# A perspective on incomplete data in longitudinal multi-arm clinical trials, with emphasis on pattern-mixture-model based methodology

by

Anna Christina (Anneke) Grobler

Submitted in fulfilment of the academic requirement for the degree of
Doctor of Philosophy
in the Discipline of Statistics
School of Mathematics, Statistics and Computer Science
THE UNIVERSITY OF KWAZULU-NATAL
Durban


April 2014

# Abstract

Missing data are common in longitudinal clinical trials. Rubin described three different missing data mechanisms based on the level of dependence between the missing data process and the measurement process. These are missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Data are MCAR when the probability of dropout is independent of both observed and unobserved data. Data are MAR when the probability of data being missing does not depend on the unobserved data, conditional on the observed data. When neither MCAR nor MAR is valid, data are MNAR.

The aim of this thesis is to discuss statistical methodology required for analysing missing outcome data and provide valid statistical methods for the MAR, MCAR and MNAR scenarios. This thesis does not focus on data analysis where covariate data are missing. Under MCAR complete and available case analyses are valid. When data are MAR multiple imputation, likelihood-based models, inverse probability weighting and Bayesian models are valid. When data are MNAR pattern-mixture, selection and shared-parameter models are valid. These methods are illustrated by an in depth analysis of two data sets with missing data.

The first data set is the SAPiT trial an open label, randomised controlled trial in HIV-tuberculosis co-infected patients. Patients were randomised to three arms; each initiating antiretroviral therapy at a different time. CD4+ count, an indication of HIV progression, was measured at baseline and every 6 months for 24 months. The primary question was whether CD4+ count trajectory over time differed for the three treatment arms. The assumption that missing data are MCAR was not supported by the observed data. We performed a range of sensitivity analyses under both MAR and MNAR assumptions.

The second data set is a placebo-controlled, randomised clinical trial conducted for 8 weeks to determine the effectiveness of hypericum or sertraline in reducing depression, measured by the Hamilton depression scale. The trial randomised 340 participants, with 28% lost to follow-up before Week 8. We performed a sensitivity analysis under different assumptions about the missing data process. The missing data mechanism was not MCAR. Under MAR assumptions, some of the sensitivity analyses found no difference between either of the treatment arms and placebo, while some found a significant difference between sertraline and placebo, but not between hypericum and placebo. This re-analysis contributed to the literature around the effectiveness of St John's Wort because it changed the conclusions of the original analysis.

# Preface

This study represents original work by the author and have not otherwise been submitted in any form for any degree or diploma to any University. Where use has been made of the work of others it is duly acknowledged in the text.

The research work was done in Durban and was supervised by Dr Matthews and Dr Molenberghs.

_____

Signed:
A. Grobler (candidate)

_____

Signed:
Professor G Matthews (supervisor)

_____

Signed:
Professor G Molenberghs (co-supervisor).

# DECLARATION 1 - PLAGIARISM

I, A Grobler declare that

1. The research reported in this thesis, except where otherwise indicated, is my original research.

2. This thesis has not been submitted for any degree or examination at any other university.

3. This thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.

4. This thesis does not contain other persons' 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.

5. 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 sections.

Signed: ……

# DECLARATION 2 - PUBLICATIONS

DETAILS OF CONTRIBUTION TO PUBLICATIONS that form part and/or include research presented in this thesis

**Publication 1 and 2:**
The primary results of the SAPiT study (Chapter 5) have been published prior to work on this thesis commencing. A. Grobler was the statistician who designed the study and analysed the data for these publications. The primary objective of these papers was analysing mortality, which is not the objective focused on in this thesis.

Abdool Karim Q., Abdool Karim S.S., Baxter C., Friedland G., Gengiah T., Gray A., **Grobler A**., Naidoo K., Padayatchi N., El-Sadr W. The SAPIT trial provides essential evidence on risks and benefits of integrated and sequential treatment of HIV and tuberculosis. *South African Medical Journal*, **2010,** 100(12): 808-809

Abdool Karim S.S, Naidoo K**., Grobler A**., Padayatchi N., Baxter C., Gray A., Gengiah T., Gengiah S., Naidoo A., Jithoo N., Nair G., El-Sadr W.M, Friedland G., Abdool Karim Q. Integration of antiretroviral therapy with tuberculosis treatment. *New England Journal of Medicine,* **2011**, 365: 1492-501

**Publication 3:**
Results from the St John's Wort trial was published online. This paper is based on Chapter 6 of this thesis and is given in Appendix 4.

**Grobler, A.C**, Matthews, G, Molenberghs, G. The impact of missing data on clinical trials: a re-analysis of a placebo controlled trial of Hypericum perforatum (St Johns wort) and sertraline in major depressive disorder. Psychopharmacology, 2013, DOI 10.1007/s00213-013-3344-x

**Publication 4:**
A paper was submitted to the South African Statistics Association student competition. This paper has not been published and will be submitted to a journal for consideration for publication.

**Grobler, A.C**, Matthews, G, Molenberghs, G. CD4+ counts in a 3-arm longitudinal clinical trial with substantial missing data: a sensitivity analysis

**Oral presentations:**
2011: SASA conference, Pretoria
**The use of pattern-mixture models in the analysis of longitudinal studies with missing data**

2013: Presentation at Hasselt Statistics Seminar:
Missing data in a longitudinal trial. Example, A 3-armed clinical trial treating HIV and TB co-infection in South Africa.

