y_i and R_i, with distribution

$$f(y_i, r_i \mid X_i, \theta, \gamma) = f(y_i \mid X_i, \theta) f(r_i \mid y_i, X_i, \gamma), \tag{5.1}$$

where X_i is the design matrix of covariates for the ith individual. As mentioned earlier, the current chapter focuses only on missing data caused by dropout. This gives rise to a monotone missing data pattern, meaning that if Y_{ij} is missing, then $Y_{i(j+1)}$,...,Y_{in} are also missing. According to Rubin (1976) and Little and Rubin (1987), three general categories of dropout mechanisms can be distinguished. The taxonomy of Rubin (1976) is based on the second factor of (5.1), i.e., $f(r_i \mid y_i, X_i, \gamma)$. First, the dropout mechanism is defined as missing completely at random (MCAR) if the probability of non-response is independent of the response; that is, $f(r_i \mid y_i, X_i, \gamma) = f(r_i \mid X_i, \gamma)$. Second, missing at random (MAR) means that the probability of non-response is dependent on the set of observed values of the response; that is, $f(r_i \mid y_i, X_i, \gamma) = f(r_i \mid y_i^o, X_i, \gamma)$. Third, missing not at random (MNAR) means that the probability of non-response depends on the missing outcomes. However, an MNAR process is also allowed to depend on the observed outcomes; that is, $f(r_i \mid y_i, X_i, \gamma) = f(r_i \mid y_i^o, y_i^m, X_i, \gamma)$. In the context of a likelihood formulation, inference is based on

$$L(\theta, \gamma \mid X_i, y_i, r_i) \propto f(y_i^o, r_i \mid X_i, \theta, \gamma) = f(y_i^o, r_i \mid \theta, \gamma) = \int f(y_i, r_i \mid X_i, \theta, \gamma) dy_i^m. \tag{5.2}$$

Therefore,

$$f(y_i^o, r_i \mid \theta, \gamma) = \int f(y_i^o, y_i^m \mid X_i, \theta) f(r_i \mid y_i^o, y_i^m, X_i, \gamma) dy_i^m. \tag{5.3}$$

Under dropout MAR process, the likelihood contributions factor is:

$$f(y_i^o, r_i \mid \theta, \gamma) = \int f(y_i^o, y_i^m \mid X_i, \theta) f(r_i \mid y_i^o, X_i, \gamma) dy_i^m = f(y_i^o \mid X_i, \theta) f(r_i \mid y_i^o, X_i, \gamma). \tag{5.4}$$

The likelihood in (5.4) factorizes into two components of the same functional form as the general factorization of the full data (Y_i, R_i) given in (5.1). If, further, the parameters θ and γ are disjoint (i.e. orthogonal) which is to say the parameter space of the full vector vector $(\theta', \gamma')'$ is the product of the individual parameter spaces, the so-called separability condition, then inference can be based on the marginal observed data density only. Hence, when the separability condition is satisfied via a likelihood framework, ignorability is equivalent to MAR and MCAR. However, an MNAR mechanism is defined as a "non-ignorable" mechanism in the context of the likelihood framework. See, Little and Rubin (2002) for details on the derivation of the contribution to the likelihood attributable to the missingness mechanisms.

5.4 Dropout analysis strategies in non-Gaussian longitudinal data

There are a limited range of statistical methods for handling incomplete non-Gaussian longitudinal data, particularly when the missingness is not MCAR. The methods of analysis to deal with dropout comprise three broad strategies: semi-parametric regression, multiple imputation (MI) and maximum likelihood (ML). In what follows, we utilize three common statistical methods in practice, namely WGEE, MI-GEE and GLMM. First, we compare the performance of the two types of GEE approach, and then show how they compare to the likelihood-based GLMM approach.

5.4.1 Weighted generalized estimating equations (WGEE)

Next, we follow the description provided by Verbeke and Molenberghs (2005) in formulating the WGEE approach, thereby illustrating how WGEE can be incorporated into the conventional GEE method. Generally, if inferences are restricted to the population averages, the marginal expectations $E(Y_{ij}) = \mu_{ij}$ can be modelled with respect to covariates of interest. This can be done using the model $h(\mu_{ij}) = x'_{ij}\beta$, where $h(.)$ denotes a known link function, for example, the logit link for binary outcomes, the log link for counts, and so on. Further, the marginal variance depends on the marginal mean, with $Var(Y_{ij}) = v(\mu_{ij})\Omega$, where $v(.)$ and Ω denote a known variance function and a scale (overdispersion) parameter, respectively. The correlation between Y_{ij} and Y_{ik}, where $j \neq k$ for $i, j = 1, 2, ..., n_i$, can be given through a correlation matrix $C_i = C_i(\rho)$, where ρ denotes the vector of nuisance parameters. Then, the covariance matrix $V_i = (V_i, \rho)$ of Y_i can be decomposed into the form $\Omega A_i^{1/2} C_i A_i^{1/2}$, where A_i is a matrix with the marginal variances on the main diagonal and zeros elsewhere. Without missing data, the GEE estimator for β is based on solving the equation

$$S(\beta) = \sum_{i=1}^{N} \frac{\partial \mu_i}{\partial \beta'} (A_i^{1/2} C_i A_i^{1/2})^{-1} (y_i - \mu_i) = 0, \tag{5.5}$$

in which the marginal covariance matrix V_i contains a vector ρ of unknown parameters. For practical purposes, the vector ρ is replaced by a constant estimator. Now, assume that the marginal mean μ_i has been correctly modelled, then it can be shown that using Equation (5.5), the estimator $\hat{\beta}$ is asymptotically normally distributed with mean equal to β and covariance matrix equals

$$Var(\hat{\beta}) = I_0^{-1} I_1 I_0^{-1}, \tag{5.6}$$

where

$$I_0 = \left(\sum_{i=1}^{N} \frac{\partial \mu_i'}{\partial \beta} V_i^{-1} \frac{\partial \mu_i}{\partial \beta'} \right), \tag{5.7}$$

and

$$I_1 = \left(\sum_{i=1}^{N} \frac{\partial \mu_i'}{\partial \beta} V_i^{-1} Var(y_i) \frac{\partial \mu_i}{\partial \beta'} \right). \tag{5.8}$$

For practical purposes, in Equation (5.6), $Var(y_i)$ can be replaced by $(y_i - \mu_i)(y_i - \mu_i)'$ which is unbiased on the sole condition that the mean was again correctly specified (Birhanu et al., 2011). This leads to the so called empirical standard errors for inference purposes. Note that the GEE arises from non-likelihood inferences, therefore "ignorability" discussed above, cannot be invoked to establish the validity of the method when dropout under MAR holds (Liang and Zeger, 1986). Only, when the dropout are MCAR; that is, $f(r_i \mid y_i, X_i, \gamma) = f(r_i \mid X_i, \gamma)$, the estimating Equation (5.5) yields consistent estimators (Liang and Zeger, 1986). However, when the dropout depends on the outcome Y_i, which is the case for MAR, the classical GEE estimator is biased (Fitzmaurice et al., 1995; Liang and Zeger, 1986). Under MAR, Robins et al. (1995) proposed the WGEE approach to make the GEE possible to model data under the MAR missingness mechanism. The weights used in WGEE, also termed inverse probability weights, reflect the probability for an observation of subject to be observed (Robins et al., 1995). Therefore, the incorporation of these weights reduces possible bias in the regression parameter estimates. As illustrated by Molenberghs and Verbeke

(2005), we discus the idea of what these weights are. According to these authors, such a weight can be calculated as

$$\omega_{ij} \equiv P[D_i = j] = \prod_{k=2}^{j-1}(1 - P[R_{ik} = 0 \mid R_{i2} = ... = R_{i,k-1=1}]) \times \\ P[R_{ij} = 0 \mid R_{i2} = ... = R_{i,j-1} = 1]^{I\{j \leq n_i\}}, \quad (5.9)$$

where $j = 2, 3, ..., n_i + 1$, $I\{\}$ is an indicator variable, and D_i is the dropout variable. The weight is obtained from the inverse probability providing the actual set of measurements are observed. In terms of the dropout variable D_i, the weight probabilities are written as

$$\omega_{ij} = \begin{cases} \\ P(D_i = j \mid D_i \geq j) & \text{for } j\text{=}2 \\ P(D_i = j \mid D_i \geq j)\prod_{k=2}^{j-1}[1 - P(D_i = k \mid D_i \geq k)] & \text{for } j = 3, ..., n_i \\ \prod_{k=2}^{n_i}[1 - P(D_i = k \mid D_i \geq k)] & \text{for } j = n_i + 1. \end{cases} \quad (5.10)$$

Now, from Section 2 recall that we partitioned Y_i into the unobserved components (Y_i^m) and the observed components (Y_i^o). Similarly, the mean μ_i can be partitioned into observed (μ_i^o) and missing components (μ_i^m). In the WGEE approach, the score equations to be solved are:

$$S(\beta) = \sum_{i=1}^{N}\sum_{d=2}^{n_i+1} \frac{I(D_i = d)}{\omega_{id}} \frac{\partial \mu_i}{\partial \beta'}(d)(A_i^{1/2}C_iA_i^{1/2})^{-1}(d)(y_i(d) - \mu_i(d)) = 0, \quad (5.11)$$

where $y_i(d)$ and $\mu_i(d)$ are the first $d-1$ elements of y_i and μ_i respectively. In Equation (5.11), $\frac{\partial \mu_i}{\partial \beta'}(d)$ and $(A_i^{1/2}C_iA_i^{1/2})^{-1}(d)$ are defined analogously, in line with the definitions of Robins et al. (1995). Provided that the ω_{id} are correctly specified, the WGEE provides consistent estimates of the model parameters under a MAR mechanism (Robins et al., 1995).

5.4.2 Multiple imputation after GEE (MI-GEE)

MI is a simulation-based technique for filling in the missing values multiple times to construct multiple complete data sets. Details of this method can be found in Rubin

(1987), Schafer (1999) and Schafer and Graham (2002). Following is a brief description of MI and its application. According to Rubin (1987), MI consists of three steps:

Step 1: Each missing value is replaced by $M > 1$ simulated values.

Step 2: Each of the M complete data sets are analysed using standard statistical methods, such as logistic regression.

Step 3: The results from the M analyses have to be combined into a single inference by means of the method laid out in (Rubin, 1978).

The MI method requires the missingness mechanism to be MAR (Rubin, 1987; Molenberghs et al., 1997). The use of the number of the imputation (M) need not be very large since, in practice, 3 - 10 imputed data sets often provided satisfactory results. See, for example, Schafer (1997, 1999) and Schafer and Olsen (1998). Figure 5.1 shows an example of multiple imputation where $M = 5$.

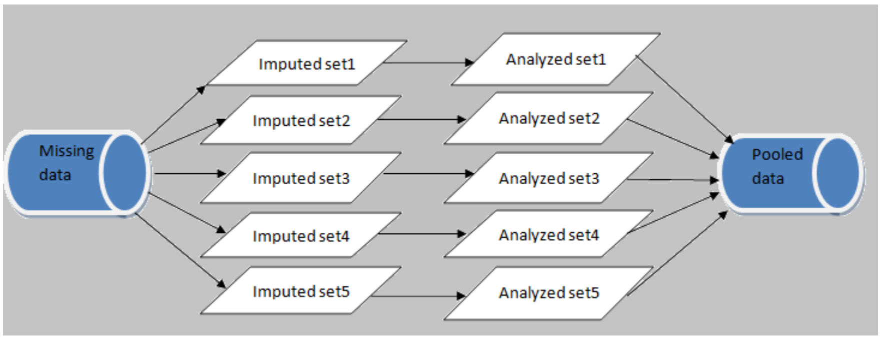


Figure 5.1: *Multiple imputation flow Chart: Imputation - Impute the missing entries 5 times, drawing from distribution. Analysis - Analyze each of the 5 imputed data sets using standard complete-data techniques. Pooling - Combine the 5 estimates into a final results, accounting for within-and-between-imputation variance*

We assume that the vector of repeated measurements Y_i is described by the parameter vector β. In the first imputation step, the objective is to impute the missing values with draws from the conditional distribution $f(y_i^m \mid y_i^o, \beta)$. Since β is unknown, an estimate for it denoted by $\hat{\beta}$, has to be obtained from the data, after which $f(y_i^m \mid y_i^o, \hat{\beta})$ is used to fill in the missing values, meaning that in the process, we generate draws from the distribution of $\hat{\beta}$, thus taking sampling uncertainty of estimating β into account. Alternatively, a Bayesian approach in which uncertainty about β is incorporated by means of using some prior distribution for β. After formulating the posterior distribution of β, the following imputation algorithm is used: A random β^* is first drawn from the posterior distribution of β, then a random Y_i^m is selected from $f(y_i^m \mid y_i^o, \beta^*)$. This posterior distribution is approximated by the normal distribution. The so-imputed missing values are next augmented to the observed data, yielding complete data, $Y = (Y_i^o, Y_i^m)$, which are then used to obtain $\hat{\beta}$ and its variance, $V = \hat{Var}(\hat{\beta})$. The steps mentioned above are independently repeated a number of times, say M times, yielding β^{*m} and V^m, for $m = 1, ..., M$. Finally, in the last step, the results of the analysis from the M completed (imputed) data are combined into a single inference. The overall estimated parameter for β and its estimated variance V are

$$\bar{\hat{\beta}} = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta^m}, \tag{5.12}$$

and

$$V = W + \left(\frac{M+1}{M}\right)B, \tag{5.13}$$

where

$$W = \sum_{m=1}^{M}\frac{V^m}{M}, \tag{5.14}$$

and

$$B = \sum_{m=1}^{M}\frac{(\hat{\beta}^m - \bar{\hat{\beta}})(\hat{\beta}^m - \bar{\hat{\beta}})'}{M-1}, \tag{5.15}$$

with W and B representing the average within-imputation variance and the between-imputation variance, respectively (Rubin, 1987). Since we consider the MI-GEE method, the M imputed data sets combined with GEE on the imputed data, is an alternative technique to likelihood inference and WGEE. It requires MAR for valid inferences. Note that the term $(1/M)B$ term in Equation (5.13) allows the overall variance to account for simulation error. Obviously, this error is minimal for large M. Thus the term is crucial for small M.

5.4.3 Generalized linear mixed model (GLMM)

When outcomes are of a non-Gaussian type, an alternative approach to deal with dropout under MAR is to use the likelihood-based inference (Verbeke and Molenberghs, 2000). A commonly encountered random effects (or subject-specific) model is the GLMM which is based on specifying a regression model for the responses conditional on an individual's random effects and assuming that within-subject measurements are independent, conditional on the random effects. The marginal likelihood in the GLMM is used as the basis for inferences for the fixed effects parameters, complemented with empirical Bayes estimation for the random effects (Molenberghs and Kenward, 2007). As pointed out by Alosh (2010), the random effects can be included as a subset of the model for heterogeneity from one individual to another. Integrating out the random effects induces marginal correlation between the responses through the same individual (Laird and Ware, 1982). Next, we briefly introduce a general framework for mixed effects models provided by Jansen et al. (2006b) and Molenberghs and Kenward (2007). It is assumed that the conditional distribution of each Y_i, given a vector of random effects b_i can be written as follows

$$Y_i \mid b_i \sim F_i(\theta, b_i), \tag{5.16}$$

where Y_i follows a prespecified distribution F_i, possibly depending on covariates, and is parameterized via a vector θ of unknown parameters common to all individuals. The term b_i denotes the $(q \times 1)$ vector of subject-specific parameters, called random effects, which are assumed to follow a so-called mixing distribution Q. The distribution Q depends on a vector of unknown parameter, say ψ; that is, $b_i \sim Q(\psi)$. In terms of the distribution of Y_i, the b_i reflect the between unit-heterogeneity in the population. Further, given the random effects b_i, it is assumed that the components Y_{ij} in Y_i are independent of one another. The distribution function (F_i) provided in model (5.16) becomes a product over the n_i independent elements in Y_i. Inference based on the marginal model for Y_i can be obtained by integrating out the random effects across their distribution $Q(\psi)$, provided one is not following a fully Bayesian approach. Now, assume that the $f_i(y_i \mid b_i)$ represents the density of the random effects function and corresponds to the distribution F_i, while $q(b_i)$ represents the density function and corresponds to the distribution Q. Thus, the marginal density function of Y_i can be written as follows

$$f_i(y_i) = \int f_i(y_i \mid b_i) q(b_i) db_i. \tag{5.17}$$

The marginal density is dependent on the unknown parameters θ and ψ. By assuming the independence of the units, the estimates of $\hat{\theta}$ and $\hat{\psi}$ can be obtained using the maximum likelihood function that is built into model (5.17). The inferences can be obtained following the classical maximum likelihood theory. The distribution Q is assumed to be of a specific parametric form, for example a multivariate normal distribution. An integration in model (5.17), depending on both F_i and Q_i, may or may not be analytically possible. However, there are some proposed solutions based on Taylor series expansions of either $f_i(y_i \mid b_i)$ or on numerical approximations of the integral, for example, adaptive Gaussian quadrature. Verbeke and Molenberghs (2000) noted that for the classical linear mixed model, $E(Y_i)$ equals $X_i\beta$, meaning that the fixed effects have a subject-specific as well as a population-averaged interpretation. However, for nonlinear mixed models, the interpretation of random effects has important ramifications for the interpretation of the fixed effects regression parameters. The fixed effects only reflect the conditional effect of covariates, and the marginal effect is difficult to obtain, as $E(Y_i)$ is given by

$$E(Y_i) = \int y_i \int f_i(y_i \mid b_i) q(b_i) db_i dy_i. \tag{5.18}$$

In the GLMM, a general formulation can be expressed as follows. It assumes that the elements Y_{ij} of Y_i are conditionally independent, given a $(q \times 1)$ vector of random effects b_i, with density function based on a classical exponential family formulation with conditional mean depending on both fixed and random effects. This leads to the conditional mean $E(Y_{ij} \mid b_i) = a'(\eta_{ij}) = \mu_{ij}(b_i)$, and the conditional variance is assumed to depend on the conditional mean according to $V = (Y_{ij} \mid b_i) = \Theta a''(\eta_{ij})$. One needs a link function, say h (for binary data, a canonical link is the logit link), and typically uses a linear regression with parameters β and b_i for the mean, i.e., $h(\mu_i(b_i)) = X_i\beta + Z_i b_i$. Here, we note that the linear mixed model is a special case with an identity link function. The random effects b_i are again assumed to be sampled from a multivariate normal distribution, with mean 0 and $(q \times q)$ covariance matrix, Cov. The canonical link function is usually used to relate the conditional mean of Y_{ij} to η_i; that is, $h = a'^{-1}$, such that $\eta_i = X_i\beta + Z_i b_i$. In principle, any suitable link function can be used (Fitzmaurice et al., 2004). In considering the link function of the logit form and assuming the random effects to be normally distributed, the familiar logistic-linear GLMM follows. For a more detailed overview, see, Jansen et al. (2006b) and Molenberghs and Verbeke (2005).

5.5 Simulation study

5.5.1 Design

The main objective of this study was to compare WGEE, MI-GEE and GLMM for handling dropout missing at random in non-Gaussian longitudinal data. To do so, we used the following steps:

Step 1: An empirical data set was generated in the form of non-Gaussian longitudinal binary data. A marginal logistic regression was fitted to the complete data, thus producing true regression coefficients.

Step 2: Once the complete data sets were generated, 1000 random samples of N=250 and 500 subjects were drawn.

Step 3: The missing at random data were generated for various dropout rates.

Step 4: The above methods were applied to each simulated data set. The results from the simulated data were then compared with those obtained from the complete data.

Step 5: The performances of WGEE, MI-GEE and GLMM were evaluated in terms of bias, efficiency and mean square error (MSE). The GLMM estimates were first adjusted for comparability before this evaluation of performance.

5.5.2 Data generation

Simulated data were generated in order to emulate data typically found in longitudinal non- Gaussian clinical trials data. The non-Gaussian longitudinal data with dropout were simulated by first generating complete data sets. Then, 1000 random samples of sizes N =250 and 500 subjects were drawn. We assumed that subjects were assigned to two arms (Treatment=1 and Placebo=0). We also assumed that measurements were taken under four time points ($j = 1; 2; 3; 4$). The outcome (Y_ij) which is the measurement of subject i, measured at time j, was defined as 1 if the measurement is positive, and 0 if otherwise. The two levels of the outcome represent anything, but generically we labeled one outcome "success", i.e., "1" and the other "failure", i.e., "0". Then, we looked at logistic regression as modeling the success probability as a function of the explanatory variables. The main interest here is in the marginal model for each binary outcome Y_ij, which we assumed follows a logistic regression. Consequently, longitudinal binary data were generated according to the following random effects model with linear predictor

$$logitE(y_{ij=1|T_j,trt_i,b_i}) = \beta_0 + b_i + \beta_1 T_j + \beta_2 trt_i + \beta_3(T_j * trt_i), \tag{5.19}$$

where $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$, and the random effects b_i's are assumed to account for the variability between individuals and assumed to be $i.i.d$ with a normal distribution, i.e., $b_i \sim N(0, \sigma^2)$. In this model, fixed categorical effects include treatment (trt), times (T) and treatment-by-time interaction ($T * trt$). For this model, throughout, we fixed $\beta_0 = -0.25$, $\beta_1 = 0.5$, $\beta_2 = 1.0$ and $\beta_4 = 0.2$. We also set a random intercept $b_i \sim N(0, 0.07)$. For each simulated data set, after the complete longitudinal binary data had been generated, dropouts were created by deleting different rates of the responses chosen stochastically (conditional on time), namely the dropouts were imposed on Y_{ij}. To achieve MAR mechanism, dropouts were generated from the full (i.e., no missing) simulated data, using the following strategy. We assumed that the dropout can occur only after the second and third time points. Consequently, there are three possible dropout patterns. That is, dropout at the third time point, dropout at the fourth time point, or no dropout. The dropouts were generated at time j and the subsequent times were assumed to be dependent on the values of outcome measured at time $j-1$. Under model (5.19), we simulated a case where the MAR specification was different for the two treatments. In particular, for time point, j=3, we retained the criterion that if the dependent variable (Y_{ij}) was positive (i.e., $Y_{ij} = 1$), then the subject dropped out at the next time point, i.e., $j+1$. Dropouts were selected to yield approximate rates of 10%, 20% and 30%. A monotone missing pattern (i.e., data for an individual up to a certain time) was considered, thus simulating a trial where the only source of dropout was an individual's withdrawal.

5.5.3 Analysis

In the analysis, different strategies were used to handle dropout: by weighting, by imputation and analyzing the data as they are (i.e., without the need to impute or weight), consistent with MAR assumption, for WGEE, MI-GEE and GLMM, respectively.

5.5.3.1 WGEE

As discussed above, the WGEE method requires a model for the dropout mechanism. Consequently, we first fitted the following dropout model using a logistic regression,

$$logitP(D_i = j \mid D_i \geq j) = \gamma_0 + \gamma_1 y_{i,j-1} + \gamma_2 trt_i, \tag{5.20}$$

where the predictor variables were the outcomes at previous occasions ($y_{i,j-1}$), supplemented with genuine covariate information. Model (5.20) is based on logistic regression for the probability of dropout at occasion j for individual i, conditional on the individual still being in the study. Note that mechanism (5.20) allows for the one used to generate the data and described in Section 5.5 only as a limiting case. This is because our dropout generating mechanism has a deterministic flavor. Strictly speaking, the probabilities of observation in WGEE are required to be bounded away from zero, to avoid issues with the weight. The effect of our choice is that WGEE is subjected to a severe stress test. It will be seen in the results section that, against this background, WGEE performs rather well. To estimate the probabilities for dropout as well as to pass the weights (predicted probabilities) to be used for WGEE, we used the "DROPOUT" and "DROPWGT" macros described in Molenberghs and Verbeke (2005). Note that no modification needed to apply these macros. The "DROPOUT" macro is used to construct the variables *dropout* and *previous*. The outcome *dropout* is binary and indicates if individual had dropped out of the study before its completion, whereas, the *previous* variable refers to the outcome at previous occasions. After fitting a logistic regression, the "DROPWGT" macro is used to pass the weights to the individual observations in the WGEE. Such weights, calculated as the inverse of the cumulative product of conditional probabilities, can be estimated as $w_{ij} = 1/(\lambda_{i1} \times ... \times \lambda_{ij})$, where

λ_{ij} represents the probability of observing a response at time j for the ith individual, conditional on the individual being observed at the time $j-1$. Once the dropout model (5.20) was fitted and the weight distribution was checked, we included the weights by means of the WEIGHT statement in SAS procedure GENMOD. As mentioned earlier, the marginal measurement model for WGEE should be specified. Therefore, the model that we considered takes the form of

$$logitE(y_{ij}) = \beta_0 + \beta_1 T_j + \beta_2 trt_i + \beta_3 (T_j * trt_i). \tag{5.21}$$

Here, we used the compound symmetry (CS) working correlation matrix. A random intercept b_i was excluded when considering WGEE.

5.5.3.2 MI-GEE

The analysis was conducted by imputing missing values using the SAS procedure MI, which employs a conditional logistic imputation model for binary outcomes. For the specification of the imputation model, an MAR mechanism is considered; that is, the imputation model comprises dichotomous covariates as well as longitudinal binary outcomes values at times $j = 1, 2, 3, 4$. To be precise, for the multiple imputation as well as for the MAR imputation model, we used a logistic regression with measurements at the second time point as well as the two key covariates to fill in the missing values occur at the third time point. In a similar way, the imputation at the fourth time point is done using the measurements at the third time point including both imputed and observed, as predictors, as well as the measurements at the second time point which is always observed and the two key covariates. Note that we describe here multiple imputation in a sequential fashion, making use of the time ordering of the measurements. Therefore, the next value is imputed based on the previous values, whether observed or already imputed. This is totally equivalent to an approach where all missing values are imputed at once based on the observed sub-vector. This implies that the dropout process was accommodated in the imputation model. It appears that there is potential for misspecification here. However, multiple imputation is valid under MAR. Whether missingness depends on one or more earlier outcomes, MAR holds, so the validity of the method is guaranteed. In terms of the number of the imputed data sets, we used $M = 5$ imputations. This is often sufficient to obtain satisfactory results (Rubin, 1987; Schafer, 1999; Verbeke and Molenberghs, 2007). The GEE was then fitted to each completed data set using SAS procedure GENMOD to estimate the overall parameters and their variances. The GEE model that we considered is based on (5.21). The results of the analysis from these 5 completed (imputed) data sets were combined into a single inference using Equations (5.12), (5.13), (5.14) and (5.15). This was done by using SAS procedure MIANALYZE. Details of implementation of this method are given in Molenberghs and Kenward (2007) and also in Beunckens et al. (2008).

5.5.3.3 GLMM

Conditionally on a random intercept b_i, the logistic regression model is used to describe the mean response, i.e., the distribution of the outcome at each time point separately. Specifically, we considered fitting model (5.19). This model assumed that there is natural heterogeneity across individuals and accounted for the within-subject dependence in the mean response over time. Model (5.19) was fitted using the likelihood method by applying the macro NLMIXED in SAS software. This procedure relies on numerical integration and includes a number of optimization algorithms (Molenberghs and Verbeke, 2005). Given that the evaluation and maximization of the marginal likelihood for GLMM needs integration, over the distribution of the random effects, the model was fitted using maximum likelihood (ML) together with adaptive Gaussian quadrature (Pinheiro and Bates, 2000) based on numerical integration which works quite well in procedure NLMIXED. The procedure is valid as long as the dropouts are MAR (Stubbendick and Ibrahim, 2006). The NLMIXED procedure allows the use of the Newton-Raphson instead of Quasi-Newton algorithm to maximize the marginal likelihood, and adaptive Gaussian quadrature was used to integrate out the random effects. The adaptive Gaussian quadrature approach makes Bayesian approaches quite appealing because it is based on numerical integral approximations centered around the empirical Bayes estimates of the random effects, and permits maximization of the marginal likelihood with any desired degree of accuracy (Anderson and Aitkin, 1985).

An alternative strategy to fit mixed models is the penalized quasi-likelihood (PQL) algorithm (Stiratelli et al., 1984). However, in this study this algorithm is not used as it often provides highly biased estimates for the longitudinal binary responses (Breslow and Lin, 1995). Also, we ought to keep in mind that the GLMM parameters need to be re-scaled in order to have an approximate marginal interpretation and to become comparable to their GEE counterparts.

5.5.4 Evaluation criteria

In the evaluation, the inferences are made on the data before dropouts are created and the results used as the main standard against those obtained from applying WGEE, GLMM and MI-GEE approaches. We evaluated the performance of the methods being studied largely using criteria which included the evaluations of bias, efficiency and mean square error (MSE). These criteria are recommended in Collins et al. (2001) and Burton et al. (2006). (1) Evaluation of bias: we defined the bias as the difference between the average estimate and the true value; that is, $\pi = (\bar{\hat{\beta}} - \beta)$, where β is the true value for the estimate of interest, $\bar{\hat{\beta}} = \Sigma_{i=1}^{S} \hat{\beta}_i / S$, S is the number of simulations performed and $\hat{\beta}_i$ is the estimate of interest within each of the $i = 1, ..., S$ simulations. (2) Evaluation of efficiency: we defined the efficiency as the variability of the estimates around the true population coefficient. In the current chapter, it was calculated by the average width of the 95% confidence interval. The 95% interval is approximately four times the magnitude of the standard error. Therefore, a narrower interval is always desirable because it leads to more efficient methods. (3) Evaluation of accuracy: the MSE provides a useful measure of the overall accuracy, as it incorporates both measures of bias and variability (Collins et al., 2001). It can be calculated as follows: MSE=$(\bar{\hat{\beta}} - \beta)^2$+$(SE(\hat{\beta}))^2$, where $SE(\hat{\beta})$ denotes the empirical standard error of the estimate of interest over all simulations (Burton et al., 2006). Generally, small values of MSE are desirable (Schafer and Graham, 2002).

5.6 Results

A few points about the parameter estimates obtained by the proposed methods through the three evaluation criteria may be noted for each estimate in Tables 5.1 and 5.2. First, the greatest bias, also the worst, are highlighted. Second, for the efficiency criterion, the widest confidence interval, also the worst, 95% interval are highlighted. Third, for the evaluation of MSEs, the greatest values, also the worst, are highlighted.

5.6.1 Sample size, N=250

The results of WGEE, MI-GEE and GLMM in terms of bias, efficiency and MSEs, under N=250 sample size are presented in Table 5.1. By looking at this table, we observed that for 10% dropout rate, bias was least in the estimates of MI-GEE than in both WGEE and GLMM. In particular, the worst performance of WGEE and GLMM on bias permeated through the estimates of β_2 and $(\beta_0, \beta_1, \beta_3)$, respectively, indicating a discrepancy between the average and the true parameter (Schafer and Graham, 2002). Between the two MI-GEE and WGEE methods, the WGEE estimates were slightly different from those obtained by MI-GEE, although the degree of these differences was not very large. The efficiency performance was acceptable for both methods and comparable to each other, but low for most parameters under WGEE. The efficiency estimates associated with GLMM were larger than with WGEE and MI-GEE. In terms of MSEs, both WGEE and MI-GEE outperformed GLMM as they tend to have smallest MSEs. In addition, they yielded MSEs much closer to each other, but in most cases, were smallest for WGEE.

Considering the 20% dropout rate, the results shown in Table 5.1 revealed that in nearly all cases, GLMM consistently produced the most biased estimates (the only exception to this rule occurred for estimate of β_2). However, some bias was also evident under WGEE in estimating the parameter β_2. For estimating all parameters, efficiency estimates by WGEE and MI-GEE were similar to each other and smaller than GLMM's estimates. The WGEE outperformed MI-GEE and GLMM on the MSE criterion. Comparing both WGEE and MI-GEE, the MSEs associated with both methods were closer to each other and in one case - MSE of β_3 - they gave the same values. The lone exception to these MSEs occurred for MI-GEE regarding the estimate of β_0. In comparison with WGEE and MI-GEE, GLMM gave larger MSEs in magnitude than the two, except for estimate of β_0.

For the 30% dropout rate, Table 5.1 showed that the results based on GLMM typically displayed greater estimation bias than did WGEE and MI-GEE, indicating a difference between the average estimate and the true values. Compared to those of WGEE, the MI-GEE estimates were less substantially biased, but was somewhat more biased for β_3. For the effciency criterion, the MI- GEE based results were smaller than those from WGEE. In general, both WGEE and MI-GEE yielded estimates which did not differ too much to each other, with a preference for MI-GEE in all cases. Particularly, MI-GEE was more efficient than WGEE, yet more efficient than GLMM. The MSEs by both WGEE and MI-GEE for all cases were smaller than those from GLMM, as the latter yielded the largest values in most cases. MI-GEE tends to have the smallest MSEs. In general, across various dropout rates and when increasing dropout rate, the performance of MI-GEE was better than that for WGEE and GLMM. However, the method was less efficient than did WGEE with respect to efficiency criterion, yet more efficient and accurate than GLMM.

5.6.2 Sample size, N=500

Table 5.2 displays the results of bias, efficiency and MSE when the sample size was 500. For 10% dropout, when compared with the results based on the WGEE and GLMM, treating the data with MI-GEE appeared to have resulted in fairly less bias. However, GLMM notably produced the most biased estimates. The exception for GLMM estimates occurred for β_2 which was more biased under WGEE. An examination of efficiency revealed that MI-GEE yielded more efficient estimates. As was the case with N=250, most GLMM estimates were less efficient, compared to those obtained by WGEE and MI-GEE. A comparison of efficiency between WGEE and MI-GEE showed no systematic differences in their estimates, although the performance of MI-GEE was better. Moreover, in all cases, GLMM led to large MSEs than those from WGEE and MI-GEE. Overall, the MSEs were smaller for MI-GEE.

Table 5.1: *Bias, Efficiency and Mean square error of the WGEE, MI-GEE and GLMM Methods, under MAR mechanism over 1,000 samples: N=250 subjects*

dropout rate	Parameter	Bias			Efficiency			MSE		
		WGEE	MI-GEE	GLMM	WGEE	MI-GEE	GLMM	WGEE	MI-GEE	GLMM
10%	β_0	0.094	0.061	**0.099**	0.005	0.012	**0.018**	0.028	0.018	**0.041**
	β_1	-0.099	-0.030	**-0.107**	0.003	0.013	**0.084**	0.018	0.086	**0.097**
	β_2	**0.053**	0.039	0.050	0.004	0.004	**0.011**	0.051	0.093	**0.107**
	β_3	0.018	0.012	**0.023**	0.002	0.004	**0.005**	0.007	0.008	**0.015**
20%	β_0	0.047	0.026	**0.052**	0.012	0.012	**0.031**	0.027	**0.060**	0.032
	β_1	0.033	0.048	**0.141**	0.012	0.014	**0.028**	0.048	0.020	**0.052**
	β_2	**0.131**	0.122	0.130	0.005	0.011	**0.017**	0.051	0.091	**0.102**
	β_3	-0.076	-0.038	**-0.080**	0.006	0.007	**0.009**	0.008	0.008	**0.016**
30%	β_0	-0.065	-0.036	**-0.085**	0.026	0.003	**0.041**	0.071	0.072	**0.087**
	β_1	0.167	0.143	**0.169**	**0.023**	0.011	0.013	**0.089**	0.035	0.044
	β_2	0.178	0.171	**0.182**	0.015	0.005	**0.019**	0.069	0.032	**0.073**
	β_3	0.033	**0.104**	0.079	0.013	0.005	**0.016**	0.025	0.014	**0.047**

Note: The largest bias, efficiency and mean square error for each given estimate presented in bold. MI-GEE=multiple imputation after generalized estimating equations; WGEE=wieghted generalized estimating equations; LMM=linear mixed model; GLMM=generalized linear mixed model; MSE=mean square error.

With respect to 20% dropout, the results revealed the same findings as in the case of 10% dropout. Across all cases, MI-GEE outperformed both WGEE and GLMM in terms of bias criterion, except for β_2. The greatest bias for the WGEE and GLMM methods occurred for the estimates of β_3 and (β_0, β_1), respectively. The efficiency estimates associated with MI-GEE were typically smaller than those associated with the WGEE and GLMM methods. Thus, the MI-GEE method was more efficient. In case of the MSEs, the results suggested that GLMM's values were larger than were those for MI-GEE and WGEE. Its MSEs tended to be worse, with a larger MSE for WGEE in terms of the value of β_2.

A comparison of 30% dropout rate again suggested that the estimates associated with MI-GEE were less biased than for WGEE and GLMM. Specifically, estimates which showed most bias were (β_0, β_3) and (β_1, β_2) for WGEE and GLMM, respectively. Efficiency by MI-GEE appeared to be independent of the dropout rate, meaning the MI-GEE method yielded more efficient estimates across all dropout rates. Comparing the efficiency results, WGEE resulted in smaller estimates than estimates of GLMM, with one exception: the estimate of β_3. With respect to MSE, results that are computed by GLMM yielded largest values, showing no substantial improvement over GLMM under different dropout rates when compared with the results computed by WGEE and MI-GEE . It can also be observed that, in terms of the estimate of β_3, the MSE value for WGEE was equal to that based on GLMM, and they gave larger MSEs than did MI-GEE, whereas compared to WGEE, the MI-GEE still resulted in smaller MSEs.

5.7 Discussion and conclusion

In this chapter, we investigated the performance of different families of approaches for handling dropout that are MAR in non-Gaussian longitudinal outcomes. Our focus was on binary outcomes. Similar results will likely apply to other data types as well but should, ideally, be the subject of additional research. The models considered include WGEE, MI-GEE and GLMM representing different strategies to deal with dropout under MAR. In the analysis, different ways were used to handle dropout: by weighting, by analyzing the data as they are (i.e., without need to weight or impute),

Table 5.2: *Bias, Efficiency and Mean square error of the WGEE, MI-GEE and GLMM Methods, under MAR mechanism over 1,000 samples: N=500 subjects*

Dropout rate	Parameter	Bias			Efficiency			MSE		
		WGEE	MI-GEE	GLMM	WGEE	MI-GEE	GLMM	WGEE	MI-GEE	GLMM
10%	β_0	0.043	0.011	**0.051**	0.156	0.144	**0.162**	0.019	0.016	**0.059**
	β_1	-0.149	-0.142	**-0.179**	0.057	0.054	**0.068**	0.048	0.044	**0.053**
	β_2	**0.221**	0.211	0.220	0.093	0.086	**0.129**	0.097	0.082	**0.101**
	β_3	0.047	0.010	**0.056**	**0.036**	0.032	0.034	0.009	0.009	**0.017**
20%	β_0	0.080	0.078	**0.091**	0.154	0.138	**0.161**	0.130	0.141	**0.145**
	β_1	-0.195	-0.139	**-0.201**	0.068	0.053	**0.073**	0.052	0.037	**0.082**
	β_2	0.265	**0.293**	0.289	0.099	0.089	**0.153**	**0.120**	0.118	0.119
	β_3	**0.067**	0.020	0.064	**0.041**	0.032	0.034	0.009	0.007	**0.014**
30%	β_0	**0.136**	0.117	0.121	0.131	0.164	**0.173**	0.139	0.193	**0.198**
	β_1	-0.232	-0.218	**-0.243**	0.072	0.048	**0.074**	0.066	0.066	**0.091**
	β_2	0.342	0.184	**0.351**	0.084	0.093	**0.107**	0.186	0.136	**0.193**
	β_3	**0.067**	0.066	0.064	**0.097**	0.029	0.068	**0.012**	0.010	**0.012**

Note: The largest bias, efficiency and mean square error for each given estimate presented in bold. MI-GEE=multiple imputation after generalized estimating equations; WGEE=wieghted generalized estimating equations; LMM=linear mixed model; GLMM=generalized linear mixed model; MSE=mean square error.

consistent with the MAR assumption and by imputation, for WGEE, GLMM and MI-GEE, respectively. We compared the MI-GEE method under an imputation model based on regression of the dropout measurement on previous observed measurement with the WGEE method for the correctly specified dropout probability model, in the sense that both the dropout and the measurement models are correctly specified. The methods were compared on simulated data in the form of binary longitudinal clinical trial data. The correlated binary variables were generated from a random effects model. The dropouts missing at random were generated under several different dropout rates as well as samples sizes. The comparisons were made through the evaluation of bias, efficiency and mean square error. Based on the results of the comparative analysis, we reached the following conclusions.

The findings in general favoured MI-GEE over both WGEE and GLMM. This MI-GEE advantage is well documented in Birhanu et al. (2011) and Molenberghs and Kenward (2007). Furthermore, the bias for MI-GEE based estimates in this study was fairly small, demonstrating that the imputed values did not produce markedly more biased results. This was to be expected as many authors, for example, Beunckens et al. (2008) noted that the MI-GEE method may provide less biased estimates than a WGEE analysis when the imputation model is correctly specified. From an extensive small and high sample sizes (i.e., N=250 and 500) simulation study, it emerged that MI-GEE is rather efficient and more accurate than other methods investigated in the current chapter, regardless of dropout rate which also shows that the method does well as the dropout rate increases. Overall, the MI-GEE performance appeared to be independent of the sample sizes. However, in terms of efficiency, in some cases, it was less efficient than WGEE, yet more efficient and accurate than GLMM. This was specially true for WGEE when the rate of dropout was small and the sample size was small as well. In summary, the results further recommended MI-GEE over WGEE. However, both MI-GEE and WGEE methods may be selected as the primary analysis methods for handling dropout under MAR in non-Gaussian longitudinal outcomes, but convergence of the analysis models may be affected by the discreteness or sparseness of the data.

Molenberghs and Verbeke (2005) stated that the parameter estimates from the

GLMM are not directly comparable to the marginal parameter estimates, even when the random effects models are estimated through marginal inference. They also suggested that the GLMM parameter estimates can be approximately transformed to their corresponding GEE, using a ratio that makes the parameter estimates comparable. Therefore, an appropriate adjustments need to be applied to GLMM estimates in order to have an approximate marginal interpretation and to become comparable to their GEE counterparts. Using this ratio in the simulation study, the findings showed that, although all WGEE, MI-GEE and GLMM are valid under MAR, there were slight differences between the parameter estimates and never differed by a large amount, in most cases. As a result, it appeared that for both sample sizes, the GLMM based results were characterized by the larger estimates for nearly all cases, although the degree of the difference in magnitude was not very large. In addition, it did not appear that the magnitude of this difference differed between the three dropout rates.

Although there was a discrepancy between the GLMM results on the one hand, and both the WGEE and MI-GEE results on other hand, there are several important points to consider in the GLMM analysis of incomplete non-Gaussian longitudinal data. The fact is that the GLMM may be applicable in many situations and offers an alternative to the models that make inferences about the overall study population when one is interested in making inferences about individual variability to be included in the model (Verbeke and Molenberghs, 2000; Molenberghs and Verbeke, 2005). Furthermore, it is important to realize that GLMM relies on the assumption that the data are MAR, provided a few mild regularity conditions hold, and it is as easy to implement and represent as it would be in contexts where the data are complete. Consequently, when this condition holds, valid inference can be obtained with no need for extra complication or effort, and the GLMM assuming an MAR process, is more suitable (Molenberghs and Verbeke, 2007). In addition, the GLMM is very general and can be applied for various types of discrete outcomes when the objective is to make inferences about individuals rather than population averages, and is more appropriate for explicative studies.