Signed:

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| ACMV | available case missing values |
| ANOVA | Analysis of Variance |
| AIC | Akaike's Information Criteria |
| AIDS | acquired immune deficiency syndrome |
| BDI | Beck Depression Inventory |
| BIC | Bayesian Information Criteria |
| CCMV | complete case missing values |
| CGI-S | Clinical Global Impressions scale for severity |
| CGI-I | Clinical Global Impressions scale for improvement |
| CHMP | committee for medicinal products for human use |
| CI | confidence interval |
| DIC | Deviance Information Criterion |
| EM | expectation maximization |
| EMEA | European Medicines Agency |
| GAF | Global Assessment of Functioning |
| GEE | generalised estimating equations |
| HAM-D | Hamilton Depression scale |
| HIV | human immunodeficiency virus |
| ICH | International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use |
| IRIS | immune reconstitution inflammatory syndrome |
| ITT | intent to treat |
| IQR | interquartile range |
| LOCF | last observation carried forward |
| MAR | missing at random |
| MCAR | missing completely at random |
| MCMC | Markov Chain Monte Carlo |
| MI | multiple imputation |
| MICE | multivariate imputation by chained equations |
| MNAR | missing not at random |
| NCMV | neighbouring case missing values |
| NMAR | not missing at random |
| OR | odds ratio |
| REML | restricted maximum likelihood |
| SAPiT | **S**tarting **A**ntiretroviral therapy at three **P**oints **i**n **t**uberculosis |
| SSRI | selective serotonin reuptake inhibitor |
| WHO | World Health Organisation |

# Acknowledgements

Thank you to my wonderful supervisors, Glenda and Geert. Glenda, thank you for sitting with me, helping me and guiding me. Thank you for all the time you spent, even though you were always pulled in so many directions. Geert was willing to be a supervisor from half a world away without having ever met me. You were incredibly generous with your time and advice, which I appreciate to no end. You have an amazing way of working with people. Thank you for everything I learnt from both of you.

Thank you to CAPRISA for the best 10 years of my working life. This is a great place to work. Thank you for exposing me to so much and giving me so many opportunities to learn and to be part of something meaningful. Thank you to the entire SAPiT study team and the participants who made it possible. It was a privilege to be part of the study and I appreciate being given the data for this project. Thank you also for the support I received while working on my PhD.

There were countless angels who helped me with this work; from staff at Hasselt University who helped me during my visit there, to a friend of a friend who helped me with WinBugs, and colleagues who acted as sound boards, helped me think and gave support. I will not name all these angels.

Thank you to Francois, my husband. Thank you for always supporting me in everything I do. Thank you for your unwavering belief in me and willingness to take up the slack somewhere else, because I was busy working. A thank you also needs to go my father who bullied me into studying mathematics, which more than 20 years ago was the first step towards getting here. Thank you to my mother, who taught me to work hard and take academics seriously.

# Chapter 1

# Introduction

The statistical analysis of longitudinal data in a clinical trial often has the unavoidable problem of missing data. This poses the question of how one should deal with the missing data in the statistical analysis. It is vital to apply a correct statistical analysis that deals with the missing data to arrive at reliable and unbiased conclusions. This thesis discusses methods of dealing with missing data and provides an in-depth analysis of two longitudinal data sets which have missing data. In the following paragraph the specific aims of the thesis are set out.

## 1.1 Aim of the thesis

The aim of this thesis is to show how available missing data methods can be used in an appropriate statistical analysis of a real life clinical trial, with missing data. We will show how statistical methodology can be used to analyse missing outcome data and provide valid statistical methods for missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) scenarios. This thesis does not focus on scenarios where covariate data are missing, but rather on the analysis when outcome data are missing.

Missing data is a vast topic in statistics. It is not possible to describe every method that could be used in the analysis of missing data. There are valid methods that are not discussed in this thesis, for example propensity scores (Horton and Kleinman, 2007) and instrumental variables. The intention is to discuss some of the most useful methods that could be applied in mainstream statistical analysis using standard statistical software.

This thesis gives a broad overview of missing data methods, with a more in depth focus on pattern-mixture models. Frequentist, likelihood and Bayesian methods are used. The methodology for calculating the variance of estimates when combining estimates from pattern-mixture models will be extended to include the case where there are three or four patterns of missing data using the delta method (Section 3.3.2.1). This methodology has been published for two categories, and this thesis includes the extension to three and four categories, which follows directly from the two category case. These derivations form a new contribution to the application of well-known theory related to the calculation of the variance of the parameter estimates in pattern-mixture models, using the delta method.

The thesis contributes to promoting good practice in two ways. It tries to discourage the use of suboptimal methodology and promotes the use of appropriate methodology and emphasises proper sensitivity analysis.

Throughout the thesis all statistical analyses are done using SAS, with the exception of the Bayesian analyses that were done using OpenBUGS. A further aim of the thesis is to show that appropriate statistical methodology can be applied using standard statistical software. Most of the statistical methods used can also be applied in many other software packages, such as STATA, R, MLwiN and SPSS. The objective of the thesis is not to list all statistical software that can be used or to contrast implementation using different software packages, but to show that it can be done in one of these. SAS is used for the analysis in this study because it is still regarded as the industry standard in clinical trial data analysis and is widely used. SAS can also fit Bayesian models, but OpenBUGS was used to fit the Bayesian models, since it is regarded as the standard software for Bayesian analysis.

This is a thesis in the field of applied statistics. The statistical methods used in the thesis are not new or original and have been developed previously by other authors. The original contribution of the thesis lies in showing that these methods can be applied validly in the analysis of a clinical trial using standard statistical software. In addition, the sensitivity analysis juxtaposing several methods is a unique application to the two data sets presented. In the analysis in Chapter 6 the application of correct methods when data are incomplete leads to a different conclusion than what was previously found with methods that did not take missing data into account. This leads to a different conclusion about the effectiveness of St John's Wort from the conclusion previously published. The data in Chapter 5 has not previously been modelled, and as such the analysis of this data set is new.

The application of the theory to both data sets is made more complicated by the fact that there are three treatment arms. Most applications in the literature only discuss the two arm case. This thesis also shows how some problems unique to this application that are not addressed in methodological discussions can be addressed. For example, the application of the delta method in Section 3.3.2.1 was done in detail because only the application for two dropout categories was available in the literature.

This thesis focuses on continuous, normally distributed outcome data only. The challenges raised by missing data are similar for categorical and non-normal data. The mathematics and some implementations differ and as such this was regarded to be outside the scope of this thesis.

## 1.2 What is a clinical trial?

A clinical trial is a prospective study comparing the effect and value of an intervention against a control in humans. Each participant is followed forward from a well-defined point, which is called baseline. The participants are directly observed and an intervention is applied to all participants in a standardised manner. A clinical trial includes a control group, which does not receive the intervention against which the intervention group is compared. The investigator does not have control over what each individual participant actually does. The investigator can only encourage participants to follow certain procedures. Since it may be impossible to have pure intervention and control groups in the presence of everything people choose to do, an investigator may only be able to compare intervention strategies (Friedman, Furberg, & DeMets, 1998).