As a final remark, recall that MI-GEE has been the preferred method for analysis as it outperformed both the WGEE and GLMM estimations in the simulation study results. Despite this, the current study has focussed on handling dropout in the outcome variable, and the MI-GEE can be well conducted in terms of the missingness in the covariates in the context of real-life, and can yield even more precise and convincing results since the choice for the WGEE method can be ruled out. This can be justified by the fact that, in the imputation model, the covariates that are conditioned on the analysis model are not included. The other available covariates can be included in the imputation model without being of interest in the analysis model, therefore yielding better imputations as well as wider applicability. Additionally, multiple imputation methods such as MI-GEE avoid some severe drawbacks encountered using direct modelling methods such as the excessive impact of the individual weights in the WGEE estimation or potential poor fit of the random subject effect in the GLMM analysis. For further discussion, see, Beunckens et al. (2008).

In conclusion, we submit that the scope of this chapter is limited to three approaches. This work is not intended to provide a comprehensive account of analysis methods for incomplete non-Gaussian longitudinal outcome. We acknowledge that there are several methods available for incomplete non-Gaussian longitudinal outcomes under the dropouts that are MAR. However, these methods are beyond the scope of the study. This article exclusively deals with the WGEE, MI-GEE and GLMM paradigms that represent different strategies to deal with dropout under MAR.

Chapter 6

An analysis of incomplete longitudinal data with application to multi-centre trial data: A selection model for non-ignorable missingness[*]

6.1 Abstract

The present chapter deals with the analysis of longitudinal continuous measurements with incomplete data due to non-ignorable dropout. In repeated measurements data, as one solution to handle non-ignorable dropout, the selection model assumes a mechanism of outcome-dependent dropout and jointly both the measurement together with dropout process of repeated measures. We consider the construction of a particular type of selection model that uses a logistic regression model to describe the dependency of dropout indicators on the longitudinal measurement. We focus on the use of

[*] A. Satty, H. Mwambi and M. Kenward (In review). An analysis of incomplete longitudinal data with application to multi-centre trial data: A selection model for non-ignorable missingness. *South African Statistical Journal.*

the Diggle-Kenward's (1994) model as a tool for assessing the sensitivity of a selection model in terms of the modelling assumptions. Our main objective here is to investigate the influence on inference that might be exerted on the considered data by the dropout process. We restrict attention to a model for repeated Gaussian measures, subject to potentially non-random dropout. To investigate this, we analyze incomplete longitudinal clinical trial with dropout using a multi-centre clinical trial data.

Keywords: Gaussian longitudinal data, Selection models, Diggle and Kenward model, non-ignorable missingness, Missing not at random (MNAR).

6.2 Introduction

A typical characteristic of longitudinal studies is that study subjects are measured repeatedly over time. The dropout of subjects along the time scale is common. The dropout process is assumed to be stochastic in nature and dependent upon observed and unobserved outcomes. It also may depend upon covariates, such as the treatment arm an individual is allocated to. The dropout may be regarded as a "failure" outcome in certain limited settings. Of prime concern to this study, is the more general situation that characterizes the statistical behavior of the original outcome, while dropout is treated as a "nuisance" occurrence that must be tolerated. As a result of this, the distinction between the outcome and the dropout processes needs to be simultaneously maintained.

Rubin (1976) and Little and Rubin (1987) introduce different mechanisms for denoting dropout or non-response. A dropout, or non-response process is said to be missing completely at random (MCAR) if the non-response process is a random event independent of both unobserved and observed outcomes, missing at random (MAR) if, conditional upon the observed outcomes, the non-response process is independent of the unobserved outcomes and missing not at random (MNAR) when the non-response process depends only upon the unobserved outcomes. In the context of likelihood and Bayesian inferences, and when the parameters describing the measurement process are functionally independent of the those describing the non-response process, MCAR and

MAR are ignorable, while a non-random process is non-ignorable (Rubin 1976; Little and Rubin, 1987).

It is possible to consider more general models when one assumes random missingness mechanism to be untrue (Jansen et al., 2006a). Examples on work of MNAR modelling include Diggle and Kenward (1994), Molenberghs et al. (1997), Jansen et al. (2003) and Verbeke e al. (2001). These belong to the so-called selection models family (Heckman, 1976; Little and Rubin, 1987). A selection model factors the joint distribution of the measurement and dropout mechanism into two parts; that is, a marginal measurement model that describes the distribution of the underlying complete data, and a dropout mechanism that describes the distribution of the missing data indicators, conditional upon the complete data. For more details, see, for example, Diggle and Kenward (1994) and Verbeke and Molenberghs (2000). This is intuitively appealing since the marginal measurement distribution would be of interest also with complete data (Molenberghs and Kenward, 2007). Furthermore, the missing data mechanisms (MCAR, MAR and MNAR) are most easily developed within the selection setting. However, it is often argued, especially within the context of non-random missingness model, that selection models, although identifiable, should be approached with caution (Glynn et al., 1986). Indeed, one has to make untestable assumptions about the missing data process. Selection models originated from the Tobit model of Heckman (1976) in econometrics to correct for selection bias. The theoretical translation from the Heckman's (1976) model to Diggle and Kenward's (1994) selection model have been addressed by Verbeke and Molenberghs (2000). Diggle and Kenward (1994) consider a selection model for the study of longitudinal measurements when data are MNAR by letting the probability of dropout depend on unobserved measurements. They use a linear mixed model for the longitudinal measurement and logistic regression model for the dropout process to describe the dependency between dropout indicators and measurements. The dropout indicators are used to indicate participant dropout. However, the intermittent missing data is assumed to be missed at random, and it can be ignored in the model. For alternatives for the missing data processes, see, Molenberghs and Kenward (2007).

Earlier work on the selection model analysis is given by Heckman (1976) and Glynn et al. (1986). Selection models that are applied to the regression analysis of categorical variables with outcomes subject to non-ignorable non-response are applied by Baker and Laird (1988), while Robins et al. (1994) used a selection perspective for the conditional expectation model in a semi-parametric approach. For the ignorable non-response hypothesis, Robins and Gill (1997) proposed a general class of selection models under non-monotone missing data pattern. In the case of the selection models for repeated measurements, sensitivity of the conclusions to the assumptions about the dropout mechanism has been illustrated by Kenward (1998). A semi-parametric approach of missing data mechanism is proposed by Scharfstein et al. (1999) in order to avoid the impact of the parametric missing data specification in a selection model perspective. With regard to the non-monotone pattern, selection models have been extended by Troxel et al. (1998). In addition to Troxel's work, within the selection model framework, models have been proposed for non-monotone pattern as well, for instance, see, Jansen and Molenberghs (2008). In the context of categorical data, selection models have been developed by Fitzmaurice et al. (1995) and Nordheim, E.V. (1984). Additionally, a number of proposals have been made for non-Gaussian outcomes (Molenberghs and Verbeke, 2005). Further details in selection models can be found in Robins et al. (1995), Rotnitzky and Robins (1997), Robins et al. (1998), Verbeke and Molenberghs (2000), Molenberghs and Verbeke (2005) and Molenberghs and Kenward (2007).

This chapter deals with the analysis of longitudinal data when there are non-ignorable dropout. We illustrate this analysis by considering the problem of missing data that occurs with a continuous outcome. We focus on the use of the Diggle and Kenward's (1994) model as a tool for assessing the sensitivity of a selection to the modelling assumptions. We restrict attention to a model for repeated Gaussian measures, subject to dropout possibly depending upon missing outcomes, i.e., MNAR. A monotone missing pattern has been constructed in the model. Similar to Diggle and Kenward (1994), a selection model is specified that uses a logistic regression model to describe the dependency of missing data indicators upon the longitudinal response. In the current application, we modify the analysis software to accommodate the case of more than two treatment arms as a computational extension. Our main objective here is to investigate the influence that might be exerted on the considered data by the dropout process. In order to investigate our objective, we analyze incomplete longitudinal data with dropout. We outline the fitting of the selection model which is based on the linear mixed model for the measurement process as well as a logistic regression for dropout process. The model will be fitted using standard statistical software (SAS version 9.2, IML macro). This is done by using a practical example in the form of a multi-centre clinical trial data.

The remainder of the chapter is organized as follows: the data setting and modelling framework are introduced in Section 6.3. In Section 6.4, a background for the selection model is provided, followed by descriptions of the selection model based on Diggle and Kenward model frameworks as well as detailed discussion of the linear mixed model and dropout model. In Section 6.5, we present an application including a description of the data set used in the analysis. The results of the estimation of the model are then described in Section 6.6. We conclude with a discussion of the results in Section 6.7.

6.3 Modelling longitudinal data with dropout

To introduce some necessary notation, we follow the terminology provided by Verbeke and Molenberghs (2000) and Molenberghs and Kenward (2007) based on the standard modelling frameworks of Rubin (1976) and Little and Rubin (2002). So, assume that for each independent subject $i = 1, ..., N$ in the study a sequence of responses Y_{ij} is designed to be measured at a fixed set of occasions $j = 1, ..., n$. The outcomes are grouped into a vector $Y_i = (Y_{i1}, ..., Y_{in})'$. It is often necessary to split the outcome vector Y_i into two sub-vectors, Y_i^o and Y_i^m, indicating the observed and missing components, respectively. Additionally, one can define an indicator R_{ij} for each occasion j as follows: R_{ij}=1, if Y_{ij} is observed, and R_{ij}=0, if not. The indicators of missing data (R_{ij}) can be grouped into a vector R_i which is of parallel structure to Y_i. The processes generating the vectors Y_i and R_i are referred to as the measurement and missing data, respectively.

We now pay attention to the dropout setting which is a particular case of monotone pattern of missingness in which a missing value whenever it occurs to any subject in the sequence of repeated measurements of the outcome is never followed by any observed measurement on that subject. Alternatively, when dropout occurs, one could use a scalar variable D_i called the dropout indicator, rather than the missing data indicator R_i, defined as $D_i = 1 + \sum_{j=1}^{n_i} R_{ij}$, indicating the occasion at which dropout occurs. Next, we consider the density of the full data (Y_i, R_i), denoted by

$$f(y_i, r_i \mid X_i, W_i, \theta, \psi), \tag{6.1}$$

where X_i and W_i are covariate matrices for the measurement and missing data mechanism, respectively, and the parameter vectors θ and ψ describe the measurement and missingness processes, respectively. The taxonomy, constructed by Rubin (1976) and Little and Rubin (2002), is based on the following factorization

$$f(y_i, r_i \mid X_i, W_i, \theta, \psi) = f(y_i \mid X_i, \theta) f(r_i \mid y_i, W_i, \psi), \tag{6.2}$$

where the first and second factors denote the marginal density of the measurement process and the density of the missing data process, conditional upon the outcomes, respectively. Factorization (6.2) forms the basis of *selection modelling* as the second factor corresponds to the self-selection of individuals into observed and missing groups. Using the reversed factorization, an alternative taxonomy which can be considered is called *pattern mixture models*. They have the following form

$$f(y_i, r_i \mid X_i, W_i, \theta, \psi) = f(y_i \mid r_i, X_i, \theta) f(r_i, W_i, \psi). \tag{6.3}$$

In fact, Equation (6.3) can be described as a mixture of different populations, characterized by the observed missing data pattern. An initial attention of these models were provided by Little and Rubin (1987) and Glynn et al. (1986), while further attention later was provided by many authors, see, for example, Little (1993, 1994). As we mentioned above, Rubin's taxonomy (Rubin, 1976; Little and Rubin, 2002) of missing data process is based on the second factor of Equation (6.2), thus within the selection modelling framework

$$f(r_i \mid y_i, W_i, \psi) = f(r_i \mid y_i^o, y_i^m, W_i, \psi). \tag{6.4}$$

In Equation (6.4), the covariates for the measurement process are assumed measured but suppressed for simplicity sake. The form in Equation (6.4) can be discussed as follows: when the missingness is independent of the responses, i.e.,

$$f(r_i \mid y_i, W_i, \psi) = f(r_i \mid W_i, \psi), \tag{6.5}$$

then this process corresponds to the case of missing completely at random (MCAR). If the missingness process is only independent of the unobserved responses Y_i^m, but depends on the observed responses Y_i^o, consequently, assuming the form

$$f(r_i \mid y_i, W_i, \psi) = f(r_i \mid y_i^o, W_i, \psi), \tag{6.6}$$

then the process corresponds to the case of missing at random (MAR). Finally, when the missingness process depends on the missing data Y_i^m, the process corresponds to the case of missing not at random (MNAR).

As pointed out by Rubin (1976) and Little and Rubin (1987), when MAR mechanism holds, the parameters θ and ψ are functionally independent. In practice, the likelihood of interest then depends upon the factor $f(y_i^o \mid \theta)$. For this reason, when using a likelihood based analysis under the MAR assumption, the missing value mechanism is sometimes said to be “ignorable”. In contrast, if the likelihood of interest only depends upon the factor $f(y_i^m \mid \theta)$, then this is referred to as “non-ignorable” setting. Therefore, when ignorability holds, likelihood-based and Bayesian inferences are valid (Verbeke and Molenberghs, 2000; Molenberghs and Verbeke, 2005).

6.4 Selection models for non-ignorable dropout

In the framework of the selection models, it is not always reasonable to assume that MAR holds, and a wide range of modelling approaches for MNAR data have been proposed. One such is the model proposed by Diggle and Kenward (1994) for continuous outcomes with dropout. In this section, we first describe the Diggle and Kenward’s (1994) selection model for continuous longitudinal data. We then discuss in detail the linear mixed model and the dropout model.

6.4.1 Diggle and Kenward's (1994) model for continuous longitudinal outcomes

A model for longitudinal Gaussian data with non-random dropout was proposed by Diggle and Kenward (1994). Their model assumes that the missingness mechanism is MNAR which combines the multivariate normal model for longitudinal Gaussian data with a logistic regression for the dropout process. From the notation presented in Section 6.3 recall that for subject i, $i = 1, ..., N$, a sequence of responses Y_{ij} is designed to be measured at time points t_{ij}, $j = 1, ..., n$, resulting in a vector of observed outcomes $Y_i = (Y_{i1}, ..., Y_{in_i})'$ of measurements for each subject. Note that although n measurements per subject were planned the vector Y_i is of size $n_i < n$ because of missing observation. In the case of dropout, the vector Y_i is only partially observed. If we let D_i be the occasion where dropout occurs, then $D_i > 1$ since the first observation is assumed to be always observed, and Y_i can be partitioned into the $(D_i - 1)$-dimensional observed component Y_i^o and the $(n_i - D_i + 1)$-dimensional missing component Y_i^m. If no dropout occurs, then $D_i = n_i + 1$, and Y_i equal Y_i^o. For the ith subject, the observed data is (Y_i^o, d_i), thus, the likelihood contribution is proportional to the marginal density function

$$f(y_i, d_i \mid \theta, \psi) = \int f(y_i, d_i \mid \theta, \psi) dy_i^m = \int f(y_i \mid \theta) f(d_i \mid y_i, \psi) dy_i^m. \qquad (6.7)$$

In Equation (6.7), a marginal model for Y_i can be combined with a model for the dropout process, conditional upon the measurement, and the measurement process model including the vectors of unknown parameters, θ and ψ, respectively. More formally, we denote the conditional probability of dropout by $g_j(y_{ij}, h_{ij})$ at time j given the response at time j, and $h_{ij} = (y_{i1}, ..., y_{ij-1})$ which denotes a possibly observed history of subject i until time $t_{i,j-1}$. According to Diggle and Kenward (1994), the dropout process allows the conditional probability for dropout at occasion j, given that the subject was still observed at the previous occasion, to depend upon the history h_{ij} and the possibly unobserved current outcome y_{ij}, but not upon future outcomes y_{ik}, $k > j$. Now, for calculating the dropout probability for each occasion, we use the conditional probabilities $P(D_i = j \mid D_i \geq j, h_{ij}, y_{ij}, \psi)$ which can be expressed as follows:

$$\begin{aligned} P(D_i = j \mid y_i, \psi) &= P(D_i = j \mid h_{ij}, y_{ij}, \psi) \\ &= \begin{cases} P(D_i = j \mid D_i \geq j, h_{ij}, y_{ij}, \psi) & j{=}2 \\ P(D_i = j \mid D_i \geq j, h_{ij}, y_{ij}, \psi) & \\ \times \prod_{k=2}^{j-1} [1 - P(D_i = k \mid D_i \geq k, h_{ik}, y_{ik}, \psi] & j = 3, ..., n_i \\ \prod_{k=2}^{n_i} [1 - P(D_i = k \mid D_i \geq k, h_{ik}, y_{ik}, \psi)] & j = n_i + 1, \end{cases} \end{aligned} \tag{6.8}$$

assuming no missing values at occasion $j = 1$. As mentioned above, Diggle and Kenward (1994) combine a multivariate normal for the measurement process together with a logistic model for the dropout process. To obtain parameter and precision estimates from the combined measurement/dropout model, they use maximum likelihood that involves marginalization over the unobserved components, i.e., Y_i^m. In fact, under repeated measurements for the ith subject, the measurement model assumes that the vector Y_i satisfies the general linear regression model $Y_i \sim N(X_i\beta, V_i)$, where $i = 1, ...N$ in which β is a vector of population-averaged regression coefficients. Further, Verbeke and Molenberghs (2000) pointed out that the matrix V_i can be left unstructured or assumed to be of a specific form, for example, resulting from a linear mixed model, a factor-analytic structure, or spatial covariance structure. As Molenberghs and Kenward (2007), there are some advantages to using an unstructured covariance matrix. More details of these advantages can be found in Molenberghs and Kenward (2007). In the following, we introduce the measurement and dropout models that can be combined for the dropout process.

6.4.1.1 Measurement model

For continuous outcomes, Laird and Ware (1982) proposed linear mixed-effects model, which can be written as follows

$$Y_i = X_i\beta + Z_ib_i + S_i + \varepsilon_i, \tag{6.9}$$

where Y_i is the n_i-dimensional response vector for subject i, $1 \leq i \leq N$, N is the number of subjects, X_i and Z_i are $(n_i \times p)$ and $(n_i \times q)$ known design matrices, β is the p-dimensional vector containing the fixed effects, $b_i \sim N(0, G)$ is the q-dimensional vector containing the random effects. The residual components $\varepsilon_i \sim N(0, \sigma^2 I_{n_i})$, and $b_1, ..., b_n$, $\varepsilon_1, ..., \varepsilon_n$ are assumed to be independent. The serial correlation is captured by the realization of a Gaussian stochastic process, S_i which is assumed to follow a $N(0, \tau^2 H_i)$ law. Here, the serial covariance matrix H_i dependent upon i through the number n_i of observations and through the time points t_{ij} at which measurements are taken. Using the autocorrelation function $\rho(t_{ij} - t_{ik})$, the structure of the matrix H_i is determined. A first simplifying assumption is that H_i depends upon the time interval between two measurements Y_{ij} and Y_{ik}, i.e., $\rho(t_{ij} - t_{ik}) = \rho(u)$, where $u = |t_{ij} - t_{ik}|$ represents the time lag. The autocorrelation function monotonically decreases such that $\rho(0) = 1$ and $\rho(u) \to 0$ as $u \to \infty$. Finally, G is a general $(q \times q)$ covariance matrix with its (i, j) element given by $d_{ij} = d_{ji}$. The random effects in model (6.9) stem from heterogeneity between subjects, in the sense that various aspects of their behavior may exhibit inter-subject random variation. It follows from model (6.9) that, given the random effect b_i, Y_i is normally distributed with mean vector $X_i\beta + Z_i b_i$ and covariance matrix V_i. Thus, after integrating over random effects, inference for the marginal distribution of the outcome Y_i can be written as follows

$$Y_i \sim N(X_i\beta, V_i), \tag{6.10}$$

where $V_i = Z_i G Z_i' + \sigma^2 I_{n_i} + \tau^2 H_i$ is a $(n_i \times n_i)$ covariance matrix which combine both the measurement error and serial components. A simpler case of b_i is a model which includes various fixed effects, a random intercept and allowing Gaussian serial correlation. In this case the covariance matrix V_i becomes

$$V_i = dJ_{n_i} + \sigma^2 I_i + \tau^2 H_i, \tag{6.11}$$

where J_{n_i} is an $(n_i \times n_i)$ matrix with all elements equal to 1, I_i is the $(n_i \times n_i)$ identity matrix, and H_i is determined through the autocorrelation function $\rho^{u_{jk}}$, where μ_{jk} the

Euclidean distance between t_{ij} and t_{ik}, thus

$$H_i = \begin{pmatrix} 1 & \rho^{u_{12}} & \cdots & \rho^{u_{1n}} \\ \\ \rho^{u_{12}} & 1 & \cdots & \rho^{u_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{u_{1n}} & \rho^{u_{2n}} & \cdots & 1 \end{pmatrix}$$

where $\sigma^2 > 0$ and $0 \leq \rho \leq 1$. The covariance structure V_i in Equation (6.11) combines both serial autocorrelation and a shared random effect variance in the estimation. The main problem with this approach, which is due to Diggle and Kenward (1994), is that it assumes stationarity. In practice, if times of measurement are common, the unstructured matrices can be used (aside from very small trials) and for unbalanced times, a random coefficient model.