The ideal clinical trial is randomised and double blinded. Randomisation is the process by which each participant has the same chance of being assigned to either intervention or control. Chance underlies the allocation process. Neither the participant, nor the investigator, should know what the assignment will be at the time the participant decides to enter the study. Randomisation is the best method for achieving comparability between treatment arms and is the basis for statistical inference (Friedman et al., 1998).

Double-blinding means that the investigator and participant are blinded, or masked, to the identity of the assigned intervention. Several other aspects of a trial can also be blinded, namely the assessment, classification and evaluation of response variables. The goal of blinding is to prevent bias during the data collection process (Friedman et al., 1998).

## 1.3 What are missing data?

In a trial context missing data are data that was intended to be collected, but were not collected. There are different causes for missing data and different amounts of missing data. Some causes of missing data are by design and some are by chance (Horton & Kleinman, 2007). Some sources of missing data affect all data for a specific participant, other sources affect specific items only (Committee for medicinal products for human use, [CHMP], 2010). For example, a specific reading could be missing because a machine broke or a specimen got lost or all data for a particular visit could be missing because the participant did not attend the visit (Carpenter & Kenward, 2007). Some variables are not collected on all participants and some participants may decline to provide some information (Horton & Kleinman, 2007).

Two occurrences of missing data can be distinguished:

1. Intermittent missing data: A participant missed one visit, but attended at least one subsequent follow-up visit.
2. Withdrawal (also called dropout): A participant had no further visits after a certain time point (Carpenter & Kenward, 2007). This could happen because the participant refused to continue further in the study (CHMP, 2010).

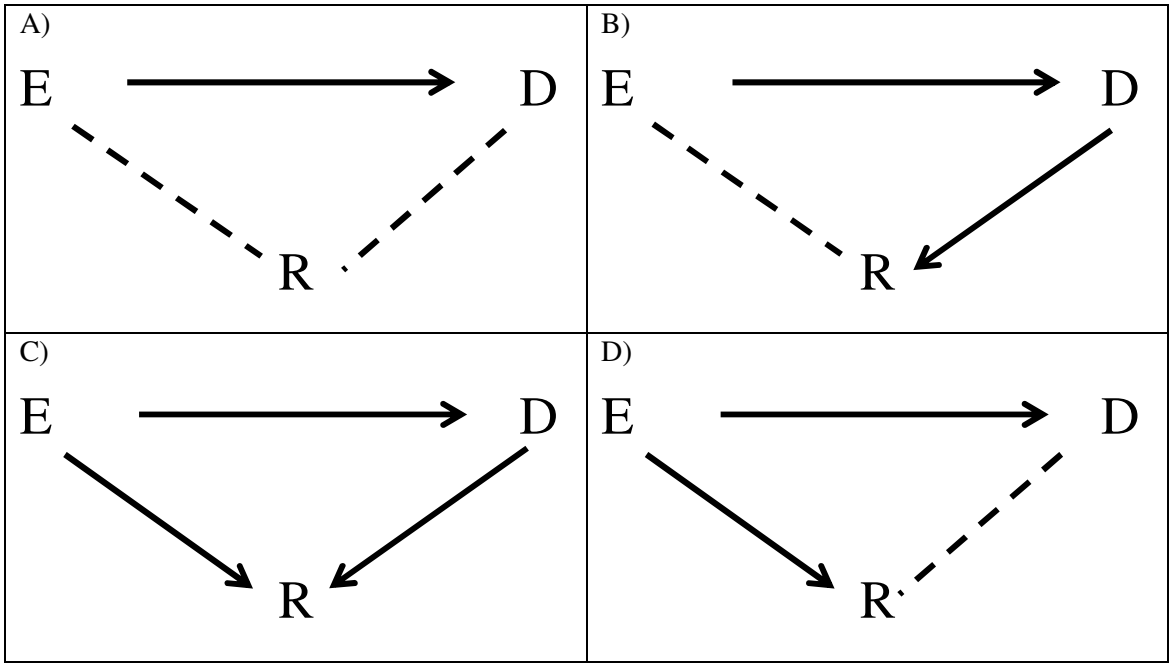
Dropouts in a clinical trial may be for treatment related reasons, for example a participant dropping out because of intolerability of the treatment or due to a lack of a drug effect (National Research Council, 2010). Dropout can also be unrelated to treatment. This can be caused by the participant moving away or if the participant developed an unrelated illness that required stopping the study treatment (Gould, 1980). Dropout can lead to bias in the analysis of a clinical trial if the participants who dropped out are systematically different from those who did not (Hogan, Roy, & Korkontzelou, 2004).

## 1.4 Why missing data matters

Interpretation of the results of a trial is problematic when the number of missing values is substantial (CHMP, 2010). Bias is one of the main concerns in a clinical trial. Bias can be defined as systematic error or the difference between the true value and the value obtained in the trial, due to causes other than sampling error (Friedman et al., 1998). Missing data are a potential source of bias when analysing clinical trials, and this bias is the most important concern resulting from missing data. If participants are excluded, this may influence the comparability of the treatment groups or the representativeness of the study sample in relation to the target population. Both of these could lead to a bias in the estimation of the treatment effect (CHMP, 2010).

An example of this is where patients who were performing less well on the treatment tended to drop out of the study more than patients who performed better on the treatment. A complete case estimation of the trial outcome would then provide a significant overestimate of the true performance of the original patients in the trial, because only the patients with favourable clinical outcome remained in the trial.

The risk of bias in the estimation of the treatment effect from the observed data depends on the relationship between missingness, treatment arm and outcome. Missing values are not expected to lead to bias if they are not related to the real value of the unobserved measurement, for example when poor outcomes are no more likely to be missing than good outcomes. This is illustrated in the causal diagram in A) of Figure 1.1. If the missing observation is related to the real value of the outcome (for example poor outcomes are less likely to be observed than good outcomes), this leads to bias even if the missing values are not related to treatment (CHMP, 2010). This situation is illustrated in B) of Figure 1.1. Missing values lead to bias if they are related to both the treatment arm and the unobserved outcome variable. This is pictured in C) in Figure 1.1. In this case missing values could be more likely in one treatment arm, if participants withdrew from the one arm since the treatment was not effective (CHMP, 2010). In Figure 1.1, A) does not lead to bias because of missing data, whereas both B) and C) can lead to bias.