6.4.1.2 Dropout model

As noted previously, we focus only on incompleteness due to dropout, and thus we assume that the first measurement Y_{i1} is measured for all subjects in the study. In agreement with notation introduced in Section 6.3, the selection model arises when the joint likelihood of the measurement process and the dropout process is factorized as follows

$$f(y_i, r_i \mid X_i, W_i, \theta, \psi) = f(y_i \mid X_i, \theta) f(r_i \mid y_i, W_i, \psi). \tag{6.12}$$

We use the linear mixed-effects model introduced in Equation (6.9) to model the measurements process, together with a logistic regression to describe the dropout process. According to Diggle and Kenward (1994), the model for dropout process is based on a logistics regression for the conditional probability of dropout at occasion j, given the subject is still in the study. Again, the $g_i(y_{ij}, h_{ij})$ denotes this probability of dropout at time j in which $h_{ij} = (Y_{i1}, Y_{i2}, ..., Y_{ij-1})$ is a vector possibly containing all observed measurements up to including occasion j-1 as well as relevant covariates in the conditional probability of dropout model. Theoretically, the dependence on future unobserved measurements is possible to justify but not straightforward, for simplicity, we model dependence only on the first order history. Therefore, modelling the dropout mechanism may be simplified by allowing dropout to depend upon the current measurement and immediately preceding measurement only with corresponding regression coefficients, i.e., ψ_1 and ψ_2. In particular, for subjects with observed measurements, dropout depends on measurement prior to the last measurement ($y_{i,j-1}$) and the current unobserved measurement (y_{ij}). A commonly used version of such a logistic dropout model is

$$logit[P(D_i = j \mid D_i \geq j, h_{ij}, y_{ij}, \psi)] = \psi_0 + \psi_c W_i + \psi_1 y_{i,j-1} + \psi_2 y_{ij}, \tag{6.13}$$

where ψ_0 and ψ_c denote the intercept and the vector of parameters for covariates W_i, respectively. The model in Equation (6.13) contains special cases corresponding to MAR and MCAR mechanism that can be obtained from ($\psi_2 = 0, \psi_1 > 0$) and ($\psi_1 = \psi_2 = 0$), respectively. As pointed out by Diggle and Kenward (1994) and Verbeke and Molenberghs (2000), a likelihood ratio test (LRT) can be used to compare the model fit under a model that assumes the missing data due to dropout are MCAR versus MAR; that is, the LRT for MCAR versus MAR has an approximate χ^2_1 distribution. The LRT statistic is used to test the hypothesis of $\psi_2 = 0$ (i.e., MAR), where dropout is no longer dependent upon the current measurement, and similarly to test the hypothesis of $\psi_1 = \psi_2 = 0$ (i.e., MCAR), where dropout is assumed to occur completely at random, where dropout therefore, does not depend upon the outcome altogether. However, the use of the LRT is inappropriate for hypothesis test for MNAR versus MAR when all the other modelling assumptions hold, due to the fact that the behavior of the LRT statistic for the MNAR parameter ψ_2 is non-standard since the availability of the information on ψ_2 is very rare and interwoven with other features of both measurement and dropout models (Jansen et al., 2006a). In addition, Rotnitzky et al. (2000) illustrated that the limiting distribution is a χ^2 mixture with characteristics controlled by the singular information matrix. Therefore, for the ψ_2 associated with MNAR model, the score equation creates a quasi-linear dependence structure in the system of score equations. This issue is studied in detail by Jansen et al. (2006a), while in the context of an onychomycosis study, Verbeke et al. (2001) have stated that excluding a small amount of measurement error can change drastically the LRT statistic for the MAR null hypothesis, see also for example, Verbeke and Molenberghs (2000). In practice, such a distinction (MAR/MNAR) can only be made using untestable modelling assumptions. (Kenward, 1998). This problem is really laid bare by Molenberghs et al. (2008) which it was show that the formal-based distinction between MAR and MNAR is not possible as for any MNAR model there exists an MAR model that fits the data equally well. The similarity of the MAR and MNAR models with respect to fitting to the observed data, may present different predictions of the unobserved outcomes, conditional upon the observed ones. Hence, it is broadly agreed that the role of such MNAR models is in sensitivity analysis; that is, if the assumptions are changed, the conclusions from the primary (typically MAR) analysis are also changed. Further detail on the precise nature of sensitivity analysis can be found in Molenberghs and Verbeke (2005) and Molenberghs and Kenward (2007).

6.5 Application to the multi-centre trial data

Below we describe the data set that is used in the analysis as well as the application schemes that are used in the analysis of the selection models based on Diggle and Kenward (1994) approach. In terms of the application of the statistical techniques considered in this study, we use the SAS statistical software.

6.5.1 Data set - multi-centre trial data

The example that is used here concerns the analysis of repeated measures designs and demonstrates how to investigate a specific scenario based on dealing with longitudinal data that has a non-ignorable dropout mechanism. The data is based on experiments that rely on the split-plot design assumptions. Such experiments which include repeated measures designs have structures that involve more than one size of experimental unit. In this case, a subject is measured over time where time is one of the factors in the treatment structure of the experiment. By measuring the subject at several different time occasions, the subject is essentially being (split) into parts (time intervals), and the response for each part is measured. The larger experimental unit is the subject or the collection of time intervals which constitute a cluster. The smaller unit is the interval of time during which the subject is exposed to a treatment or an interval just between time measurement. The only departure from the classical split-plot assumptions is because in this case the subplot treatments (time intervals) are not randomized. The data used is from a multi-centre experiment data which is a typical longitudinal example. This data is described and reported in Milliken and Johnson (2009). This example considers an experiment that involves three drugs where each subject was measured repeatedly at three different time points ($j = 1, 2, 3$), where the outcome is described only as a measure of a continuous blood component. The data were collected by three different investigators (or in three different centres) and contains 51 patients. There are 17 patients assigned to each drug. All of the 51 patients were observed at the first occasion, whereas 8 and 10 patients were not seen at the third occasion and at both the second and third occasions, respectively. In Table 6.1, we present the numbers of dropouts by time, centre and drug. The dropouts occur for all drugs and centres. It is clear that drug$_2$ contains more percentages of missing values. The observed data for all subjects are shown in Figure 6.1. The main purpose of this experiment has been to estimate the effects of the drugs on the blood component over time as well as to investigate the relationship between drugs and blood component. In this study, we restrict attention to the influence that might be caused on these effects by the dropout mechanisms. The full results of the analysis of this trial using a likelihood based linear mixed models approach have been reported elsewhere by Milliken and Johnson (2009).

Table 6.1: *Numbers of dropouts in the multi-centre trial*

time	centre-R			centre-S			centre-T		
	drug$_1$	drug$_2$	drug$_3$	drug$_1$	drug$_2$	drug$_3$	drug$_1$	drug$_2$	drug$_3$
1	0	0	0	0	0	0	0	0	0
2	1	1	0	0	0	2	0	3	3
3	2	3	0	1	0	2	3	4	3
total	3	4	0	1	0	4	3	7	6
total		7			5			16	

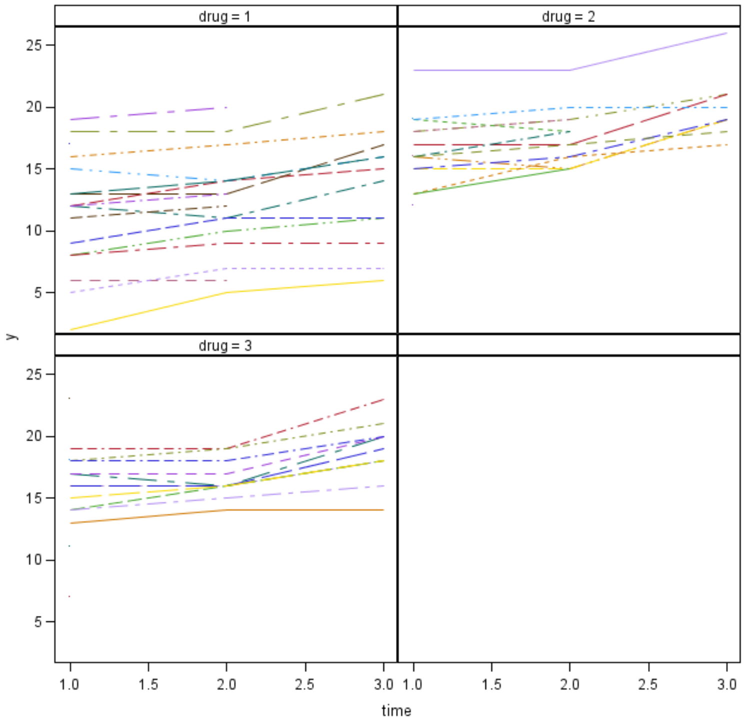


Figure 6.1: *Multi-centre data. Observed data for all subjects*

6.5.2 Diggle-Kenward model applied to the multi-centre trial data

When data are MNAR, an estimation of a selection model can be a complicated issue since the dropout process depends on the unobserved measurements. For example, based on the above mentioned selection model, the dropout process depends in part on the unobserved measurement at the time of dropout. This can be seen as a major complication in assessing a likelihood function but one that can be handled (Diggle and Kenward, 1994). To apply the selection models due to the Diggle-Kenward model based on continuous longitudinal data, in the current computations, we modified the SAS macro that was reported in Dmitrienko et al. (2005) that maximizes the log-likelihood for the model using PROC IML to the case of three drugs as opposed to most applications which are based on two drugs. We carried out an application to the above modelling strategy to the multi-centre data as earlier described. We fit the Diggle and Kenward model in accordance with the MCAR, MAR and MNAR assumptions to our own data set. The three post-baseline visits correspond to the measurements taken at times 1, 2 and 3. In the linear mixed model in Equation (6.9), we allow the inclusion of a variety of fixed effects, a random intercept and Gaussian serial correlation. Furthermore, the dropout model in Equation (6.13) is considered, assuming that the dropout does not depend upon the covariates. Apart from the explicit MCAR, MAR and MNAR versions of this model, we will also conduct an ignorable analysis; that is, an analysis based on the measurement model only, ignoring the dropout model. Firstly, we fit a linear mixed model (LMM) of the form in Equation (6.9) in order to obtain initial values for the parameters estimation of the measurement model. Assuming that the first measurement Y_{i1} is observed for every subject in the study. We thus assume a linear time trend of the response within each drug group. This implies that each profile can be described using two parameters, namely the intercept and a slope. The error matrix is chosen to be of the form (6.11). Since the multi-centre trial data contains fifty-one subjects ($i = 1, ..., 51$) observed at three time points ($j = 1, 2, 3$) for three drugs ($p = 1, 2, 3$), the model can be written as follows

$$Y_{ijp} = \beta_0 + A_p + \beta_1 T_j + \beta_{1p}(TA)_{jp} + \varepsilon_{ijp}, \tag{6.14}$$

where Y_{ijp} is the blood component of subject i at time j on drug p, A_p denotes the pth drug effect, T_j denotes the jth measurement time effect, $(AT)_{jp}$ denotes the interaction effect between time and drug and $\varepsilon_{ip} \sim N(0, V_i)$, where $V_i = dJ_3 + \sigma^2 I_3 + \tau^2 H_i$, with

$$H_i = \begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix}.$$

Using the set to zero constraint (A_1=0), β_0 is the intercept for the drug$_1$ group, $(\beta_0+\alpha_2)$ is the intercept for the drug$_2$ group and for drug$_3$, the intercept is $(\beta_0 + \alpha_3)$, where α denotes the drug fixed effects. These are, respectively, referred to as β_{01}, β_{02} and β_{03} in the results, as we will see in Tables 6.3 and 6.4. The slopes are β_1, $(\beta_1 + \beta_{12})$ and $(\beta_1 + \beta_{13})$ for drug$_1$, drug$_2$ and drug$_3$, respectively, referred to as β_{11}, β_{12} and β_{13} in the results presented in Tables 6.3 and 6.4. The SAS PROC MIXED with REPEATED statement can be used to obtain the initial values. In conforming to the model introduced in Equation (6.13), we use the following logistic regression model for the dropout model probabilities

$$logit[P(D_i = j \mid D_i \geq j, h_{ij}, y_{ij}, \psi)] = \psi_0 + \psi_1 y_{i,j-1} + \psi_2 y_{ij}, \qquad j = 1, 2, 3, \quad (6.15)$$

where ψ_1 and ψ_2 denote the logistic regression coefficients for current and immediately previous observations, respectively, and j denotes the time points. In practice, the combined model for measurement and dropout can be fitted to the data using a generic function maximization routine in the maximum likelihood (Kenward, 1998). In doing so, Diggle and Kenward (1994) used the simplex algorithm of Nelder and Mead (1965) to maximize the log-likelihood. However for the same purpose, we use another optimization method that is available in SAS software, so-called Newton-Raphson ridge optimization. For more detail of this method, see, Dmitrienko et al. (2005). Therefore, we use SAS IML macro which maximizes the likelihood for the model, so as to fit the selection models for the dropout process. The results of initial values for the parameter estimates of the logistic dropout model can be obtained as in Table 6.2 (Dmitrienko et al., 2005).

Table 6.2: *Initial values for the parameters of the dropout model*

	parameter		
Dropout mechanism	ψ_0	ψ_1	ψ_2
MCAR	1		
MAR	$\widehat{\psi_0},MCAR$	1	
MNAR	$\widehat{\psi_0},MAR$	$\widehat{\psi_1},MAR$	1

6.6 Results

Next, we present the results of the application that was discussed earlier. The initial values for the parameters of the linear mixed model are listed in Table 6.3. The results of maximum likelihood for the parameter estimates (standard errors) from the measurement model as well as the results of the variance model under the three missingness mechanisms are presented in Table 6.4. Examining these results, we see that as expected, the parameters estimation and corresponding standard errors of

Table 6.3: *Multi-centre data. Parameter estimates of the linear mixed model, used as initial values for the Diggle-Kenward model*

Effect	Parameter	Estimate	Rounded to initial value
Fixed-effects parameters			
drug_1 intercept	β_{01}	13.9102	13.91
drug_2 intercept	β_{02}	-3.6667	-3.67
drug_3 intercept	β_{03}	0.6853	0.69
drug_1 slope	β_{11}	1.1980	1.20
drug_2 slope	β_{12}	1.5146	1.51
drug_3 slope	β_{13}	1.3481	1.35
Variance parameters			
Random-intercept variance	d	8.9976	9.00
Serial process variance	τ^2	3.4068	3.41
Serial process correlation	ρ	1.0000	1.00
Measurement error variance	σ^2	0.7423	0.74
***p*-value**			
drug_1 effect		0.0061	
drug_2 effect		0.6004	

the fixed effects of the measurement model and the variance model were the same under ignorability, MCAR and MAR mechanisms. This confirms what is expected in theory, see, Molenberghs and Kenward (2007), for example. We now study factors that influence dropout. As discussed above we fit the three dropout models in turn, under the mechanisms MCAR (ψ_1=ψ_2=0), MAR ($\psi_2 = 0$), and MNAR, respectively. Table 6.5 shows the results of the three dropout models that were considered. Here, the evidence for the MNAR setting is only borderline. Thus, under the MNAR assumption, the maximum likelihood estimates for ψ_1 (-0.29) and ψ_2 (0.30) were more or less equal, but with opposite signs, pointing to a relationship between the incremental change and probability of dropout. This finding agrees with the theoretical findings of Molenberghs and Kenward (2007), noting that the dropout often depends upon the increment $y_{ij} - y_{i,j-1}$. This can be justified by the fact that two subsequent measurements are usually

Table 6.4: *Multi-centre data: Maximum likelihood for the parameter estimates (standard errors) under MCAR, MAR and MNAR assumptions without covariate in the dropout model*

Effect	Parameter	MCAR	MAR	MNAR
Measurement model				
drug$_1$ intercept	β_{01}	13.91 (0.92)	13.91 (0.92)	13.90 (0.92)
drug$_2$ intercept	β_{02}	-3.67 (1.30)	-3.67 (1.30)	-3.71 (1.30)
drug$_3$ intercept	β_{03}	0.69 (1.30)	0.69 (1.30)	0.61 (1.32)
drug$_1$ slope	β_{11}	1.20 (0.17)	1.20 (0.17)	1.24 (0.17)
drug$_2$ slope	β_{12}	1.51 (0.19)	1.51 (0.19)	1.60 (0.20)
drug$_3$ slope	β_{13}	1.35 (0.18)	1.35 (0.18)	1.38 (0.18)
Variance model				
Random-intercept variance	d	8.99 (2.63)	8.99 (2.63)	8.99 (2.63)
Serial process variance	τ^2	3.41 (0.56)	3.41 (0.56)	3.35 (0.57)
Serial process correlation	ρ	1.00 (0.00)	1.00 (0.00)	1.00 (0.01)
Measurement error variance	σ^2	0.74 (0.12)	0.74 (0.12)	0.76 (0.12)
-2ℓ		596.99	591.43	595.56

positively correlated (Kenward and Molenberghs, 1999). Furthermore, as can be seen in the dropout model, the parameter estimate ($\psi_2 = 0.30$) in our model is positive indicating a strong association between the dropout and the increment in the outcome variable (blood component) between two successive times. In addition, as mentioned previously, the maximum likelihood estimates of ψ_1 and ψ_2 have different signs, and furthermore, although there is a strong positive association between ψ_1 and ψ_2, the likelihood based 95% confidence interval for these two parameters (ψ_1, ψ_2) is largely contained in the negative-positive quadrant; that is, the intervals for the parameter space where $\psi_1 < 0$ and $\psi_2 > 0$. The full dropout model estimated from the MNAR process is as follows:

$$logit[P(dropout)] = -2.03 - 0.29y_{i,j-1} + 0.30y_{ij}. \tag{6.16}$$

One of our interests is to investigate whether the dropout process is MAR or MCAR,

Table 6.5: *Dropout model: Comparison of the Parameter estimates (standard errors) for MCAR, MAR and MNAR models*

	Dropout mechanism		
Parameter	MCAR	MAR	MNAR
ψ_0	-1.41 (0.31)	-1.52 (0.85)	-2.03 (0.89)
ψ_1		0.01 (0.12)	-0.29 (0.40)
ψ_2			0.30 (0.41)

in other words, whether or not ψ_1=ψ_2=0 in Equation (6.15). The likelihood ratio test is used to compare the model fit under a model that assumes the missing data due to dropout are MCAR versus MAR. The maximum likelihood parameter estimates and minus twice the maximized log-likelihood from the MCAR, MAR and MNAR models appears in Table 6.4. Comparing the log-likelihood estimates from the MAR and MCAR models, we see that the likelihood ratio for the null hypothesis ψ_1=ψ_2=0 is 596.99-591.43=5.56 which is significant with $p < 0.01$ on 1 degree of freedom. The test suggests that an MAR dropout process cannot be ruled out, i.e., there is an evidence in favour of the MAR; that is, dropouts are not completely at random in the context of the assumed model. However, great care has to be taken regarding the sensitivity of the MNAR model to modelling assumptions fit here. To assess the mechanism that the dropout are MNAR, a problem occurs in that neither an LRT statistic between the models that assume the dropout is MAR against MNAR nor an assessment of ψ_2 relative to its standard error is reliable (Jansen et al., 2006a). Therefore, it is impossible to verify that the dropout mechanism is MNAR (Molenberghs et al., 2008).

From the dropout model in Equation (6.15), it is possible to extend the model by using more observed outcomes. According to Diggle and Kenward (1994) and Molenberghs et al. (1997), the dropout in the non-ignorable models tends to depend upon the increment (i.e., the difference between the current and previous measurements, $y_{ij} - y_{i,j-1}$). Including this effect implies a switch to the MAR framework. Some insight into this fitted model can be obtained by rewriting it in terms of the increment. In our case, we obtain the following

$$logit[P(dropout)] = -2.03 + 0.30(y_{ij} - y_{i,j-1}) + 0.01y_{i,j-1}, \tag{6.17}$$

which indicate that dropout is related to the increment $y_{ij} - y_{i,j-1}$, rather than to any of the actual observations y_{ij} or $y_{i,j-1}$, and such that individuals that improve most (large increments) are very likely to dropout from the study. On the other hand, it is useful also to rewrite this with respect to the increment and the sum of the successive measurements. Thereby, by rewriting Equation (6.15), the fitted dropout model equals

$$logit[P(D_i = j \mid D_i \geq j, h_{ij}, y_{ij}, \nu)] = \nu_0 + \nu_1(y_{i,j} + y_{i,j-1}) + \nu_2(y_{ij} - y_{i,j-1}), \qquad j = 1, 2, 3, \tag{6.18}$$

where $\nu_1 = (\psi_1 + \psi_2)/2$ and $\nu_2 = (\psi_1 - \psi_2)/2$. The parameters ν_1 and ν_2 represent dependence on level and increment in the outcome (blood component), and these quantities are likely to be much less strongly correlated than $y_{i,j}$ and $y_{i,j-1}$. Thus, from the fitted MNAR model in Equation (6.18), we have

$$logit[P(dropout)] = -2.03 + 0.005(y_{i,j} + y_{i,j-1}) - 0.295(y_{ij} - y_{i,j-1}), \tag{6.19}$$

which is to say that the probability of dropout increases with larger negative increments. In the other words, those patients who showed or would have shown a greater decrease in the overall level of the blood component from the previous time have a higher probability of dropout. This is said, given the fact that those patients who have a large improvement compared with the previous time and, a sudden shift in profile, are more likely to drop out of the study.

In terms of the significance of the drug effects, the corresponding p-values are displayed in Table 6.6. The p-values of the drug effects at the first point in time does not change much, it being significant in all the three models. However for all cases, the p-values of the drug$_2$ effects were not statistically significant. It is clear from the different dropout models that the drug effects do not differ to a large extent, the impact caused by drugs might be only on the dropout rate through their effects on the blood component. This is similar to the results from Diggle and Kenward (1994) which stated that the drug effects should be integrated directly into the dropout model, either by using it as constants or allowing the relationship between dropout and outcome to differ between the drugs.

Table 6.6: *Multi-centre data: p-values for drug effects under MCAR, MAR and MNAR assumptions*

p-value	MCAR	MAR	MNAR
drug$_1$ effect	0.0062	0.0061	0.0025
drug$_2$ effect	0.6002	0.6006	0.6012

6.7 Discussion and conclusion

In this chapter we have discussed the performance of the selection models based on the Diggle-Kenward approach in terms of the analysis of longitudinal continuous measurements when the dropouts are MNAR. We considered the use of the Diggle and Kenward (1994) model as a tool to assess the sensitivity of a selection model with regard to the modelling assumptions. A model for repeated Gaussian measures, subject to a possibly MNAR assumption were considered. Similar to Diggle and Kenward (1994), a selection model is specified that uses a logistic regression model to describe the dependency of missing data indicators on the longitudinal measurement. In particular, we have investigated the influence on inference that might be caused by the dropout process. In doing so, we carried out an application for analyzing incomplete longitudinal data with dropout. The model was fitted by using an example from a multi-centre clinical trial data involving three treatment arms.

The application notably reveals that dropout increases with one element, i.e., large increments. This implied an occurrence of unfavorable values at the previous time. In fact, this case is, in practical terms, very common in fitting selection models of Diggle-Kenward, we refer to, Verbeke and Molenberghs (2000), Diggle and Kenward (1994) and Molenberghs et al. (1997). Our findings were similar to those of Verbeke and Molenberghs (2000), Diggle and Kenward (1994) and Molenberghs et al. (2001b) in that the example followed in the study yielded parameter estimates for the dropout model that present different signs for current and previous observations indicating the relationships between incremental changes and the probability of dropping out. The results further suggested that there is an evidence in favour of the prevalence of an MAR process rather than an MCAR process in the context of the assumed model. However, Verbeke and Molenberghs (2000), Molenberghs and Verbeke (2005) and Diggle and Kenward (1994) advise one to take care in interpreting the evidence for such conclusions using only the data under analysis.

On the other hand, when all the other modelling assumptions can be guaranteed to hold, the use of the LRT, in a well-defined sense, is inappropriate for hypothesis test for MNAR versus MAR (Jansen et al., 2006a). This is certainly true for the model based on Diggle and Kenward (1994) who investigated the tests of MAR null hypothesis against MNAR, but it is important to note that their tests are conditional on the alternative model holding. Kenward (1998) noted that, in practice, such a distinction (MAR/MNAR) can only be made using untestable modelling assumptions such as a distributional forms. This problem is really laid bare in Molenberghs et al. (2008) which showed that for any MNAR model there exists an MAR model that fits the data equally well, but they differ in the prediction of what is unobserved. Further, they stated that it is not possible to use the fit of an MNAR model for or against an MAR model, unless one puts a priori belief in the posited MNAR model. In other words, as the original MNAR model, the MAR model can give the same estimates of predictions to the observed data, and depending on the same parameter vector. This in line with previous study conducted by Gill et al. (1997). Therefore, this problem of model identifiability can be seen as a complicated issue when considering models for the dropout mechanism. As Molenberghs and Kenward (2007) stated, one suggestion is to conduct a sensitivity analysis of the parameter estimates of the longitudinal model across varying model assumptions about the dropout. Hence, the role of MNAR is in sensitivity analysis; that is, if the assumptions are changed, the conclusions from the primary (typically MAR) analysis are also changed, as the nature of sensitivity comes due to the non-verifiability in the MNAR model from the data. For more discussions of examination the differences between an MNAR model and its MAR counterpart, we recommend Molenberghs et al. (2008) and Kenward (1998) articles.

Finally, in line with previous studies, for example, Verbeke and Molenberghs (2000), Molenberghs and Kenward (2007), Kenward and Molenberghs (1999) and Molenberghs et al. (2003), the selection model of Diggle and Kenward is viewed as a member of the sensitivity analysis framework. An alternative approach to modelling incomplete longitudinal data under a non-ignorable assumption has frequently been proposed in the literature are the pattern mixture models (Little, 1993, 1994). There is also what is known as *influence tools* to deal with incomplete longitudinal data with non-ignorable missingness and these are useful for detecting subjects that cause non-ignorable dropout as well as other subjects that lead to non-random missingness. Hence, in order to assess sensitivity it is useful to obtain further insight into the data by comparing both the selection and the pattern mixture models, for instance, see, Kenward and Molenberghs (1999) and Molenberghs et al. (2003). While it is not the focus of our current study, sensitivity analysis is an important issue of modelling incomplete longitudinal data and should be routinely conducted. To this end, special attention should go to the comparisons between the various sensitivity analysis frameworks.

Chapter 7

Selection and pattern mixture models for modelling longitudinal data with dropout: An application study*

7.1 Abstract

Incomplete data is unavoidable in studies that involve data measured or observed longitudinally on individuals, regardless of how they are well designed. Dropout can potentially cause serious bias problems in the analysis of longitudinal data. In the presence of dropout, an appropriate strategy for analyzing such data would require the definition of a joint model for dropout and measurement processes. This chapter is primarily concerned with selection and pattern mixture models as modelling frameworks that could be used for sensitivity analysis to jointly model the distribution for the dropout process and the longitudinal measurement process. We demonstrate the application of these models for handling dropout in longitudinal data where the dependent variable is missing across time. We restrict attention to the situation in which

[*] A. Satty and H. Mwambi (Accepted). Selection and pattern mixture models for modelling longitudinal data with dropout: An application study. *International Journal of Biostatistics.*

outcomes are continuous. The primary objectives are to investigate the potential influence that dropout might have or exert on the dependent measurement process based on the considered data as well as to deal with incomplete sequences. We apply the methods to a data set arising from a serum cholesterol study. The results obtained from these methods are then compared to help gain additional insight into the serum cholesterol data and assess sensitivity of the assumptions made. Results showed that additional confidence in the findings was gained as both models led to similar results when assessing significant effects, such as marginal treatment effects.

Keywords: Identifying restrictions, Under-identification, Selection models, Pattern mixture models, Sensitivity analysis.

7.2 Introduction

In most longitudinal studies where data are collected over a sequence of time points, missing data are caused by individuals dropping out of the study prior to the time at which the primary endpoint data would be collected. Missingness for longitudinal data often occurs as dropout that is a particular case of missing data. Furthermore, the resulting data obtained from such studies would have a particular type of missing data pattern; that is, a monotone missingness pattern, in which if an individual has missing values for a given time, no data can be obtained for all subsequent times for that individual. In this chapter, our focus will be on this type of missing data pattern. Other types of missingness patterns are possible, such as intermittent missingness, but we focus on dropout which occurs most often in longitudinal studies. The mechanisms that lead to missing data are varied. Rubin (1976) and Little and Rubin (1987) classified these mechanisms into three possible categories, namely data missing completely at random, at random, or not at random. For longitudinal data, when data are missed at random or completely at random, available cases analysis, such as mixed models can be used. In contrast, when data are missed not at random, then using a standard mixed model without accounting for the missingness may lead to biased and inconsistent assessment of study results. Standard strategies of analysis currently assess non-random dropout by performing sensitivity analysis using analytical methods that incorporate non-random dropout in longitudinal data with and without a non-random component. Common families of models for data that are subject to dropout are selection and pattern mixture models.

Selection and pattern mixture models are two alternative and important approaches for dealing with longitudinal data when there are dropouts. They make empirically unverifiable assumptions and require extra constraints to identify the parameter estimates. Both models differ in the way the joint distribution of the measurement and dropout processes are factorized. However, other models that drive both the measurement process and dropout process, such shared-parameter models by Wu and Carrol (1988) and Wu and Bailey (1988, 1989) are also available. We restrict ourselves to the selection and pattern mixture models with dropout that falls under the monotone missing data pattern. A selection model factors the joint distribution into the marginal measurement model that describes the distribution of the complete measurements, and the dropout model that describes the conditional distribution of the dropout indicators, given the observed and unobserved measurements (Diggle and Kenward, 1994). However, in many discussions, for example, Diggle and Kenward (1994), Verbeke and Molenberghs (2000) and Molenberghs and Verbeke (2005), the conclusions obtained from selection models depend on the assumptions made some of which cannot be investigated from the data under analysis. Early reference to such models is found in Heckman (1976) in the econometrics area. The use of pattern mixture models, on the other hand, was originally proposed by Little (1993, 1994) as a viable alternative to selection models. In this approach, models are under-identified; that is, for each dropout pattern the observed data does not provide direct information to identify the distributions for the incomplete patterns. Therefore, to overcome this problem, Little (1993, 1994) solves the under-identification problem through the use of identifying restrictions. Early applications concerning selection and pattern mixture models can be found in Marini et al. (1980) and Glynn et al. (1986). Selection and pattern mixture models are somewhat opposite to each other. That is, these models exploit the conditional probability rule, but they do so in opposite ways. The marginal estimates in selection models can be derived directly, while pattern mixture models estimate the marginal parameters as a weighted average through pattern specific estimates (Little, 1995).

There are several studies in the literature which provide a comprehensive review of these models. The differences between selection models and pattern mixture models have been discussed in many works, for example, Glynn et al. (1986) and Little (1993, 1994). Little (1995) also made an important distinction between selection and pattern mixture models. A comparison of the conclusions based on the selection model with those based on the pattern mixture models have been discussed in Verbeke et al. (2001) and Michiels et al. (2002). Molenberghs et al. (1998a) contrast selection and pattern mixture models. Further discussion of these models can be found in McArdle and Hamagami (1992), Little and Wang (1996), Hedeker and Gibbons (1997), Hogan and Laird (1997), Kenward and Molenberghs (1999), Verbeke and Molenberghs (2000), Molenberghs and Verbeke (2005), Molenberghs and Kenward (2007) and Daniels and Hogan (2008). However, the approach by Daniels and Hogan (2008) is Bayesian based which is not the focus of the current study.

This chapter is primarily concerned with two attractive modelling frameworks to account for non-random dropout, namely selection and pattern mixture models. We demonstrate the application of selection and pattern mixture models for handling dropout in longitudinal data where the dependent variable is missing across time. In particular, we illustrate the application and results of analysis with these models. The under-identification in pattern mixture models is addressed through identifying restrictions, while the use of the selection model is based on Diggle and Kenward's (1994) model. We restrict our attention to the situation in which linear models are used and the outcomes are continuous. The primary objectives are to investigate the potential influence that dropout might exert on the dependent measurement on the considered data as well as how to deal with incomplete sequences. We relate the identified restrictions estimates using a pattern mixture model framework to their corresponding estimates using a selection model framework. We apply the methods to a data set arising from a serum cholesterol study.

Section 7.3 describes the notation and general concepts based on the selection and pattern mixture models. In Section 7.4, we give a discussion of the two families of models that are used in the analysis, namely selection and pattern mixture models. An application study is provided in Section 7.5 including the description of the serum cholesterol data to which our methods will be applied. Full analysis and results of the application is given in Section 7.6. Section 7.7 presents concluding remarks and discussion.

7.3 Notation and concepts

We introduce modelling incompleteness notation which is largely due to Rubin (1976) and Little and Rubin (1987). Let y_{ij} be the response of interest, for the ith study subject, where $i = 1, ...N$, designed to be measured at occasion t_j, where $j = 1, ..., n$. In other words, the original intention was to have n observations per individual. However, due to dropout some individuals end up contributing less than n intended observations. Therefore, generally, we can assume that the ith individual is actually observed n_i times. For subject i and occasion j, define R_{ij}=1, if y_{ij} is observed, and 0, if not. We split y_{ij} into two sub-vectors, y_i^o and y_i^m, representing those y_{ij} for which R_{ij}=1, and R_{ij}=0, respectively. In addition, suppose the missing data occur due to dropout, then, the measurements for each subject can be recorded up to a certain time point, after which all data are unobserved. In this case, a dropout indicator can then be defined as D_i, given by $D_i = 1 + \sum_{j=1}^{n} R_{ij}$, denoting the occasion at which dropout first occurs. In modelling a missing data process, it is often necessary to consider a joint model for the measurement process together with the dropout process. Therefore, we assume the full data density is given by

$$f(y_i, r_i \mid X_i, Z_i, \theta, \psi), \tag{7.1}$$

where X_i denotes the design matrix for fixed effects, Z_i denotes the design matrix for random effects, while θ and ψ represent the vectors of parametrization for the joint distribution. In considering the above model in expression (7.1), we can factorize this joint density function in two possible ways that can facilitate modelling. Specifically, the selection and pattern mixture models mentioned earlier are defined by the conditional factorizations of the joint distribution of Y and R, and both are discussed in more detail in Little (1995) and stated briefly below. A selection model is based on the following factorization

$$f(y_i, r_i \mid X_i, Z_i, \theta, \psi) = f(y_i \mid X_i, Z_i, \theta) f(r_i \mid y_i, X_i, \psi), \tag{7.2}$$

where the first factor in the above factorization represents the marginal density of the measurement process, while the second factor represents the density of the dropout process, conditional on the measurements. An alternative factorization based on the pattern mixture models (Little 1993, 1994) is of the form

$$f(y_i, r_i \mid X_i, Z_i, \theta, \psi) = f(y_i \mid r_i, X_i, Z_i, \theta) f(r_i \mid X_i, \psi). \tag{7.3}$$

This factorized density (7.3) can be seen as a mixture of the conditional distributions, and the model for the measurements depends on the particular missing data pattern. An excellent review of these models is given in Glynn et al. (1986), Little and Rubin (1987), Little (1993, 1994), Hogan and Laird (1997) and Ekholm and Skinner (1998). The missing data processes have been developed by Rubin (1976) and Little and Rubin (1987) through the selection model framework. They make distinctions among different missing data processes. These processes can be formulated based on the second factor of model (7.2), i.e,

$$f(r_i \mid y_i, X_i, \psi) = f(r_i \mid y_i^o, y_i^m, X_i, \psi). \tag{7.4}$$

Thus, if the distribution of missingness process is reduced to $f(r_i \mid y_i, X_i, \psi) = f(r_i, X_i, \psi)$, i.e., the process is independent of the measurements, then the process is defined as missing completely at random (MCAR). If the missingness probability depends on the observed measurement y_i^o, but not on the unobserved measurements y_i^m, i.e, $f(r_i \mid y_i, X_i, \psi) = f(r_i \mid y_i^o, X_i, \psi)$, then the process is termed missing at random (MAR). Finally, data are missing not at random (MNAR) or exhibiting an informative process, when the missingness probability depends on the unobserved measurement, y_i^m, and possibly on the observed measurement, y_i^o, i.e., $f(r_i \mid y_i, X_i, \psi) = f(r_i \mid y_i^o, y_i^m, X_i, \psi)$. In other words, an informative process in expression (7.4) cannot be reduced.

7.4 Selection and pattern mixture models for modelling dropout

We consider the comparison between the selection and pattern mixture models concerning the significant characteristics, such as marginal treatment effects since such a comparison is a useful form of a sensitivity analysis. Specifically, we are interested in parametric selection and pattern-mixture models for modelling dropout. In the following, we briefly review these models.

7.4.1 Selection models for dropout

As mentioned above, a selection model factors the joint distribution into two parts: the marginal measurement model that describes the distribution of the complete measurements and the missingness model that describes the conditional distribution of the response indicators given the observed and unobserved measurements. In other words, in a selection model, we first specify a distribution for the measurement, then propose a manner in which the probability of being observed depends on the data. For continuous outcomes, using a selection model formulation as in equation (7.2), Diggle and Kenward (1994) combine the multivariate Gaussian linear model together with the dropout model. Similarly, we consider the measurement model to be of the linear mixed effects model (Laird and Ware, 1982). Recall that y_{ij} is the response of interest for the ith study subject, where $i = 1, ...N$, at time point j, where $j = 1, ..., n_i$. More generally, the model for y_i the $(n_i \times 1)$ vector of responses for the ith subject can be written as

$$y_i = X_i\beta + Z_ib_i + \varepsilon_i, \tag{7.5}$$

where X_i and Z_i are known $(n_i \times p)$ and $(n_i \times q)$ design matrices for fixed and random effects, respectively, β is the $(p \times 1)$ vector of fixed effects, b_i is the $(q \times 1)$ vector of the random effects distributed as $N(0, G)$, ε_i is the $(n_i \times 1)$ vector of the residual components distributed independently as $N(0, \Sigma_i)$, G is the general $(q \times q)$ covariance matrix with (i, j)th element $d_{ij} = d_{ji}$ and Σ_i is the $(n_i \times n_i)$ error covariance matrix. Then, marginally, the responses y_i are distributed as independent normal

$y_i \sim N(X_i\beta, Z_iGZ_i' + \Sigma_i)$. Here, $\Sigma_i = \sigma^2 H_i + \tau^2 I$, where σ^2 denotes the variance of the serially correlated process, $H_i = (h_{jk}) = (\rho(t_j, t_k))$ denotes the associated correlation matrix, τ^2 pertains to the measurement error variability and I is a $(n_i \times n_i)$ identity matrix.

We assume the missingness is due to dropout only, and that the first measurement y_{i1} is observed for each individual. Again, recall that D_i was defined as the dropout indicator which denote the occasion at which dropout first occurs. Now, let $D_i = d_i$ identify the dropout time for subject i, where $D_i = n{+}1$, if the sequence of measurement is complete. Therefore, the selection models introduced in equation (7.2) arise when the joint likelihood of the measurement and dropout processes is factorized as following

$$f(y_i, D_i \mid X_i, Z_i, \theta, \psi) = f(y_i, \mid X_i, Z_i, \theta) f(d_i \mid y_i, \psi).$$

The model for dropout process is based on a logistic regression for the probability of dropout at occasion j, given the subject was still in the study at the previous occasion. Let $g_i(y_{ij}, h_{ij})$ denote this probability, where h_{ij} represent the history of the measurement process. Thus, one can assume that $g_i(y_{ij}, h_{ij})$ satisfies the model

$$logit[g(h_{ij}, y_{ij})] = logit[p(D_i = j \mid D_i \geq j, h_{ij}, y_{ij})] = \eta(h_{ij}, y_{ij}), \tag{7.6}$$

where $\eta(h_{ij}, y_{ij})$ is the linear predictor depending on h_{ij} and y_{ij}. Modelling the dropout mechanism may be simplified in the expression in equation (7.6) by assuming $\eta(h_{ij}, y_{ij})$ depends only on the current measurement and the previous measurement y_{ij-1}, but not on future measurements or higher order history, with corresponding regression coefficients, ψ_1 and ψ_2. Dependence on future unobserved measurements is not easy to justify therefore it is not modelled here. Higher order history can be included, but we assume first order history for simplicity. This leads to the following logistic expression

$$logit[g(y_{i,j-1}, y_{ij})] = logit[p(D_i = j \mid y_{i,j-1}, y_{ij})] = \psi_0 + \psi_1 y_{i,j-1} + \psi_2 y_{ij}. \tag{7.7}$$

Note that the linear predictor in equation (7.7) may depend on other covariates but in the current model we only include the constant ψ_0. According to Little and Rubin's (1987) terminology introduced in the previous section, and based on the expression in equation (7.7), it is clear that when both parameters ψ_1 and ψ_2 are equal to 0, the dropout mechanism should be MCAR. However, when ψ_1 is not equal to 0, but ψ_2 equal to 0, the dropout mechanism is referred to as MAR, and finally, when ψ_2 is not equal to 0, dropout mechanism is referred to as MNAR. Here, we note that a likelihood ratio test (LRT) can be used to compare model fit under a model that assumes the missing data due to dropout are MCAR versus MAR (Diggle and Kenward, 1994). The LRT test statistic follows a null asymptotic χ_1^2 distribution. See, Diggle and Kenward (1994) and Molenberghs et al. (1997) for details on the derivation of this statistic. When the LRT test statistic is significant, then it suggests that the least restrictive of the two models is preferred; that is, the model that assumes the dropout is MAR. However, based on the argument of Jansen et al. (2006), we restate that the test for MAR against MNAR is not recommended using the LRT statistic via a model based on the Diggle and Kenward's (1994) type. This is because the behavior of the LRT statistic for the MNAR parameter ψ_2 is non-standard since the availability of the information on ψ_2 is very rare and interwoven with other features of both measurement and dropout models (Jansen et al., 2006). This is specially true when one considers the model based on Diggle and Kenward type, but it is important to realize that their tests are conditional on the alternative model holding. According to Kenward (1998), such a distinction, between a MAR mechanism or a MNAR mechanism, can only be made using untestable modeling assumptions, such as the distributional form. Molenberghs and Kenward (2007) stated that the assumption giving arise to the dropout in a sample cannot be verified by the observed measurements and any test regarding the dropout process can be invalidated. This can be justified by the fact that parameters of the dropout model are dependent in part on dropout. Furthermore, unless one puts a priori belief in the posited MNAR model, the distinction (MAR/MNAR) is not possible, due to the fact that for any dropout model that assumes dropout are MNAR, there is a MAR model that provides exactly the same fit to the data, but the two models differ in the prediction of what is unobserved (Molenberghs et al., 2008). This problem of model identifiability poses a major complication when considering models for the dropout mechanism. Thus, one recommendation is to conduct a sensitivity analysis of the parameters of the measurement model across models that make different assumptions about the dropout process (see, Molenberghs and Kenward, 2007). Therefore, although the dropout process cannot be known via empirical examination, the analysis can be carried out to study differences in parameters estimates of the measurement process across varying assumptions about the dropout.

7.4.2 Pattern mixture models for dropout

Now, we shift our attention to the pattern mixture models that stratify subjects according to their missingness pattern. Under these models, the thinking is that, a separate model is fit for each pattern and then the results can be combined across the different patterns in order to derive an average estimate of the model parameters. Thus, in these models the joint distribution of the longitudinal measurements as well as the missing data indicators is divided into response pattern so that the distribution of the longitudinal measurements depends on the pattern of responses. As mentioned earlier, pattern mixture models are under-identified, or possess non-estimable parameters. Therefore, some identifying constraints are required. Little (1993, 1994) proposed the use of the identifying restrictions in which inestimable parameters of the incomplete patterns are set equal to (functions of) the parameters describing the distribution of the completers to deal with under-identifiability of these models. In fact, there is an alternative major strategy simplified to deal with the under-identifiability of pattern mixture models, called model specification in which the different pattern allows for sharing of certain parameters so that the missing pattern can borrow information from patterns with more data points (Verbeke and Molenberghs, 2000). The advantage of this strategy is that the number of parameters decreases which is in general an issue with pattern mixture models. Detailed strategies of pattern mixture modelling are given in Verbeke and Molebergh (2000), Molenberghs et al. (2003) and Molenberghs and Kenward (2007).