E: Treatment arm, D: Outcome variable, R: An indicator of whether or not data are missing
Arrows indicate a relationship between variables, whereas dotted lines indicate no relationship

**Figure 1.1: Causal diagrams showing the relationship between treatment arm, outcome and missingness**

Several authors test whether missing values are related to treatment by checking whether missing values are equally likely in all treatment arms. Imbalance in participant withdrawal by intervention arm is not itself a problem, but could be a possible indicator of other problems (Carpenter & Kenward, 2007). Dropout can differ among treatment groups without invalidating the analysis, if the dropout is not related to the outcome. This is illustrated in D) in Figure 1.1.

It is often difficult to determine whether any relationship between missing values and the unobserved outcome is absent (CHMP, 2010). It is therefore not always possible to know from the collected data whether the outcome would be biased.

The extent to which missing data bias results is influenced by many factors. These include the relationship between missingness, treatment assignment and outcome; the type of measure employed to quantify the treatment effect and the expected direction of changes over time for participants in the trial. The strategy employed to handle missing values might in itself provide a source of bias (CHMP, 2010).

Missing data lead to a loss in precision, even if the data are missing in such a way that conclusions are valid, since less data are available to use to make decisions (Carpenter & Kenward, 2007). Missing data imply a smaller sample size. The greater the number of missing values the greater the likely reduction in power (CHMP, 2010).

Non-completers may be more likely to have extreme values (either because a very good response or a very bad response might lead to loss to follow-up). The exclusion of the non-completers could lead to an underestimate of the variability in the data because participants who remain are often more similar and thus artificially narrow the confidence interval for estimation of the treatment effect (Carpenter & Kenward, 2007; CHMP, 2010). Missing data may thus potentially alter conclusions.

To draw proper inferences from the trial, dropouts should be included correctly in the analysis. The way in which dropouts, especially treatment related dropouts, are included in the analysis can influence the conclusions drawn from the trial dramatically (Gould, 1980). It is important to know whether these conclusions are sensitive to the missingness mechanism; as this determines how valid the conclusions of the trial are (Carpenter & Kenward, 2007).

The International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) E9 document states the following on missing data, Section 5.3:

> "*Missing values represent a potential source of bias in a clinical trial. Hence, every effort should be undertaken to fulfil all the requirements of the protocol concerning the collection and management of data. In reality, however, there will almost always be some missing data. A trial may be regarded as valid, none the less, provided the methods of dealing with missing values are sensible, and particularly if those methods are predefined in the protocol. Definition of methods may be refined by updating this aspect in the statistical analysis plan during the blind review. Unfortunately, no universally applicable methods of handling missing values can be recommended. An investigation should be made concerning the sensitivity of the results of analysis to the method of handling missing values, especially if the number of missing values is substantial*." (ICH E9 Expert Working Group, 1999)

Analysing missing data correctly versus just dropping it from the analysis increases efficiency and decreases bias (Horton & Kleinman, 2007). Increasingly the importance of considering missing data is being acknowledged (CHMP, 2010).

## 1.5 Intent to treat (ITT) analysis

The definition of an ITT analysis is that all participants are analysed as randomised, in the groups they were randomised to, whether or not these participants completed the study according to protocol or adhered to the treatment they were randomised to. The E9 guidelines (ICH E9 Expert Working Group, 1999) define the ITT population as "*The set of participants that is as close as possible to the ideal implied by the intention-to-treat principle. It is derived from the set of all randomised participants by minimal and justified elimination of participants.*" The ITT analysis has two principles. The first is that all participants randomised to a treatment arm are included in the analysis, even if they drop out, and the second is that participants are analysed according to the treatment they were randomised to, and not according to the treatment they actually received (Little & Yau, 1996). The principal advantage of the ITT analysis is that it avoids the selection bias introduced by the non-random losses of participants or when participants take a different treatment than the one assigned to them and thus preserves the randomisation, providing a basis for a valid inference (ICH E9 Expert Working Group, 1999; Little & Yau, 1996). This is stated by Fleming (2011) as "*In order to preserve the integrity of randomization, all patients should be followed until the complete capture of trial outcomes, even after patients have discontinued randomized treatment or initiated other interventions*." In order to do an ITT analysis as intended, complete follow-up of all participants is needed (ICH E9 Expert Working Group, 1999).

The ITT analysis estimates the benefits of a treatment policy relative to control and reflects the effectiveness of the treatment policy for the population. This observed estimate includes the effect of the treatment assigned and changes to this treatment over time (National Research Council,

2010). An ITT analysis that includes time off treatment in the analysis evaluates the entire intervention, including auxiliary care provided and rescue therapy offered. It is argued that this is most relevant in answering the question about the real life effect of the intervention should it become policy (Fleming, 2011). The ITT analysis asks one specific question, while a patient taking his medicine might want to know what the on-treatment efficacy of a treatment is, not just the ITT efficacy (Keene, 2011).

With no missing data the ITT analysis includes every participant who was randomised, regardless of adherence to the protocol, to estimate the effect of intending to give an intervention. Thus if participants are followed up after they cease to adhere to the intervention protocol, the data needed for an ITT analysis is available (Carpenter & Kenward, 2007).

Participants who are lost to follow-up create special problems in the ITT analysis, since potentially incomplete profiles would need to be analysed as randomised. The target estimand of the ITT analysis is the as randomised analysis of the hypothetical complete data, if outcomes were available for the participants who dropped out (Little & Yau, 1996). Failure to include these participants in the analysis may seriously undermine the ITT approach. The E9 guidelines mention that imputation techniques can be used to compensate for missing data. The E9 guidelines were written long ago before missing data theory was well developed. The guidelines state: "*Other methods employed to ensure the availability of measurements of primary variables for every subject in the full analysis set may require some assumptions about the subjects' outcomes or a simpler choice of outcome (for example, success/failure).*" (ICH E9 Expert Working Group, 1999).

The requirement for an ITT analysis implies that the data analyst needs a response from each randomised participant in order to include each randomised participant in the analysis. The ITT analysis intends to analyse the outcome for everyone who entered the trial, even if they do not have complete data. This interpretation of the ITT analysis makes missing data a particularly troublesome issue in clinical trials. The analyst cannot ignore the missing data if he or she wants to follow the requirement of an ITT analysis.

In practice, a simplistic imputation method, last observation carried forward (LOCF) is often seen as a solution to this problem. When implementing LOCF every missing value for a participant is replaced by the last observation observed before the missing value. However, several publications highlight the problems with LOCF as an analysis technique and this is not regarded as a feasible alternative in analysing missing data (Carpenter et al., 2004; Mallinckrodt, Clark, & David, 2001; Molenberghs & Kenward, 2007; Molenberghs et al., 2004). LOCF might inflate the Type I error and creates bias in the estimation of mean change from baseline while creating standard errors that

are too small (Mallinckrodt et al., 2001). LOCF has also been shown to force the missing data mechanism to be future dependent (Kenward & Molenberghs, 2009).

Missing data also violate the strict ITT principle which requires measurement of all participant outcomes regardless of protocol adherence (CHMP, 2010). With incomplete data there is no unequivocal ITT analysis (Carpenter & Kenward, 2007). Various analyses are possible, as discussed later.

## 1.6 Missing data and per protocol analysis

The per protocol population is a subset of the ITT population and includes those participants who were more compliant to the protocol; patients who completed a pre-specified minimal exposure to the treatment regimen (for example, received 80% of all doses), have the primary endpoint variable available and have no major protocol violations. The per protocol analysis maximises the opportunity for a new treatment to show efficacy in the analysis. It also most closely reflects the scientific model underlying the protocol. It is possible that a per protocol analysis could be biased, and this is seldom regarded as the primary analysis of a trial (ICH E9 Expert Working Group, 1999).

In a per protocol analysis non-adherers to the protocol are simply eliminated (Rothman, Greenland, & Lash, 2008). The per protocol analysis regards a participant's data as effectively missing from the time the participant ceased to adhere to the protocol. The per protocol analysis thus does not maintain the comparability of treatment groups that should have been guaranteed by randomisation, because it excludes randomly assigned participants on the basis of outcomes after baseline. The answer provided by the per protocol analysis is not applicable to the entire set of eligible patients and not interpretable in real life settings (Fleming, 2011). A per protocol analysis could be excluding participants because they do badly on the regimen, since participants doing badly on a regimen are more likely to discontinue the regimen and not completing the full treatment regimen, thus biasing the results (Keene, 2011).

An "as treated" analysis is defined as a subset of the per protocol analysis, where only participants who adhered to the intervention are included (Hulley et al., 2007). This analysis is based on the actual treatment received, rather than the treatment randomised to. This is subject to bias from factors that influence whether a participant in the study adheres to treatment. One such factor could be the treatment itself; since certain treatments might be easier or harder to adhere to than others (Rothman et al., 2008).

Because of the different hypotheses underlying per protocol and ITT analyses, it is not a surprise that there will be a difference in the way missing data are handled. There must be a difference between the conditional distributions of the missing data given the observed when addressing the ITT and per protocol hypotheses (Carpenter & Kenward, 2007).

## 1.7 What is the correct analysis in the face of missing data?

Before modelling the data it is important for the statistician to take the time, together with the investigators, to understand why observations might be missing, using the data (Carpenter & Kenward, 2007). Any data analysis should start with a critical discussion of the number, timing, pattern, reason for and implications of missing values. There is no universally applicable method of handling missing values and different approaches may lead to different results (CHMP, 2010).

The CHMP report emphasizes that when proposing a method to handle missing data it is important that an analysis is provided which does not have a bias favouring the experimental treatment, this, in their words, is a conservative analysis. This is the main justification they require for selecting a particular method. They do not consider the properties of the method under particular assumptions, only whether it provides a conservative estimate of the efficacy of the experimental treatment (CHMP, 2010).

This primary focus of the CHMP report is not helpful in many instances. The document is written as a guideline in drug development and drug licensing and in that context the requirement for a conservative approach is easy to implement. Drug manufacturers have a motivation to get new drugs on the market and it protects the public to require that drugs can only be shown superior in a conservative analysis. If effectiveness can be shown in the presence of bias against the drug, the licensing authority can be certain that the drug is even more effective than claimed by the drug manufacturer. A biased, conservative analysis might also lead to an effective drug not being licensed, which is also not in the best interest of the public. However, many other settings do not have such a clear experimental arm against which a conservative analysis should be applied. For example, in public health research, where governments are looking for the best possible treatment protocol to implement in state hospitals it is not clear whether a conservative bias would be more prudent. In cost effectiveness analysis of medications already registered, it is also not clear in which direction one wants conservative bias to be applied, since there is no clear ‘active’ arm that should not be advantaged by any bias.

Trying to perform a conservative analysis is not the solution to analysing missing data. In the first place, it can be argued that a conservative analysis might do as much harm as a liberal analysis by keeping an effective treatment from being adopted. Rather, one should apply an analysis that is

valid, neither conservative nor liberal. Whether an analysis is valid, depends on the assumptions one is willing to make regarding the missing data mechanism. Keene (2011) also argued that it is important to aim for the most relevant approach, the one that answers the question you are asking, not for the most conservative approach.

There is no universal statistical method to use when data are missing, but there are universal principles which apply to every situation. With missing data extra assumptions are required. These assumptions specify the mechanism by which the data became missing and highlights differences in the distribution of the data between participants who did and did not complete the trial. It is also necessary to show that the inferences about the intervention are robust to different assumptions about the reason for missing data (Carpenter & Kenward, 2007).

Principles for handling missing data highlighted by the panel on handling missing data in clinical trials (National Research Council, 2010) are:

1. Minimise the occurrence of missing data through proper screening of participants, good trial conduct and data collection efforts.
2. A missing value hides a true underlying value. This true underlying value is important for analysis.
3. Reasons for missing data should be documented, since these reasons can inform the assumptions about the missing data mechanism.
4. The missing data mechanism needs to be assumed and this assumption should be transparent.
5. A statistically valid analysis should be done under the missing data assumption. This analysis should account for sampling variability and uncertainty associated with missing observations.
6. A sensitivity analysis should be done to test the robustness of the analysis.