Our primary concern in this study is to apply a pattern mixture model including the identifying restriction strategy. In doing so, we follow Verbeke and Molenberghs (2000) in illustrating the use of this strategy based on the results obtained by Molenberghs et al. (1998b). We are restricting attention to dropout which is a special case of monotone missingness. Let us assume that there are $t = 1, ..., T$ dropout patterns, where the dropout indicator, introduced in Section 2, is $d = t + 1$. The complete data density, for pattern t, can be expressed as

$$f_t(y_1, ..., y_T) = f_t(y_1, ..., y_t) f_t(y_{t+1}, ..., y_T \mid y_1, ..., y_t). \tag{7.8}$$

It is clear from equation (7.8) that the first factor $f_t(y_1, ..., y_t)$ is identified from the observed data assuming the first factor is known, and modeled using the observed data. Whereas the second factor is not identifiable from the observed data. In order to identify the second component, the identifying restriction can be applied (Verbeke and Molenberghs, 2000). It is often necessary to base identification on all patterns for which a given component is identified. We denote this component by y_s. Thus, this can be described as

$$f_t(y_s \mid y_1, ..., y_{s-1}) = \sum_{j=s}^{T} \omega_{sj} f_j(y_s \mid y_1, ...y_{s-1}), \qquad s = t + 1, ..., T. \tag{7.9}$$

We denote the set of ω_{sj} used by the vector ω_s, components of which are typically non-negative. Every ω_s that sums to 1 provides a valid identification scheme. Hence, by incorporating equation (7.9) into (7.8), we have

$$f_t(y_1, ..., y_T) = f_t(y_1, ..., y_t) \prod_{s=0}^{T-t-1} \left[\sum_{j=T-s}^{T} \omega_{T-s,j} f_j(y_{T-s} \mid y_1, ..., y_{T-s-1}) \right] \tag{7.10}$$

To establish the complete data density, it is clear in equation (7.10) whose information can be used to complement the observed data density in pattern t. There are three sets of identifying restrictions associated with such choices of ω_s. Complete case missing values (CCMV) that were proposed by Little (1993) use the following identification

$$f_t(y_s \mid y_1, ..., y_{s-1}) = f_T(y_s \mid y_1, ..., y_{s-1}), \qquad s = t + 1, ..., T,$$

corresponding to $\omega_{sT} = 1$ and all others equal 0, which is to say that identification is always done from the completers's pattern. Alternative restrictions are based on so called neighboring case missing values (NCMV). In these restrictions, the nearest identified pattern can be used as follows

$$f_t(y_s \mid y_1, ..., y_{s-1}) = f_s(y_s \mid y_1, ..., y_{s-1}), \qquad s = t + 1, ..., T.$$

The NCMV restriction follows from setting $\omega_s = 1$ and all others equal 0. Finally, the third case for equation (7.10) is the available case missing values (ACMV). With regard to the corresponding ω_s for ACMV, there always is a unique choice. Molenberghs et al. (1998b) show that the corresponding ω_s can have the following components

$$\omega_{sj} = \frac{\alpha_j f_j(y_1, ..., y_{s-1})}{\sum_{\ell=s}^{T} \alpha_\ell f_\ell(y_1, ..., y_{s-1})}, \qquad j = s, ..., T, \tag{7.11}$$

where α_j is the fraction of observations in pattern j. Clearly, ω_{sj} defined by (7.11) contains positive components and sum to 1. That is, a valid density function is defined. The selection and pattern mixture families can be connected using this MAR-ACMV link. The ACMV is reserved for a counterpart of MAR in the pattern mixture setting. A specific counterpart to MNAR selection models has been studied by Kenward et al. (2003).

7.5 Application to the NCGS data

7.5.1 The data

In this section, we describe the application of the selection and pattern mixtures models to data from the National Cooperative Gallstone Study (NCGS). Further background details of this experiment are given in Schoenfield and Lachin (1981) and in its accompanying discussion. In this study, 103 patients were randomly assigned to three treatment groups corresponding to two doses; that is, high-dose (750 mg per day), low-dose (375 mg per day) and placebo, and were to be treated for four weeks. The current analysis is based on a subset of the data on patients who had floating gallstones and who were assigned to the high-dose and placebo groups. In the NCGS it was suggested that chenodiol would dissolve gallstones but in doing so might increase levels of serum cholesterol. As a result, serum cholesterol (mg/dL) was measured at baseline and at 6, 12, 20 and 24 weeks of follow-up. In this experiment, many cholesterol measurements contain missing values because of missed visits, laboratory specimens were lost or inadequate, or patient follow-up was terminated. In addition, all subjects have observed values at time 6. One group of individuals received study treatment (drug and placebo), but dropped out of the study before the scheduled post-baseline time. These individuals dropped out of the study at time point 12. However, other individuals dropped out of the study either at time point 20 or 24. Therefore, the data presents three possible dropout patterns (dropout at time points 12, 20, or 24). All 103 patients are observed at the first occasion, whereas there are 93, 78 and 67 patients seen at the second, third and fourth weeks, respectively. The percentage of patients that are still in the study after each week is tabulated in Table 7.1 by treatment arm. Figure 7.1 represents the means across weeks by treatment group. A primary objective of this trial was to study the safety of the drug chenodiol for the treatment of cholesterol gallstones. In what follows, we restrict our attention to examination of more than just this association between treatment and cholesterol. That is, we investigate the potential influence of dropout on the outcome of interest, the serum cholesterol, as well as the interactive effect of dropout with week and treatment-related influences on the serum cholesterol. The focus here will be on the parameter estimates, standard errors and p-values.

Table 7.1: NCGS data: Percentage of patients still in study, by treatment arm (Drug=high-dose (750 mg per day))

week	drug	placebo
6	100	100
12	45	62
20	57	63
24	46	69

7.5.2 Fitting selection model

First, we consider fitting the selection model. In line with Diggle and Kenward (1994), we fit the selection models to the serum cholesterol data by combining the measurement model with the logistic regression for dropout model. The combined model for measurement/dropout will be fitted to the serum cholesterol by maximum likelihood using a generic function maximization routine. We use the linear mixed effects model

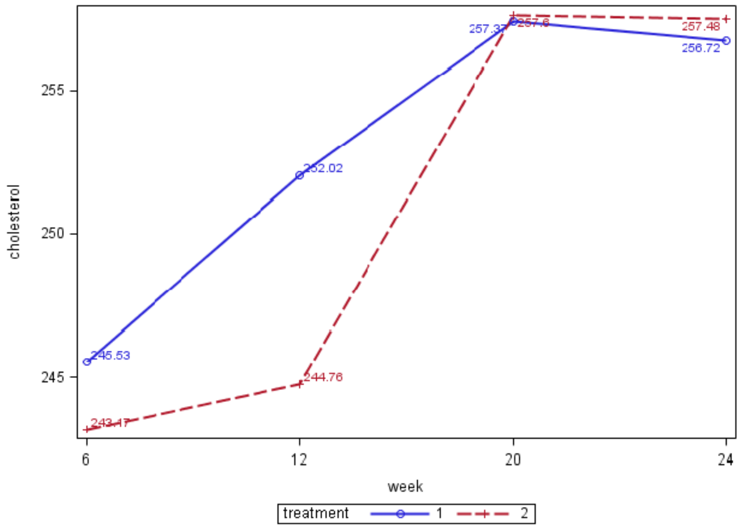


Figure 7.1: *Serum cholesterol data. Means across weeks by treatment*

of the form in Equation (7.5) in order to obtain initial values for the parameters of the measurement model. In the fitted model, we assume different intercepts and treatment effects for each of the four time points, with a (4×4) unstructured variance-covariance matrix. Specifically, we consider a multivariate normal model, with unconstrained time trend under placebo and an occasion-specific treatment effect. Since serum cholesterol data consist of 103 subjects (i=1,...,103) on four time points (j=6, 12, 20 and 24), the model can be written as

$$Y_{ij} = \beta_{j1} + \beta_{j2} G_i + \varepsilon_{ij}, \tag{7.12}$$

where $G_i = 0$ for placebo and $G_1 = 1$ for active drug. In this way, the parameter estimates and standard errors as well as p-values for the eight mean model parameters can be obtained. To fit this model, we use SAS procedure MIXED with REPEATED statement. Next, we consider the dropout model. The dropout will be allowed to be independent of covariates. We fit the model with an intercept, an effect for previous outcome and an effect for the current unobserved measurement, corresponding to MCAR, MAR and MNAR, respectively. Dependence on future unobserved measurements is theoretically possible, but for simplicity, we model dependence on the current unobserved measurements. The probability of serum cholesterol is assumed to follow the logistic regression model (a commonly used model for dropout process, see, Molenberghs and Verbeke, 2005) in Equation (7.7). Therefore, the logistic regression model consists of three parameters; that is, an intercept (ψ_0), the effect of the measurement prior to dropout (ψ_1) and the effect of the measurement at the time of dropout (ψ_2). Consequently, for the four time points j=6, 12, 20 and 24, the model can be expressed as follows

$$logit[g(y_{ij-1}, y_{ij})] = logit[p(D_i = j \mid y_{ij-1}, y_{ij})] = \psi_0 + \psi_1 y_{ij-1} + \psi_2 y_{ij}, \qquad j = 6, 12, 20, 24. \tag{7.13}$$

Estimation of a selection model for MNAR can be seen as major complication as the dropout indicators depend on the unobserved measurement. For example, in the selection model mentioned above, the dropout indicators depend in part on the unobserved longitudinal measurements at the time of dropout. This leads to complexity in assessing the likelihood function, however, one that can be handled (Diggle and Kenward, 1994). Virtually, the parameters were estimated using a code written in SAS provided by Dmitrienko et al. (2005) that maximizes the log-likelihood for the model using PROC IML.

7.5.3 Fitting pattern mixture models

Now, we turn our attention to fitting the pattern mixture models using the strategy outlined in Section 7.4, making CCMV, NCMV and ACMV identifying restrictions. To fit pattern mixture models through identifying restrictions, three steps in the analysis procedure are needed (for details of implementation, see Molenberghs and Kenward, 2007).

Step 1: Fit the initial model to the observed data within each of the patterns:

$$f_t(y_1, ..., y_t), \tag{7.14}$$

where $t = 1, ..., T$ indicate the observed dropout times in the data set. In this step, we fit a separate model within each pattern, then the resulting parameter estimates and their estimated variance-covariance matrices were used to extrapolate the patterns.

Step 2: Select an identification scheme to determine the conditional distributions of the unobserved measurements, given the observed ones

$$f_t(y_{t+1}, ..., y_T \mid y_1, ..., y_t). \tag{7.15}$$

As stated earlier, each of such conditional distributions is a mixture of known normal densities for continuous repeated measures. According to the weights w_s introduced in Equation (7.9), an easy way to simulate values from the mixture distribution is to randomly select a component of the mixture and then draw from it. In this regard, we choose an identifying restriction, mentioned earlier, to define the conditional distributions of the unobserved measurements, conditional upon the observed ones.

Step 3: Fit a model to the so-augmented data. Multiple imputation (MI) can be used to fit such models by aiding to draw values for the unobserved components, conditional upon the observed outcomes and correct pattern-specific density in model (7.15). Here, we notice that MI is a technique that imputes the dropouts multiple times in order to construct multiple complete data sets. For more detail of this technique, we recommend Rubin's (1987) book. Analytically, MI involves three steps, imputation, analysis and combination. Thus, the identifying step corresponds to the so-called imputation step, and the final model corresponds to the analysis step.

Step 4: The combination step, is where the inferences from a number of imputations are drawn together and combined into a single one. The goal being to pool the simplicity of imputation strategies, without bias in both point estimates and measures of precision. After applying each of the three restrictions, as above introduced, the same model as before being fitted (7.12) is analyzed. The model is parameterized as follows: different intercepts and treatment effects for each of the four time points, with (4×4) unstructured covariance matrix for each pattern. The number of imputations (M) must be decided. The established advice by Rubin (1987) is that the number of imputations can generally be constrained to fewer than 10 under most realistic circumstances. Also, many researches tend to support Rubin's (1987) results (Schafer, 1999). We draw multiple imputations 5 times, since a small number of imputed data sets (typically 5) can be substantial enough to materially affect conclusions (Schafer, 1999). In addition, Schafer and Olsen (1998) recommended the use of M=5 before the results are combined. Thus the option M=5 gives the desired relative efficiency in estimating the parameters for the analysis model (7.12). In this way, we ended up with totally five multiply-imputed data sets for each choice of identifying restriction strategy which can be analysed using several possible models. Once the imputations have been generated, the final analysis model from each completed data sets is fitted and MI inference conducted. The parameter and precision estimates can be obtained using classical MI machinery. In particular, the asymptotic covariance matrix of the form

$$V = W + \left(\frac{M+1}{M}\right) B, \tag{7.16}$$

where W denotes the average within-imputation variance and B the between-imputation variance (Rubin, 1987). The analysis of identifying restrictions, fitting of imputed data, and a combination of the results into a single inference was implemented using the SAS macro. This SAS macro dealt with the analysis of the three types of identifying restrictions as follows. First, fit the linear mixed model per pattern using SORT and MIXED procedures. Second, complete the data using ACMV, CCMV and NCMV restrictions. This is done by using IML and MI procedures. Third, analyze the 5 completed data sets using a linear-mixed model using PROC MIXED. Fourth, combine the results from the 5 analyzed models using PROC MIANALYZE.

7.6 Results

Table 7.2 shows the parameter estimates, standard errors and p-values of the fixed effects for the selection model, including the eight mean model parameters, all into the marginal measurement model as well as in the logistic dropout model. Interestingly, the comparison of the MCAR and MAR produces the same results when compared to those of the ignorable analysis, except for negligible differences, as seen in the standard errors. These results are in line with theoretical findings, see, for example, Molenberghs and Kenward (2007). In the context of the assumed model, when examining the statistical significance of the results in the dropout model, the LRT test statistic for comparing the MAR and MCAR models is 17.1. The corresponding tail probability from χ^2 on 1 degree of freedom is $p < 0.001$ which is significant. This indicates that there is a significant evidence for MAR. In other words, dropout completely at random can be ruled out in the context of the assumed model. However, great care has to be taken with such a conclusion (Molenberghs et al., 1997; Molenberghs and Verbeke, 2005). To assess the mechanism that the dropout are MNAR, a problem occurs in that neither an LRT statistic between the models that assume the dropout is MAR against MNAR nor an assessment of ψ_2 relative to its standard error is reliable (Jansen et al., 2006a). Consequently, it is not possible to verify the mechanism that the dropout is MNAR (see, Molenberghs, et al., 2008). One of our interests lies in the marginal treatment effect. There is no overall treatment effect and p-values between the three models do not vary too much. However, the situation is different for the occasion-specific treatment effects considered here. For all weeks, all four p-values for the treatment effects indicate non-significance, whereas for all cases the p-values are certainly highly significance ($p < 0.0001$) for all intercepts. Now, we discuss factors which influence dropout. In doing so, in the full selection models, the logistic regression for dropout is modelled based on (7.13). As can be seen in Table 7.2, the maximum likelihood estimates for ψ_1 (0.04) and ψ_2 (-0.16) are not necessarily equal, however, their signs are different. This finding is not surprising. It confirms the argument put forward by Molenberghs et al. (2001a). They pointed out that since two subsequent measurements are usually positively correlated, the dropout model can depend on the increment, i.e., $y_{ij} - y_{i,j-1}$. The dropout estimated from the MNAR model is as follows

$$logit[p(D_i = j \mid y_{ij-1}, y_{ij})] = -1.64 - 0.12y_{i,j-1} - 0.16(y_{ij} - y_{i,j-1}). \tag{7.17}$$

However, some insight into this fitted model can be obtained by the re-parameterizing the dropout parameters with respect to increment and the sum of the successive measurements. Therefore, we re-parameterize the dropout probabilities from the dropout model as in Equation (7.13) to obtain

$$logit[p(D_i = j \mid y_{ij-1}, y_{ij})] = \vartheta_0 + \vartheta_1(y_{i,j} + y_{i,j-1}) + \vartheta_2(y_{ij} - y_{i,j-1}), \qquad j = 6, 12, 20, 24. \tag{7.18}$$

Here, $\vartheta_1 = (\psi_1 + \psi_2)/2$ and $\vartheta_2 = (\psi_1 - \psi_2)/2$. These parameters represent dependence on level and increment in the serum cholesterol, and these quantities are likely to be much less strongly correlated than are y_{ij} and $y_{i,j-1}$. Rewriting the fitted MNAR model as in Equation (7.18),

$$logit[p(D_i = j \mid y_{ij-1}, y_{ij})] = -1.64 - 0.06(y_{i,j} + y_{i,j-1}) + 0.10(y_{ij} - y_{i,j-1}), \tag{7.19}$$

suggests that the probability of dropout increases with larger negative increments. In

Table 7.2: Maximum likelihood for the parameter estimates - Est. (standard errors - S.e.) and p-values, resulting from the selection model under complete cases analysis, MCAR, MAR and MNAR.

		Complete cases		MCAR		MAR		MNAR	
Effect	Parameter	Est.(S.e.)	P-value	Est.(S.e.)	P-value	Est.(S.e.)	P-value	Est.(S.e.)	P-value
Measurement model									
intercept$_6$	β_{11}	243.17 (6.74)	< 0.0001	243.17 (6.74)	< 0.0001	243.17 (6.74)	< 0.0001	243.17 (6.74)	< 0.0001
intercept$_{12}$	β_{21}	244.93 (6.46)	< 0.0001	244.93 (6.46)	< 0.0001	244.93 (6.45)	< 0.0001	243.98 (6.46)	< 0.0001
intercept$_{20}$	β_{31}	258.92 (6.70)	< 0.0001	258.92 (6.69)	< 0.0001	258.91 (6.70)	< 0.0001	258.13 (6.70)	< 0.0001
intercept$_{24}$	β_{41}	257.08 (8.03)	< 0.0001	257.08 (8.02)	< 0.0001	257.08 (8.03)	< 0.0001	256.28 (8.99)	< 0.0001
treatment effect$_6$	β_{12}	2.36 (8.69)	0.786	2.36 (8.69)	0.784	2.36 (8.68)	0.788	2.36 (8.69)	0.732
treatment effect$_{12}$	β_{22}	6.41 (8.36)	0.445	6.41 (8.32)	0.478	6.41 (8.36)	0.461	6.54 (8.36)	0.441
treatment effect$_{20}$	β_{32}	-5.77 (8.78)	0.512	-5.77 (8.78)	0.535	-5.77 (8.79)	0.427	-5.83 (8.79)	0.456
treatment effect$_{24}$	β_{42}	-2.06 (10.70)	0.847	-2.06 (10.70)	0.883	-2.06 (10.74)	0.856	-2.59 (10.79)	0.912
dropout model process									
intercept	ψ_0			-1.88(0.11)		-1.73(0.14)		-1.64(0.27)	
previous measurement	ψ_1					-0.20(0.05)		0.04(0.02)	
current measurement	ψ_2							-0.16(0.08)	
-2 log-likelihood		3313.3		3346.4		3329.3		3327.7	

Note: MCAR=missing completely at random; MAR=missing at random; MNAR=missing not at random other words, those patients with a greater increase in the overall level of the serum cholesterol from the previous week have a higher probability of dropping out of the experiment.

Turning to the pattern mixture model, the results of the three types of identifying restrictions are listed in Table 7.3. Examining these results we see that the estimates for the corresponding parameters are comparable and their numerical values are indeed very close to each other under the three possible restrictions, namely ACMV, CCMV and NCMV. It can be seen from the analysis that the association p-values for the marginal effect assessments are all nonsignificant, their p-values being all greater than 0.05. However, the association p-values for the intercepts are highly significant ($p < 0.0001$), in line with the p-values obtained from the selection model analysis. In summary, no significant treatment effect is obtained. These findings confirm those obtained from the selection model formulation which gives more weight to this conclusion. These results can be justified by the fact that pattern mixture models using identifying restrictions play a very similar role to the modelling assumptions in the selection model case (Michiels et al., 1999). Furthermore, the parameter estimates and standard errors for the first marginal effect are equal for all the three restrictions CCMV, NCMV and ACMV, see the effects for intercept$_6$ and treatment$_6$. Such results should be expected considering the fact that the first outcome contained observed data for all subjects that were considered in the analysis.

As shown in the results in Table 7.3, the model building using CCMV, NCMV and ACMV restrictions in contrast to selection model did not allow an estimation of whether the dropout process is MNAR or not, because of differences in the modelling assumptions. This agrees with previous studies (see, for example, Molenberghs et al., 1998b), in that the identifying restrictions in a pattern mixture models context can be used only to relate the model to a MAR mechanism. Consequently, an important issue is to equate results for both the ACMV and MAR to make a clear and useful connection between the selection model and the pattern mixture model frameworks (Verebeke and Molenberghs, 2000; Kenward et al., 2003). With this in mind, the same is true for the selection model, MAR-based ACMV restrictions indicating nonsignificant treatment effects at all weeks. This result means that the treatment effects

Table 7.3: Multiple imputation parameter estimates - Est. (standard errors - S.e.) and p-values resulting from the pattern mixture model using identifying restrictions ACMV, CCMV and NCMV.

		ACMV		CCMV		NCMV	
Effect	Parameter	Est.(S.e.)	p-value	Est.(S.e.)	p-value	Est.(S.e.)	p-value
intercept$_6$	β_{11}	243.17 (6.74)	-	243.17 (6.74)	-	243.17 (6.74)	-
intercept$_{12}$	β_{21}	245.44 (7.06)	< 0.0001	245.36 (6.51)	< 0.0001	245.86 (6.55)	< 0.0001
intercept$_{20}$	β_{31}	255.78 (6.71)	< 0.0001	255.88 (6.83)	< 0.0001	257.99 (6.78)	< 0.0001
intercept$_{24}$	β_{41}	256.59 (8.10)	< 0.0001	256.99 (8.21)	< 0.0001	256.99 (8.08)	< 0.0001
treatment effect$_6$	β_{12}	2.36 (8.69)	-	2.36 (8.69)	-	2.36 (8.69)	-
treatment effect$_{12}$	β_{22}	6.23 (8.39)	0.540	6.16 (8.37)	0.539	5.41 (6.45)	0.716
treatment effect$_{20}$	β_{32}	-5.98 (8.85)	0.484	-5.13 (8.81)	0.475	-6.73 (8.82)	0.290
treatment effect$_{24}$	β_{42}	-2.18 (11.03)	0.627	-2.12 (11.64)	0.565	-1.87 (10.13)	0.629

appear to be independent of the ACMV (MAR) assumption. Although corresponding models include the same effects, the estimates for ACMV are slightly different to those for MAR. These slight differences are to be expected as argued in Kenward et al. (2003) that both models are similar in spirit but not necessarily identical. On the other hand, the parameter estimates and standard errors for the treatment effects obtained by applying NCMV are smaller than those of CCMV and ACMV as seen in some cases. This is to be expected as somewhat CCMV and ACMV pattern mixture models use data from different patterns to multiply impute new values, whereas in NCMV, pattern mixture models take information from the neighboring case patterns only. Further, ACMV and CCMV estimates are closer to each other since many more completers are available than does NCMV. Therefore, additional variability may be introduced because, depending on the nature of the conditional distributions sampled from, data have been borrowed from more distant patterns.