## 1.8 How much missing data are acceptable?

It is often assumed that the amount of missing data determines whether missing data are biasing results. However, there is no universal rule regarding the maximum number of missing values that could be acceptable. The nature of the outcome variable determines the absolute number of missing values. If the outcome variable is mortality, less missing data are expected than if the outcome is difficult to measure. The length of the trial also determines the amount of missing data. The longer the trial the more missing data are to be expected (CHMP, 2010).

It is not only the amount of missing data that determines whether missing data bias results, although more missing data certainly raises more suspicion of possible bias than few missing data

points. Rather it is determined by the question, information in the observed data and the reason for the missing data. Context determines whether error will be induced by the missing data. If an event is rare then missing data on few participants can alter the estimated event rates dramatically. If the proportion of participants withdrawing varies by intervention arm, estimated intervention effects are more likely to be affected than if participants withdraw independently of treatment arm. Even if no bias is introduced by missing data, missing data certainly leads to efficiency loss (Carpenter & Kenward, 2007). The anticipated effect size in the study and the likelihood that a sensitivity analysis would confirm the findings of the trial also influences the amount of missing data that would be deemed acceptable (National Research Council, 2010).

In Chapter 2 we discuss under what circumstances missing data can lead to bias (Carpenter & Kenward, 2007).

## 1.9 How missing data are currently handled

Burton and Altman (2004) did a review of the use of missing data methods in 100 cancer prognostic studies published in 2002. Of these 81 had missing data and 32 stated how the missing data were analysed; 12 used complete case approach, 23 available case, 6 omitted variables, 4 used a missing indicator approach, 3 used ad hoc imputation procedures and one used multiple imputation. They found that missing data were generally handled inadequately and included some guidelines for the handling of missing data in their paper. These included:

- Description of the completeness of data. The number of cases with missing data should be stated. The frequency of missingness should be given for every variable.
- Sufficient detail should be provided on the missing data analysis methods used.
- Any imputation method used should be referenced.
- For each analysis the number of cases included should be given.
- Missing data should also be explored by giving the known reasons for missingness and comparing the characteristics of cases with and without missing data.

A review by Wood et al. (2004) showed that the description of missing data and the methods used are inadequate. Horton and Switzer (2005) reviewed 311 papers published between 2004 and 2005 in the New England Journal of Medicine, only 26 (8%) reported missing data methods. This means there is a disjoint between the theoretical statistical methods available and the practical implementation of these methods in applied settings.

A trial should be designed in such a manner that variables are collected on all participants prior to withdrawal. These variables can be collected at baseline and during follow-up and can then be

used to predict or describe withdrawal (Carpenter & Kenward, 2007). The amount of missing data and the strategies selected to handle missing data can influence the required sample size, the estimate of treatment effect and the confidence with which data can ultimately be interpreted. How to minimise the amount of missing data is a critical issue that must be considered during the design of a trial (CHMP, 2010).

The National Research Council (2010) report on missing data discusses ways in which trials could be designed in order to minimise the impact of missing data. Data collection should be strengthened. They emphasise that trial outcome data should be collected for participants who stop the study treatment or had protocol non-compliance. Where possible, outcome data after withdrawal should be collected. The reason for dropout should also be collected (CHMP, 2010). They also suggest that outcomes be defined in a manner that is measurable in as many participants as possible; this includes the use of composite outcomes such as time to death or worsening of disease. They suggest that study duration be shorter in order to reduce missing data due to attrition. Attention should be given to how missing data due to death is to be handled. It is worth noting from their discussion that the handling of missing data is much larger than simply accounting for this in the statistical analysis. Researchers should be proactive in preventing missing data in the first place. It is also important to collect covariate data that is predictive of withdrawal and the study outcome, since this improves the adjustments made with incomplete data. Fleming (2011) also stated that the preferred approach to addressing missing data is to prevent it. He also argues that studies should distinguish between non-adherence (not receiving randomised therapy) and non-retention (not attending study visits).

The CHMP made the following points regarding the handling of missing data (CHMP, 2010), and these were reinforced by Carpenter and Kenward (2007):

1. One should take care in the design and implementation of a trial by minimising the amount of missing observations
2. One should consider how to cope with missing data when drawing up the data analysis plan
3. One should specify in advance the nature and scope of any sensitivity analysis
4. One should look at the proportion of missing data by time of withdrawal and treatment arm

The panel on handling missing data in clinical trials (National Research Council, 2010) summarises points on which guidelines for the handling of missing data in clinical trials agree. These are

1. One should plan how missing data will be handled during the design of the study.
2. Complete data should be collected for all randomised participants, even if they discontinue study treatment.

3. The CONSORT guidelines (Schulz et al., 2010) should be adhered to.
4. The use of single imputation methods is criticised.
5. Emphasis is placed on doing sensitivity analysis, since no single way of handling missing data exists.

The issues raised by missing data are wider than the technical details of statistical analysis (Carpenter & Kenward, 2007). There is no best approach for all situations. The acceptability of an approach depends on the assumptions made and whether it is reasonable to make these assumptions in the particular case (CHMP, 2010). Missing data provide a real challenge for the correct analysis of clinical trials. The National Research Council (2010) start their recommendations with: "*Missing data in clinical trials can seriously undermine the benefits provided by randomization into control and treatment groups*".

Mallinckrodt et al. (2003) summarised this as follows:

> "*When determining a suitable approach to modelling longitudinal data, it is important to realize that no single "best" method currently exists. This implies that an analysis must be individually tailored for a given situation. It is, therefore, crucial that the desired attributes of the analysis are clear, and that the characteristics of the missing and non-missing data are understood.*"

This thesis addresses the analysis of clinical trials when a considerable amount of data are incomplete. However, it should be remembered that the best way to deal with missing data is to prevent it from occurring in the first place by conducting the trial well, screening of participants and data collection. The National Research Council (2010) stressed the importance of continuing to collect data even when participants are not receiving the study treatment.

The rest of this thesis is structured as follows. In Chapter 2 we discuss different missing data mechanisms and introduce the MCAR, MAR and MNAR terminology of Rubin (1976). Monotone and non-monotone missing data are described and a taxonomy of missing data is given. In Chapter 3 different approaches to dealing with missing data are given. These include the complete case analysis, single and multiple imputation, selection models, pattern-mixture models, shared-parameter models, weighting methods and Bayesian methods. Likelihood-based methods are discussed in detail.

Chapter 4 highlights the importance of sensitivity analyses and describes the principles that should be followed when a sensitivity analysis is done. Chapter 5 introduces the first example data set, the SAPiT study. The study is explained and the data are analysed using the methods discussed in Chapter 3. In Chapter 6 the second example data set, the St John's Wort trial, is introduced and the

data are analysed using the methods discussed in Chapter 3. This re-analysis of the data provided a contribution to the literature around the effectiveness of St John's Wort because it changed the conclusions of the original analysis where missing data were not taken into account. Chapter 7 discusses the practical implications of missing data when analysing the results of a clinical trial.

# Chapter 2

# Types of missing data

In a clinical trial either baseline covariates, covariates over time or outcome data can be missing. Baseline covariates, which are often included in the statistical analysis, are seldom missing, because study staff go to great lengths to collect this data and some variables need to be collected at baseline in order to determine the participant's eligibility, for example age or disease status. The most likely baseline covariates to be missing are laboratory variables and sometimes interview data that participants do not want to divulge, such as salary. Missing date are often only created over time when participants do not attend all visits. Because baseline data are often collected with very little missing observations, the focus of the missing data discussion will be on missing outcome data and baseline covariates are assumed to be complete. Missing data in covariates can be handled conveniently using multiple imputation, which is discussed in Section 3.2.

## 2.1 Notation

We assume $N$ independent participants in a trial, indicated by $i = 1,\ldots,N$. We plan to collect a set of $n$ measurements $Y_{ij}$ with $j = 1,\ldots,n$. Participant is indicated by $i$ and measurement by $j$. Because of the possibility of missing data, we assume that we have $n_i$ observations for participant $i$.

$Y_{ij}$: Response for $i$th participant at $j$th occasion

$\boldsymbol{Y}_i = (y_{i1}, \ldots, y_{in_i})'$ : Vector of outcomes

$\boldsymbol{Y}_i^{obs}$: Observed component of $\boldsymbol{Y}_i$

$\boldsymbol{Y}_i^{mis}$: Missing component of $\boldsymbol{Y}_i$

$\boldsymbol{R}_i = (R_{i1}, \ldots, R_{in_i})'$: Vector of missing data indicators; $R_{ij} = 1$ if $Y_{ij}$ is observed and *0* otherwise

$D_i = 1 + \sum_{j=1}^{n_i} R_{ij}$: Dropout indicator. If the missing data indicator consists of all $R_{ij}$ equal to 1 up to time *j* and *0* thereafter, then the missing data can be represented by the dropout-time, the first time point where the data are missing.

$\boldsymbol{x}_i$: Vector of covariates for participant *i*

$t_{ij}$: The value of time (day, visit, week) for the $j^{th}$ measurement of participant *i*

$\boldsymbol{\theta}$: Parameter vector describing the measurement process; relating the outcomes $\boldsymbol{Y}$ to the covariates $\boldsymbol{x}$

$\boldsymbol{\psi}$: Parameter vector describing missingness process. $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ may have some parameters in common.

$z_i$: An indicator variable for treatment; *1* if the participant is in the active treatment arm and *0* if the participant is in the control arm. This is constant across time for a given participant

## 2.2 Missing data mechanisms

Rubin (1976) described three different missing data mechanisms. This taxonomy is based on the level of dependence between the missing data process and the measurement process. These are MCAR, MAR and MNAR. Most subsequent work has used these terms and has built on these concepts. Each is discussed in turn.

Independence across participants is assumed. The density of the full data is

$$f\,(\boldsymbol{y}_i, \boldsymbol{r}_i|\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{x}_i) = f(\boldsymbol{y}_i^{obs}, \boldsymbol{y}_i^{mis}, \boldsymbol{r}_i|\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{x}_i)\,.$$

This can be factored as

$$f\big(\boldsymbol{y}_i^{obs}, \boldsymbol{y}_i^{mis}, \boldsymbol{r}_i\big|\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{x}_i\big) = f\big(\boldsymbol{y}_i^{obs}, \boldsymbol{r}_i\big|\boldsymbol{\psi}, \boldsymbol{x}_i\big)\, f\big(\, \boldsymbol{y}_i^{mis}\big|\boldsymbol{y}_i^{obs}, \boldsymbol{r}_i, \boldsymbol{x}_i, \boldsymbol{\theta}, \boldsymbol{\psi}\,\big).$$

This factorisation makes it clear that the first factor on the right hand side can be inferred from observed data. By making modelling assumptions the analyst can estimate this distribution from the observed data. The second factor above is the distribution of missing data and cannot be inferred from the observed data. Thus, to analyse the full data model when data are incomplete, the analyst has to specify a model for the distribution of the observed data and to combine this with assumptions about the missing data. These assumptions cannot be tested using the observed data.

### 2.2.1 Missing completely at random (MCAR)

When missing data are MCAR the missingness is not related to any factor and is unrelated to any inference one wishes to draw. This means that the probability of dropout is independent of both observed and unobserved variables, including the outcome variable, given covariates.

This type of missing data arises when observations are missing because of equipment failure or because the staff member who collected the data was ill. These events are as likely to occur for one participant as for another, whatever their disease severity or intervention. It is as if after randomising the participants to intervention we randomly decide who to observe (Carpenter & Kenward, 2007). Participants who drop out of a study for this reason could be considered a random sample from the total study population and their characteristics and outcomes are similar to that of the study population (CHMP, 2010). MCAR data implies that

$$f\ (\boldsymbol{r}_i|\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{\psi}) = f(\boldsymbol{r}_i|\boldsymbol{x}_i, \boldsymbol{\psi})$$

(Horton & Kleinman, 2007; Molenberghs & Kenward, 2007).