7.7 Discussion and conclusion

In this study, we demonstrated the application of two families of models for analyzing incomplete longitudinal data, where the dependent variable is missing across time. In particular, we illustrated the application and compared results of analysis using these models. We focused on the situation in which outcomes are continuous. The models that were considered were the selection model and the pattern mixture model. Many authors have recommended fitting both families of models to be able to gain extra insight into the data to assess sensitivity to the modelling assumptions and to assess the extent of agreement in results as well (see, Molenberghs and Verbeke, 2005). The study focused on the specific cases of selection model and pattern mixture models; that is, a Diggle and Kenward's (1994) model and an identifying restrictions strategy (Little, 1993, 1994), respectively. In applying the selection model, we used logistic regression for modelling dropout, however, a number of other probabilities can be used, for example, using survival analysis techniques, the length of duration on treatment or placebo before dropout can also be modelled. However, in this study, the survival model for dropout cannot be used because the time to event (dropout) is not exactly determined by design. For example, if someone is not seen at week 12, the exact time to dropout could theoretically be any time between week 6 and 12. The objective was to investigate the potential influence that dropout might have or exert on the dependent measurement on the considered data and to deal with incomplete sequences. The results from the pattern mixture models were analogous to those from the selection model to obtain additional insights into the serum cholesterol data. The application was based on an example from a longitudinal clinical trial data.

Findings in general suggested that the conclusion obtained under both modelling frameworks practically coincide. Thus, one can put more confidence in these results as argued by many authors. For example, Michiels et al. (1999, 2002) have argued that greater confidence in a conclusion can be reached when the analysis of joint applications of these models leads to essentially similar inference. Both families of models were compared and noticeable similarities in results were found. Hence, this begs the question as how, depending on the scientific question of interest such as conditional measurement probabilities, to choose between them. Michiels et al. (1999) argued that the selection model can be recommended as a natural choice when the interest is in the population as a whole, i.e., marginal effects. Whereas, pattern mixture models can be considered, when investigating the differences between subgroups that are identified by their measurement patterns, i.e., pattern-specific.

The selection models suggested that the dropout mechanisms were not completely at random. In other words, in the context of the assumed model, there was a lot of evidence in favour of the prevalence of an MAR rather than an MCAR dropout process. However, many authors, Diggle and Kenward (1994) and Molenberghs and Verbeke (2005) for example, stated that careful consideration is necessary with such a conclusion when using only the data under analysis. A problem arises for dealing with dropout that are MNAR. Given this problem in a longitudinal study, it is important to realize that this assumption gives rise to the dropout that is not likely to be known in the application setting. Therefore, any of the different proposed application methods to address dropout that are MNAR cannot easily be verified. For example, one often does not know if the dropout process is accurately captured by a particular method used. Molenberghs and Kenward (2007) suggested that one should apply several approaches to the same data problem. This is the case when the sensitivity of parameters estimates to the different mechanisms about the dropout process may be investigated, for example, by applying models that allow for the dropout to be MNAR. According to Xu and Blozis (2011), if parameter estimates are comparable under different methods, this can indicate that the dropout process may be ignored. However, if different methods give different parameters estimates of the longitudinal model, this can indicate that the dropout process is a vital element for the description of the data in the analysis.

The structure of the selection dropout model adopted that dropout increases with a unit change in the serum cholesterol; that is, the dropout is related to the larger negative increments $(y_{ij} - y_{i,j-1})$ rather than to any actual observation $(y_{ij} + y_{i,j-1})$, which implies that patients with a greater decrease in the overall level of the serum cholesterol from the previous week have a higher probability of dropping out of the experiment. This situation is very common in practice within a model of the Diggle and Kenward type, and we refer to Molenberghs and Kenward (2007), Diggle and Kenward (1994) and Molenberghs et al. (1997) as examples. Under the modeling scheme applied in this study, it can be seen from the analysis that the treatment effects over all weeks under all ACMV, CCMV and NCMV restrictions were non-significant, and the same is true for the selection model analysis. Therefore, it is clear that there is a strong evidence for no significant treatment in the context of serum cholesterol data. It appeared that the non-significant treatment effects were not conditional upon any dropout mechanism holding. As a result, the conclusions obtained from CCMV, NCMV and NCMV restrictions did not differ considerably. As argued in Molenberghs et al. (2008), the choice between them is not always clear. Although they fit the observed data equally well, the difference between them only becomes clear with respect to estimation of the missing data, conditional upon the observed data.

On the other hand, the use of different models in which the data were analysed, can be considered as a sensitivity analysis. In particular, the use of pattern mixture models including identifying restrictions can be seen as a first tool for assessing the sensitivity of the assumptions made. Further, other more complex or flexible sensitivity analysis are also possible, under new models for the probability of dropout. The analysis conducted here is a typical sensitivity analysis as the serum cholesterol data were analyzed using different assumptions concerning the longitudinal measurements and dropout mechanisms. In particular, both models compared well concerning some aspects, for example, marginal treatment effects. Such comparisons as these can play a vital role in sensitivity analysis by providing additional motivation, for example, when considering the choice between selection and pattern mixture models. In conclusion, because the true model and measurement process as well as dropout process are often unverifiable, the recommendation that in many settings, multiple strategies or models such as selection and pattern mixture models be applied to the same data set in order to investigate the impact of assumption on dropout or missingness is supported.

Chapter 8

Conclusions and recommendations for future work

8.1 Summary and conclusions

This thesis discussed some of the key modelling strategies and basic issues in statistical data analysis to address the missing data problem. The study dealt with both the problem of missing covariates and missing outcomes. The main focus of this thesis was on missing data with a monotone pattern. The different methodologies presented in this thesis have been considered to highlight two ways. On the one hand, Chapters 2, 3, 4 and 5 served to demonstrate comparison of existing approaches providing useful and important information regarding their applications. Chapters 6 and 7, on the other hand, provided techniques that might serve as tools in the context of a sensitivity analysis thereby broadening the possibilities under such. The thesis has been divided into two parts. The first part placed strong emphasis on ignorable missingness area which is often used synonymously with MCAR and MAR. In principle, the MCAR assumption is too strong to generally hold (Molenberghs and Kenward, 2007). In this regard, the scope of the first part was limited to MAR missingness rather than the much stronger assumption, MCAR. This was not intended to be an extensive investigation including conventional methods, such as deletion methods which are valid under MCAR. Rather, the discussion paid attention to modern procedures that can be useful at least under most circumstances in missing data analysis. The aim of this part was to investigate various modelling techniques using application studies, and to specify the most appropriate techniques as well as gain insight into the appropriateness of these techniques for handling incomplete data. The second part focused on MNAR modelling that can be used to deal with the change over time in the outcome score and factors that influence this change in modelling incomplete longitudinal data with continuous outcomes. The aim of the second part was to deal with non-random dropout by explicitly modelling the assumption that caused the dropout and incorporated this additional model into the model for the measurement data, and to assess the sensitivity of the modelling assumptions. What follows is a brief overview of the resulting conclusions from the pertinent chapters.

The focus of Chapter 2 of this thesis was on the comparison of various techniques for the imputation of missing covariates with monotone missingness to estimate regression parameters. The imputation methods investigated were: Markov chain monte carlo (MCMC), regression, propensity score (PS) and last observation carried forward (LOCF). The results of the regression analysis of the imputed models by these imputation routines were assessed not only with respect to resulting estimates for the prescribed model, but also in terms of the imputation bias, efficiency and coverage. The application study was carried out under different missing data rates. The missing data mechanism was assumed to be MAR and the missingness was imposed only on covariates. A concise conclusion of this chapter is that we have universally best methods to deal with missing covariates in estimating regression parameters with monotone missingness. From the results, it appeared that either the MCMC or regression methods of imputation for estimating regression models with monotone missingness are preferable to the PS method which is a non-parametric technique, and the LOCF method which is a single imputation technique. Further, the performance of these methods seemed to be independent of the amount of missing data rates as well as the amount of variability in the covariates.

The work in Chapter 3 dealt with missing data on the outcome or response variable showed that MI is the best approach when compared to IPW with regard to handling or modelling incomplete longitudinal data subject to dropout with continuous outcomes.

Generated missing data were used to investigate the performance of these methods, and the dropout assumed to depend on observed responses. This application study was carried out under different dropout rates and focused on a monotone missing data pattern. The techniques were assessed in terms of the statistical measures that focussed on bias and efficiency. The results corresponding to MI were compared to those based on the IPW. The results were also compared with those obtained from linear mixed model analysis (LMM) since LMM is appropriate for dropout under MAR. This was done so that LMM results could be reference against which MI and IPW were contrasted. The results showed that MI should be the method of choice since it has good properties regarding statistical validity, bias and efficiency. In addition, it became clear that the IPW method does not always produce the best results, even though the mechanism of dropout is MAR, unless it is used in the context of marginal models for discrete outcomes.

Chapter 4 was concerned with comparing the two techniques applied to incomplete Gaussian longitudinal data with MAR dropouts, namely direct likelihood and multiple imputation (MI) methods. For comparison, an application study was conducted under 10%, 15% and 20% dropout rates. The MAR dropout was generated from the complete original data. Furthermore, the results from the multiple imputation (MI) and direct likelihood analysis were analogous to those from the complete data. Overall, both methods offered acceptable performance and yielded similar conclusions. The chapter recommended MI and direct likelihood analysis as a solution for handling incomplete Gaussian longitudinal data due to ignorable dropout. Despite MI needing more computing effort, and it is that more difficult to implement than direct likelihood analysis, it greatly improves the results compared to other methods, such as the deletion family of methods. Furthermore, direct likelihood analysis ignores the missing values and estimates the model using only observed data, and the software tool used must be able to deal with more complicated missing data problems.

Up to this point, it has been made clear that inference of the different methodologies presented in this thesis was heavily based on the continuous outcomes assumption of normally distributed data. However, Chapter 5 dealt with discrete outcome data with missing values requiring the use of another distribution apart from the normal distribution. The chapter compared and contrasted different statistical methods for analyzing incomplete non-Gaussian longitudinal outcomes when the underlying study is subject to dropout. The methods considered included WGEE, MI-GEE and GLMM, representing different strategies to deal with dropout under MAR. These strategies dealt with dropouts using different routes: by weighting, by imputation and by analyzing the data as they were (i.e., without the need to weight or impute, consistent with MAR assumption) for WGEE, MI-GEE and GLMM, respectively. The chapter aimed to explore the performance of these methods in terms of handling dropouts that are MAR. The methods were compared using simulated data sets under several different dropout rates and sample sizes. The correlated binary variables were generated from a logistic marginal model. The comparison was made through the evaluation of bias, efficiency and mean square error. MI-GEE was robust, doing better than all the other methods in terms of small and large sample sizes, regardless of the dropout rates. In addition, care is needed when comparing the GLMM estimates to marginal estimates. Appropriate adjustments need to be applied to GLMM estimates in order to have an approximate marginal interpretation and to become comparable to their GEE counterparts.

All the above methods did not have the capacity to provide an optimal solution to the problem of non-ignorable missingness. This kind of missing data can be seen as a major complication, in particular in the context of longitudinal studies. Selection models in terms of the analysis of incomplete longitudinal continuous measurements due to non-ignorable dropout were the subject of Chapter 6. Attention was paid to sensitivity analysis framework in order to deal with MNAR modelling. A review of MNAR-based models namely selection models were conducted. For selection models, the use of the Diggle-Kenward model was also discussed as a tool to assess the sensitivity of a selection model with regard to modelling assumptions. The fitting of the selection models which is based on linear mixed model for the measurement process and the logistic regression for dropout process was outlined. The main aim of this chapter was to investigate the influence on inference that might be exerted on the considered data by the dropout process. An application study was carried out using data from a multi-centre clinical trial. In this work, the Diggle-Kenward analysis code was extended to be able to handle three treatment arms in contrast to the two arms code commonly used. The conclusions drawn from this chapter were that there was evidence in favour of the prevalence of an MAR process rather than an MCAR process in the context of the assumed model. In addition, there is need to subject incomplete data to different families of models for successful missing data estimation and inference in order to derive more insight into the data.

Chapter 6 placed emphasis on the influence of the selection models framework on the dropout mechanisms and can be seen as a first step in the direction of more formal sensitivity analysis. Selection models in Chapter 6 can be compared to other frameworks, such as pattern mixture models, since the comparison between different models using sensitivity analysis can be useful in terms of the performance and reliability of the final inference. Chapter 7 dealt with two modelling strategies for incomplete longitudinal data using selection and pattern mixture models and demonstrated their applications within the sensitivity analysis framework. Both were jointly applied under certain model assumptions. A selection model based on Diggle-Kenward's type, and the pattern mixture models based on the idea of identifying restrictions. The primary objective in this chapter was to examine the sensitivity analysis of the outcomes of interest in terms of the assumptions used in the estimation procedures. By contrasting these two families of models on a single set of data, a range of conclusions were obtained which provides insight into sensitivity analysis with regard to the assumptions made. The results of the sensitivity analysis were identical and led to coinciding estimates. This explained the assurance that the models have an important impact on the overall study conclusion. Additional confidence was gained in the results of the application as both models led to similar results in significant characteristics, such as marginal treatment effects.

8.2 Recommendations for future work

In the context of planning data collection, study designers must think of study or research designs and data collection strategies that minimize missing data since data collection plays an important role in the problem of missing data for a specific study. This means that careful planning can reduce the amount of missing data although there is no rule concerning the level of missing data that can be acceptable. Thus, at the analysis stage, how to handle the missing data and how to minimize the amount of missing data are main issues that must be considered when planning and designing a study for data collection. In the presence of missing data or dropout, knowing the reasons why the data were missing, as well as exploring the missing data pattern become very important and helpful in choosing the right statistical procedures to approximate missingness. In fact, there is no universal technique for handling all missing data situations, however, there are some rules that can be considered. As such, it is necessary to design a study where the potential pattern of missingness is considered when specifying the primary analysis.

The scope of this thesis was limited to several modelling techniques with monotone missing data pattern. However, there is still a considerable amount of research that needs to be undertaken in the field of missing data. The following recommendations could serve as an agenda for future work. Most of these aspects have been already mentioned in the summaries of each chapter.

- Further research could be undertaken to investigate sensitivity analysis since it is an important issue in modelling incomplete longitudinal data when MNAR holds and should be routinely conducted. To this end, the main idea should be to compare the various sensitivity analysis frameworks. While not a focus of our current study, shared parameter models and local and global influence are other alternative approaches. In this context, a comparison between different sensitivity analysis models need further investigation.

- All fitted modelling techniques in this thesis are discussed under assumption of monotone missing data pattern. Alternative patterns, such as intermittent missingness pattern, could be an important future research area.

- The inference in this thesis is heavily based on the continuous outcomes assumption of normally distributed data. However, other types of data and the use of other distributions need to be further explored.

- In the context of identifying restrictions, we restricted our attention to the continuous data setting. However, an extension of this strategy in the case of categorical data deserves further research.

- When the missing data are particulary problematic, for example, when it is hard to specify that at least some of the missing data are not MNAR, sensitivity analysis that deals with specific types of missing data should be conducted.

- In conclusion, we submit that in the literature review there are several relevant incomplete data areas which seem to attract very little attention. Some of these are: identifiability issues for local and global influence techniques, multiple imputation for recurrent event data and sensitivity of inference to data transformations. In addition, a further route for sensitivity analysis should be investigated particularly the latent-class mixture models which are an extension of the shared parameter models need further examination as it can serve as another tool for sensitivity analysis.

Bibliography

[1] Afifi, A. and Elashoff, R. (1966). Missing observations in multivariate statistics I: Review of the literature. *Journal of the American Statistical Association*, **61**, 595-604.

[2] Allison, P. D. (2000). Multiple imputation for missing data: A cautionary tale. *Sociological Methods and Research*, **28**, 301-309.

[3] Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA: Sage.

[4] Alosh, M. (2010). Modeling longitudinal count data with dropouts. *Pharmaceutical Statistics*, **9**, 35-45.

[5] Anderson, T. W. (1957). Maximum likelihood estimates for the multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, **52**, 200-203.

[6] Anderson, J. A. and Aitkin, M. (1985). Variance component models with binary response: interviewer variability. *Journal of the Royal Statistical Society, Series B*, **47**, 203-210.

[7] Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides and R. E. Schumacker (Eds.), Advanced structural equation modeling: Issues and techniques (pp. 243-277). Mahwah, NJ: Erlbaum.

[8] Baker, S. G. and Laird, N. M. (1988). Regression analysis for categorical variables with outcomes subject to non-ignorable non-response. *Journal of the American Association*, **83**, 62-69.

[9] Baker, S. G. (1992). A simple method for computing the observed information matrix when using the EM algorithm with categorical data. *Journal of Computational and Graphical Statistics*, **1**, 63-76.

[10] Barnard, J. and Meng, X. L. (1999). Applications of multiple imputation in medical studies: from AIDS to NHANES. *Statistical Methods in Medical Research*, **8**, 17-36.

[11] Belin, T. R., Diffendal, G. J., Mack, S., Rubin, D. B., Schafer, J. L., and Zaslavsky, A. M. (1993). Hierarchical logistic regression models for imputation of unresolved enumeration status in undercount estimation (with discussion). *Journal of the American Statistical Assocation*, **88**, 1149-1166.

[12] Beunckens, C., Sotto, C., and Molenberghs, G. (2008). A simulation study comparing weighted estimating equations with multiple imputation based estimating equations for longitudinal binary data. *Computational Statistics and Data Analysis*, **52**, 1533-1548.

[13] Birhanu, T, Molenberghs, G., Sotto, C., and Kenward, M. G. (2011). Doubly robust and multiple-imputation-based generalized estimating equations. *Journal of Biopharmaceutical Statistics*, **21**, 202-225.

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

[15] Breslow, N. E. and Lin, X. (1995). Bias correction in generalised linear models with a single component of dispersion. *Biometrika*, **82**, 81-91.

[16] Brown, C. H. (1983). Asymptotic comparison of missing data procedures for estimating factor loadings. *Psychometrika*, **48**, 269-291.

[17] Burton, A., Altman, D. G., Royston, P., and Holder, R. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, **25**, 4279-4292.

[18] Carpenter, J. R., Kenward, M. G., and Vansteelandt, S. A. (2006). A comparison of multiple imputation and inverse probability weighting for analyses with missing data. *Journal of the Royal Statistical Society: Series A*, **169**, 571-584.

[19] Carter, R. L. (2006). Solutions for missing data in structural equation modeling. *Research and Practice in Assessment*, **1**, 1-6.

[20] Chen, H. Y. (2002). Double semiparametric method for missing covariates in cox regression. *Journal of the American Statistical Association*, **97**, 565-576.

[21] Chen, H. Y. (2004). Nonparametric and semiparametric models for missing covariates in parametric regression. *Journal of the American Statistical Association*, **99**, 1176-1189.

[22] Clayton, D. and Hills, M. (1993). *Statistical methods in epidemiology.* Oxford: Oxford University Press.

[23] Cochran, W. G. (1977). *Sampling techniques.* Ne York: John Willey and Sons, Inc.

[24] Collins, L. M., Schafer, J. L., and Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, **6**, 330-351.

[25] Craig, H. M., Todd, M. S., Sanjay, D., David, J. D., Greet, M., Raymond, J. C., William, Z. P., and Gray, D. T. (2003). Assessing and interpreting treatment effects in longitudinal clinical trials with missing data. *Society of Biological Psychiatry*, **53**, 754-760.

[26] Daniels, M. and Hogan, J. (2008). *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis.* Chapman and Hall/CRC.

[27] Day, S. (1999). *Dictionary for clinical trials.* New York: John Wiley and Sons.

[28] De Leeuw, E. D. (2001). Reducing missing data in surveys: an overview of methods. *Quality and Quantity*, **35**, 147-160.

[29] De Leeuw, E. D. , Hox, J., and Husman, M. (2003). Prevention and treatment of item nonresponse. *Journal of Official Statistics*, **19**, 153-176.

[30] Demirtas, H. and Schafer, J. L. (2003). On the performance of random-coefficient pattern-mixture models for nonignorable dropout. *Statistics in Medicine*, **22**, 2553-2575.

[31] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of Royal Statistical Society: Series B*, **39**, 1-38.

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

[33] Diggle, P. J. and Kenward, M. (1994). Informative dropout in longitudinal data analysis (with discussion). *Applied Statistics*, **43**, 49-94.

[34] Dmitrienko, A., Offen, W. W., Faries, D., Christy Chuang-Stein, J. L., and Molenberghs G. (2005). *Analysis of clinical trial data using the SAS system*, Cary, NC: SAS Publishing.

[35] Fitzmaurice, G. M., Molenberghs, G., and Lipsitz, S. R. (1995). Regression models for longitudinal binary responses with informative dropouts. *Journal of the Royal Statistical Society, Series B*, **57**, 691-704.

[36] Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2004). *Applied longitudinal analysis*. New York: John Willey.

[37] Flanders, W. and Greenland, S. (1991). Analytic methods for two-stage case-control studies and other stratified designs. *Statistics in Medicine*, **10**, 739-747.

[38] Friedman, L. M., Furberg, C. D., and DeMets, D. L. (1988). *Fundamentals of clinical trials*. New York: Springer.

[39] Fowler, J. R., F.J. (1988). *Survey research methods*. Newbury Park, CA: Sage.

[40] Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov chain monte carlo in practice*. London: Chapman and Hall.

[41] Gill, R. D., van der Laan, M. J., and Robins, J. M. (1997). Coarsening at random: characterizations, conjectures and counterexamples. In *Proc. 1st Seattle Symp. Biostatistics: Survival Analysis* (eds D. Y. Lin and T. R. Fleming), pp. 255-294. New York: Springer.

[42] Green, S., Benedetti, J., and Crowley, J. (1997). *Clinical trials in oncology*. London: Champan and Hall.

[43] Glynn, R. J., Laird, N. M., and Rubin, D. B. (1986). *Selection modeling versus mixture modeling with nonignorable nonresponse. In: Drawing Inferences from Self-Selected Samples*. New York: Springer-Verlag.

[44] Grace, Y. Yi. and Wenqing, He. (2009). Median regression models for longitudinal data with dropouts. *Biometrics*, **65**, 618-625.

[45] Graham, J. W. and Schafer, J. L. (1999). On the performance of multiple imputation for multivariate data with small sample size. In R. Hoyle (Ed.), Statistical strategies for small sample research (pp. 1-29). Thousand Oaks, CA: Sage.

[46] Hartley, H. O. and Hocking, R. (1971). The analysis of incomplete data. *Biometrics*, **27**, 783-808.

[47] Heckman, J. J. (1976). The common structure of statistical models of trucation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, **5**, 475-492.

[48] Hedeker, D. and Gibbons, R. D. (1997). Application of random-effects pattern mixture models for missing data in longitudinal studies. *Psychological Methods*, **2**, 64-78.

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

[50] Hogan, J. and Laird, N. (1997). Model-based approaches to analysing incomplete longitudinal and failure time data. *Statistics in Medicine*, **16**, 259-272.

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

[52] Horton, N. J. and Lipsitz, S. R. (2001). Multiple imputation in practice: comparison of software packages for regression models with missing variables. *Journal of the American Statistician*, **55**, 244-254.

[53] Horton, N. J. and Kleinman, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, **61**, 79-90.

[54] Horton, N. J. and Laird, N. M. (1999). Maximum likelihood analysis of generalized linear models with missing covariates. *Statistical Methods in Medical Research*, **8**, 37-50.

[55] Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663-685.

[56] Ibrahim, J. G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association*, **85**, 765-769.

[57] Jansen, I., Molenberghs, G., Aerts, M., Thijs, H., and Van Steen, K. (2003). A local influence approach applied to binary data from a psychiatric study. *Biometrics*, **59**, 410-419.

[58] Jansen, I., Hens, N., Molenberghs, G., Aerts, M., Verbeke, G., and Kenward, M. G. (2006a). The nature of sensitivity in missing not at random models. *Computational Statistics and Data Analysis*, **50**, 830-858.

[59] Jansen, I., Beunckens, C., Molenberghs, G., Verbeke, G., and Mallinckrodt, C. (2006b). Analyzing incomplete discrete longitudinal clinical trial data. *Statistical Science*, **21**, 52-69.

[60] Jansen, I. and Molenberghs, G. (2008). A flexible marginal modeling strategy for non-monotone missing data. *Journal of the Royal Statistical Society: Series A*, **171**, 347-373.

[61] Kahn, H. and Sempos, C. T. (1989). *Statistical methods in epidemiology.* New York: Oxford University Press.

[62] Kenward, M. G., Lesaffre, E., and Molenberghs, G. (1994). An application of maximum likelihood and generalized estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random. *Biometrics*, **50**, 945-953.

[63] Kenward, M. G. (1998). Selection models for repeated measurements with nonrandom dropout: An illustration of sensitivity. *Statistics in Medicine*, **17**, 2723-2732.

[64] Kenward, M. and Molenberghs, G. (1999). Parametric models for incomplete continuous and categorical longitudinal data. *Statistical Methods in Medical Research*, **8**, 51-83.

[65] Kenward, M., Molenberghs, G., and Thijs, H. (2003). Pattern mixture models with proper time dependence. *Biometrika*, **90**, 53-71.

[66] Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**, 963-974.

[67] Laird, N. M. (1988). Missing data in longitudinal studies. *Statistics in Medicine*, **7**, 305-315.

[68] Lavori, P .W., Dawson, R., and Shera, D. (1995). A multiple imputation strategy for clinical trials with truncation of patient data. *Statistics in Medicine*, **14**, 1913-1925.

[69] Li, K. H. (1988). Imputation using markov chains. *Journal of Statistical Computation and Simulation*, **30**, 57-79.

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

[71] Lilienfeld, D. E. and Stolley, P. D. (1994). *Foundations of epidemiology.* New York: Oxford University Press.

[72] Lipsitz, S. R. and Ibrahim, J. G. (1996). Using the EM algorithm for survival data with incomplete categorical covariates. *Lifetime Data Analysis*, **2**, 5-14.

[73] Lipsitz, S. R. and Ibrahim, J. G. (1998). Estimating equations with incomplete categorical covariates in the cox model. *Biometrics*, **54**, 1002-1013.

[74] Little, R. J. A. (1976). Inference about means from incomplete multivariate data. *Biometrica*, **63**, 593-604.

[75] Little, R. J. A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association*, **87**, 1227-1237.

[76] Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, **88**, 125-134.

[77] Little, R. J. A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika*, **81**, 471-483.

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

[79] Little, R. J. A. and Wang, Y. (1996). Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics*, **52**, 98-111.

[80] Little, R. J. A. and Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: John Wiley.

[81] Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd edn). New York: John Wiley and Sons.

[82] Little, R. J., DAgostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J., Frangakis, C., Hogan, J. W., Molenberghs, G., Murphy, S. A., Neaton, J. D., Rotnitzky, A., Scharfstein, D., Shih, W, J., Siegel, J. P., and Stern, H. (2012).

The prevention and treatment of missing data in clinical trials. *The New England Journal of Medicine*, **367**, 1355-1360.

[83] Liu, M., Taylor, J. M., and Belin, T. R. (2000). Multiple imputation and posterior simulation for multivariate missing data in longitudinal studies. *Biometrics*, **56**, 1157-1163.

[84] Louis, T. A. (1982). Finding the observed information when using the EM algorithm. *Journal of Royal Statistical Society: Series B*, **44**, 226-233.

[85] Madow, W. G., Nisselson, H., and Olkin, I. (Eds.) (1983). *Incomplete data in sample surveys, VOL 1: Report and case studies.* Academic Press.

[86] Mallinckrodt, C. H., Clark, W. S., and Stacy, R. D. (2001a). Type I error rates from mixedeffects model repeated measures versus fixed effects analysis of variance with missing values imputed via last observation carried forward. *Drug Information Journal*, **35**, 1215-1225.

[87] Mallinckrodt, C. H., Clark, W. S., and Stacy, R. D. (2001b). Accounting for dropout bias using mixed-effect models. *Journal of Biopharmaceutical Statistics*, **11**, 9-21.

[88] Mallinckrodt, C. H., Clark, W. S., Carroll, R. J., and Molenberghs, G. (2003a). Assessing response profiles from incomplete longitudinal clinical trial data under regulatory considerations. *Journal of Biopharmaceutical Statistics*, **13**, 179-190.

[89] Mallinckrodt, C. H., Sanger, T. M., Dube, S., Debrota, D. J., Molenberghs, G., Carroll, R. J., Zeigler Potter, W. M., and Tollefson, G. D. (2003b). Assessing and interpreting treatment effects in longitudinal clinical trials with missing data. *Biological Psychiatry*, **53**, 754-760.

[90] Marini, M. M., Olsen, A. R., and Rubin, D. B. (1980). Maximum likelihood estimation in panel studies with attrition. *Sociology Methodology*, **1980**, 314-357.

[91] McArdle, J. J. and Hamagami, F. (1992). Modeling incomplete longitudinal and cross-sectional data using laten growth structural models. *Experimental Aging Research*, **18**, 145-166.

[92] McKnight, P. E., McKnight, K. M., Sidani, S., and Figueredo, A. J. (2007). *Missing data: A gentle introduction.* The Guilford Press: New York.

[93] McLachlan, G. J. and Krishnan, T. (1997). *The EM algorithm and extentions.* New York: John Wiley and Sons.

[94] Michiels, B., Molenberghs, G., and Lipsitz, S. R. (1999). Selection models and pattern mixture models for incomplete data with covariates. *Biometrics*, **55**, 978-983.

[95] Michiels, B., Molenberghs, G., Bijnens, L., Vangeneugden, T., and Thijs, H. (2002). Selection models and pattern-mixture models to analyze longitudinal quality of life data subject to dropout. *Statistics in Medicine*, **21**, 1023-1042.

[96] Milliken, G. A. and Johnson, D. E. (2009). *Analysis of messy data. Design Experiments*, volume 1. (2nd edn). Chapman and Hall/CRC.

[97] Molenberghs, G., Kenward, M. G., and Lesaffre, E. (1997). The analysis of longitudinal ordinal data with non-random dropout. *Biometrika*, **84**, 33-44.

[98] Molenberghs, G., Michiels, B., and Kenward, M. (1998a). Pseudo-likelihood for combined selection and pattern-mixture models for missing data problems. *Biometrical Journal*, **40**, 557-572.

[99] Molenberghs, G., Michiels, B., Kenward, M., and Diggle, P. J. (1998b). Missing data mechanism and pattern mixture models. *Statistica Neerlandica*, **52**, 135-161.

[100] Molenberghs, G., Verbeke, G., Thijs, H., Lesaffre, E., and Kenward, M. (2001a). Mastitis in dairy cattle: influence analysis to assess sensitivity of the dropout process. *Computational Statistics and Data Analysis*, **37**, 93-113.

[101] Molengberghs, G., Kenward, M. G., and Goetghebeur, E. (2001b). Sensitivity analysis for incomplete contigency tables: The solvenian plebiscite case. *Applied Statistics*, **50**, 15-29.

[102] Molenberghs, G., Thijs, H., Kenward, M. G., and Verbeke, G. (2003). Sensitivity analysis of continuous incomplete longitudinal outcomes. *Statistics Neerlandica*, **57**, 112-135.

[103] Molenberghs, G., Thijs, H., Jansen, I., Beunckens, C., Kenward, M., Mallinckrodt, C., and Carroll, R. J. (2004). Analyzing incomplete longitudinal clinical trial data. *Biostatistics*, **5**, 445-464.

[104] Molenberghs, G. and Verbeke, G. (2005). *Models for discrete longitudinal data.* New York: Springer.

[105] Molenberghs, G. and Kenward, M. G. (2007). *Missing data in clinical studies.* England: John Wiley and Sons.

[106] Molenberghs, G., Beunckens, C., Sotto, C., and Kenward, M. (2008). Every missing not at random model has got a missing at random counterpart with equal fit. *Journal of Royal Statistical Soceity: Series B*, **70**, 371-388.

[107] Molnar, F. J., Hutton, B., and Fergusson, D. (2008). Does analysis using "last observation carried forward" introduce bias in dementia research?. *Canadian Medical Association Journal*, **179**, 751-753.

[108] Moodie, E. E. M., Delaney, J. A. C., LeFebvre, G., and Platt, R. W. (2008). Missing confounding data in marginal structural models: A comparison of inverse probability weighting and multiple imputation. *The International Journal of Biostatistics*, **4**, 1-13.

[109] Musil, C. M., Warner, C. B., Yobas, P. K., and Jones, S. L. (2002). A comparison of imputation techniques for handling missing data. *Western Journal of Nursing Research*, **24**, 815-829.

[110] Nelder, J. A. and Mead, R. (1965). A simplex method for function minimisation. *The Computer Journal*, **7**, 303-313.

[111] Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., and Bent, D. H. (1975). *SPSS: Statistical Package for the Social Sciences* (2nd edn). New York: McGraw-Hill.

[112] Nordheim, E.V. (1984). Inference from nonrandomly missing categorical data: an example from a genetic study on Turners syndrome. *Journal of the American Statistical Association*, **79**, 772-780.

[113] Peng, C. J., Harwell, M., Liou, S.-M., and Ehman, L. H. (2006). Advances in missing data methods and implications for educational research. In S. Sawilowsky (Ed.), Real data analysis (pp. 31-78). Greenwich, CT: Information Age Publishing.

[114] Peng, C. J. and Zhu, J. (2008). Comparison of two approaches for handling missing covariates in logistic regression. *Education and Psychological Measurement*, **68**, 58-77.

[115] Piantadosi, S. (1997). *Clinical trials: A methodologic prespective.* New York: John Wiley and Sons, Inc.

[116] Pinheiro, J. C. and Bates, D. M. (2000). *Mixed effects models in S and S-Plus.* New York: Springer.

[117] Regoeczi, W. C. and Riedel, M. (2003). The application of missing data estimation models to the problem of unknown victim/offender relationships in homicide cases. *Journal of Quantitative Criminology*, **19**, 155-183.

[118] Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, **89**, 846-866.

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

[120] Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, **90**, 122-129.

[121] Robins, J. M. and Gill. R. (1997). Non-response models for the analysis of non-monotone ignorable missing data. *Statistics in Medicine*, **16**, 39-56.

[122] Robins, J. M., Rotnitzky, A., and Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with non-ignorable non-response. *Journal of the American Statistical Association*, **93**, 1321-1339.

[123] Rosenbaum, P. R. and Rubin, D. B. (1983). The central role propensity score in observational studies for causal seffects. *Biometrika*, **70**, 41-55.

[124] Rotnitzky, A. and Robins, J. M. (1997). Analysis of semi-parametric regression models for repeated outcomes in the presence of missing data. *Statistics in Medicine*, **16**, 81-102.

[125] Rotnitzky, A, Cox, D. R., Bottai, M., and Robins, J. (2000). Likelihood-based inference with singular information matrix. *Bernoulli*, **6**, 243-284.

[126] Rubin, D. B. (1974). Characterizing the estimation of parameters in incomplete data problems. *Journal of the American Statistical Association*, **69**, 467-474.

[127] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, **63**, 581-592.

[128] Rubin, D. B. (1978). Multiple imputations in sample surveys. *Proc. Survey Research Methods Section, Am. Statist. Assoc.*, **1978**, 20-34.

[129] Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics*, **4**, 87-94.

[130] Rubin, D. B. and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, **81**, 366-374.

[131] Rubin, D. B. (1987). *Multiple imputation for non-response in surveys.* New York: John Wiley.

[132] Rubin, D. B. (1991). EM and beyond. *Psychometrika*, **56**, 241-254.

[133] Rubin, D. B., Stern, H. S., and Vehovar, V. (1995). Handling "don't know" survey responses: the case of the Solvenian plebiscite. *Journal of the American Statistical Association*, **90**, 822-828.

[134] Rubin, D. B. (1996). Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association*, **91**, 473-520.

[135] Rudas, T. (2005). Mixture models of missing data. *Quality and Quantity*, **39**, 19-36.

[136] Satty, A. and Mwambi, H. (2012). Imputation methods for estimating regression parameters under a monotone missing covariate pattern: A comparative analysis. *South African Statistical Journal*, **46**, 327-356.

[137] Schafer, J. L., Khare, M., and Ezatti-Rice, T. M. (1993). Multiple imputation of missing data in NHANES III. In Proceedings of the Annual Research Conference, pp. 459-487. Washington, DC: Bureau of the Census.

[138] Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* New York: Champan and Hall.

[139] Schafer, J. L. and Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analysts perspective. *Multivariate Behavioral Research*, **33**, 545-571.

[140] Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*, **8**, 3-15.

[141] Schafer, L. J. (2000). Multiple imputation for missing data problems. Presented January, 24-25. Durham, NC.

[142] Schafer, J. L. and Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, **7**, 147-177.

[143] Schafer, J. L. and Yucel, R. M. (2002). Computational strategies for multivariate linear mixedeffects models with missing values. *Journal of Computational and Graphical Statistics*, **11**, 437-457.

[144] Schafer, J. L. (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica*, **57**, 19-35.

[145] Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for non-ignorable dropout using semiparametric nonresponse models (with discussion). *Journal of the American Statistical Association*, **94**, 1096-1146.

[146] Schoenfield, L. J. and Lachin, J. M. (1981). The steering committee, and the NCGS group, Chenodiol (Chenadeoxycholic Acid) for Dissolution of Gallstones: The National Cooperative Gall- stone Study. *Annals of Internal Medicine*, **95**, 257-282.

[147] Schluchter, M. D. and Jackson, K. L. (1989). Log-linear analysis of censored survival data with partially observed covariates. *Journal of the American Statistical Association*, **84**, 42-52.

[148] Seaman, S. R. and White, L. R. (2011). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, **0**, 1-18.

[149] Selvin, S. (1996). *Statistical analysis of epidemiology data.* New York: Oxford University press.

[150] Siddiqui, O. and Ali, M. W. (1998). A comparison of the random-effects pattern mixture model with last observation carried forward (LOCF) analysis in longitudinal clinical trials with dropouts. *Journal of Biopharmaceutical Statistics*, **8**, 545-563.

[151] Sinharay, S., Stern, H. S., and Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods*, **6**, 317-329.

[152] Skellam, J. G. (1948). A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society, Series B*, **10**, 257-261.

[153] Stiratelli, R., Laird, N., and Ware, J. (1984). Random effects models for serial observations with dichotomous response. *Biometrics*, **40**, 961-972.

[154] Streiner, D. L. (2002). The case of the missing data: Methods of dealing with dropouts and other research vagaries. *Canadian Journal of Psychiatry*, **47**, 68-75.

[155] Stubbendick, A. L. and Ibrahim, J. G. (2006). Likelihood-based inference with nonignorable missing responses and covariates in models for discrete longitudinal data. *Statistica Sinica*, **16**, 1143-1167.

[156] Stumpf, S. A. (1978). A note on handling missing data. *Journal of Management*, **4**, 65-73.

[157] Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, **82**, 528-550.

[158] Thijs, H., Molenberghs, G., Michiels, B., Verbeke, G., and Curran, D. (2002). Strategies to fit pattern-mixture models. *Biostatistics*, **3**, 245-265.

[159] Troxel, A. B., Harrington, D. P., and Lipsitz, S. R. (1998). Analysis of longitudinal data with non-ignorable non-monotone missing values. *Applied Statistics*, **47**, 425-438.

[160] Tsiatis, A. A. (2006). *Semiparametric theory and missing data.* New York: Springer.

[161] Vach, W. and Blettber, M. (1995). Logistic regression with incompletely observed categorical covariatesinvestigating the sensitivity against violation of the missing at random assumption. *Statistics in Medicine*, **14**, 1315-29.

[162] Van Buuren, S. and Van Rijckevorsel, J. L. A. (1992). Imputation of missing categorical data by maximizing internal consistency. *Psychometrika*, **57**, 567-580.

[163] Van Buuren, S., Boshuizen, H. C., and Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, **18**, 681-694.

[164] Van der Lan, M. J. and Robins, J. (2003). *Unified methods for censored longitudinal data and causality.* New York: Springer.

[165] Van Steen, K., Molenberghs, G., Verbeke, G., and Thijs, H. (2001). A local influence approach to sensitivity analysis of incomplete longitudinal ordinal data. *Statistical Modelling*, **1**, 125-142.

[166] Verbeke, G. and Molenberghs, G. (1997). *Linear mixed models in practice. A SAS-oriented approach.* New York: Springer.

[167] Verbeke, G. and Molenberghs, G. (2000). *Linear mixed models for longitudinal data.* New York: Springer.

[168] Verbeke, G., Lesaffre, E., and Spiessens, B. (2001). The practical use of different strategies to handle dropout in longitudinal studies. *Drug Information Journal*, **35**, 419-434.

[169] Verbeke, G. and Molenberghs, G. (2005). Longitudinal and incomplete clinical studies. *International Journal of Statistics*, LXIII, n. **2**, 143-176.

[170] Wachter, K. W. (1993). Comment on hierarchical logistic regression models for imputation of unresolved enumeration status in undercount estimation. *Journal of the American Statistical Assocation*, **88**, 1161-1163.

[171] Wang-Clow, F., Lange, M., Laird, N. M., and ware, J. H. (1995). A simulation study of estimators for rates of change in longitudinal studies with attrition. *Statistics in Medicine*, **14**, 283-297.

[172] Willems, J. P., Saunders, J. T., Hunt, D. E., and Schorling, J. B. (1997). Prevalence of coronary heart disease risk factors among rural blacks: A community-based study. *Southern Medical Journal*, **90**, 814-820.

[173] William, R. M. (2000). Handling missing data in clinical trials: An overview. *Drug Information Journal*, **34**, 525-533.

[174] Wu, M. C. and Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, **44**, 175-188.

[175] Wu, M. C. and Bailey, K. R. (1988). Analysis changes in the presence of informative right censoring caused by death and withdrawal. *Statistics in Medicine*, **7**, 337-346.

[176] Wu, M. C. and Bailey, K. R. (1989). Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. *Biometrics*, **45**, 939-955.

[177] Xu, S. and Blozis, S. A. (2011). Sensitivity analysis of mixed models for incomplete longitudinal data. *Journal of Educational and Behavioral Statistics*, **36**, 237-256.

[178] Yi, G. Y. and Cook, R. J. (2002a). Marginal methods for incomplete longitudinal data arising in clusters. *Journal of the American Statistical Association*, **97**, 1071-1080.

[179] Yi, G. Y. and Cook, R. J. (2002b). Second order estimating equations for clustered longitudinal binary data with missing observations. In Recent Advances in Statistical Methods, Y. P. Chaubey (ed), 352-366. London: World Scientific Publishing Company, Inc.

[180] Yoo, B. (2009). The impact of dichotomization in longitudinal data analysis: A simulation study. *Pharmaceutical Statistics*, **9**, 298-312.

[181] Yuan, Y. C. (2000). Multiple imputation for missing data: Concepts and new development. SUGI Proceedings, pp. 267-25.

[182] Zhang, P. (2003). Multiple imputation: Theory and method. *International Statistical Review*, **71**, 581-592.

[183] Zhao, L. P. and Lipsitz, S. (1992). Design and analysis of two-stage studies. *Statistics in Medicine*, **11**, 769-782.