When missing data are MCAR, an analysis of those participants who completed the study is valid. Important instances of non-likelihood methods such as generalised estimating equations (GEE) are also valid. Validity does not necessarily imply that such methods applied to the completers only are also efficient, and there usually are good reasons to include the incompletely observed study participants as well.

MCAR is the easiest type of missing data mechanism to deal with. However, in practice missingness is often related to the outcome of interest and thus the data are not MCAR. There is often an association between the probability of having missing data (either participant withdrawal or intermittent missing data) and either the intervention, baseline variables such as baseline disease status, or the measurement just prior to the missing observation (when data are measured longitudinally). If, for example sicker participants are more likely to have missing data, an analysis that assumes MCAR is not sensible. The convenient assumption of MCAR can thus be made in rare circumstances only. When the MCAR assumption cannot be made, one cannot analyse only those participants with complete data (Carpenter & Kenward, 2007).

In spite of what observed data may suggest, one can never be sure that data are MCAR. Nevertheless the observed outcome data can rule out MCAR. We can investigate whether there is any relationship between observed outcome data and the occurrence of missing data. If there is, data are not MCAR (Carpenter & Kenward, 2007). One can test whether data are MCAR against the assumption of MAR (Diggle, 2002). In the simple setting this test reduces to a t-test comparing the means of outcome variables for complete and incomplete cases. Logistic regression can be used to identify predictors of dropout or missingness (Carpenter, Pocock, & Lamm, 2002).

### 2.2.2 Missing at random (MAR)

MAR is a less stringent requirement than MCAR. Missing data are said to follow a MAR mechanism if the probability of non-response can depend on observed data (responses and baseline covariates) but conditionally on these does not depend on unobserved responses. This assumption implies that the behaviour of the post dropout observations can be predicted from the observed variables and therefore the response can be estimated without bias using the observed data exclusively (CHMP, 2010).

The term MAR might be misleading, since missingness is not random and can actually be predicted from recorded information, although in a stochastic rather than a deterministic sense. Missingness is, however, random after controlling for the observed covariates (Horton & Kleinman, 2007). Another way of looking at this is to say that data are MAR when we can find fully observed variables which define groups within which the data are MCAR. Data are MAR, if, conditional on another variable, the data are MCAR (Carpenter & Kenward, 2007).

In the case of MAR data, the statistical distribution of potentially missing data is the same (conditionally) for all participants who share the same observed data, whether or not they have missing data. Participants who have missing data share the same conditional statistical behaviour in their unobserved future, given their observed past, as those who do not have missing data. In other words, the distribution of the endpoint value, given baseline and observed data, for participants with missing values is the same as for participants without missing data (Carpenter & Kenward, 2007).

MAR data occur, for example, when a participant was doing poorly and the physician decided to discontinue participation. In such a case dropout was related to the outcome of interest, but the observed data explained the dropout (Mallinckrodt, Sanger, et al., 2003). The MAR assumption is implausible if missing data are due to some unobserved deterioration, but plausible if missing data are simply due to loss to follow-up. The MAR assumption is more plausible than the MCAR assumption (Little & Rubin, 1987; Mallinckrodt, Clark, et al., 2003).

Mathematically MAR data are defined as follows:

$$\boldsymbol{f}\left(\boldsymbol{r_i}|\boldsymbol{x_i}, \boldsymbol{y_i}, \boldsymbol{\psi}\right) = \boldsymbol{f}(\boldsymbol{r_i}|\boldsymbol{x_i}, \boldsymbol{y_i^{obs}}, \boldsymbol{\psi})$$

(Molenberghs & Kenward, 2007).

Under MAR assumptions analysis of completers only or observed data only is not valid. The marginal average cannot be used. Imagine that worse health at baseline is associated both with

increased risk of withdrawal and poor response to intervention. In this case, analysing data from the participants who remain to the end of the trial gives an over optimistic view of the intervention effect (Carpenter & Kenward, 2007). Under MAR assumptions unweighted GEE is not valid, due to its non-likelihood nature. Likelihood-based analyses, some weighted estimation methods, multiple imputation and Bayesian methods are valid, among others (Molenberghs & Kenward, 2007).

MAR data can be analysed through joint modelling of complete and partially observed response data, conditional on fully observed data. With current methods available to analyse MAR data, single or simple imputation of data is not needed. If outcome data are MCAR conditional on treatment alone (i.e. MAR) then including treatment in the model gives a valid treatment estimate (Carpenter & Kenward, 2007).

It is argued that a MAR analysis addresses the per protocol hypothesis under which we want to estimate the distribution of responses had participants continued to adhere to the protocol. If we assume the data are MAR, sensible estimates of the intervention effect addresses this hypothesis. Within each baseline group, for example each treatment group, the distribution of responses is the same for observed and unobserved participants. This assumes that the unobserved participants continued with treatment per protocol as the observed participants did. Under MAR assumptions the estimated treatment effect is thus the per protocol treatment effect (Carpenter & Kenward, 2007). If no further data are collected when participants are withdrawn from treatment, methods that rely on MAR are estimating treatment effect under the condition that everyone remained on treatment. This does not provide a valid estimator of the ITT effect. If data are collected after withdrawal from treatment, this can be used to estimate the ITT effect, assuming MAR (National Research Council, 2010).

The process to analyse data where the missing data mechanism is believed to be MAR can be as follows:

1. Identify a fully observed variable or set of variables whose values predict the occurrence of missing data
2. Within groups identified by this variable or set of variables, assume data are MCAR. Within these groups sensible estimates can thus be obtained from the observed data
3. An overall average estimate is obtained from these separate groups, by averaging the groups and allowing for there being different numbers of participants in the different groups (Carpenter & Kenward, 2007).

Hogan et al (2004) extended the MAR concept to longitudinal data. They coined the phrase sequential MAR (or S-MAR). The S-MAR assumption is that given $R_{i,t-1} = 1$ the future responses $(Y_{it}, \ldots, Y_{in})'$ are independent of $R_{it}$ conditional on the past. The difference between the MAR and S-MAR assumption is that under MAR, dropout at $R_{it}$ can depend on elements of the covariates (*x* variables) and observed responses (*Y* variables) before, at and after *t*, while under S-MAR dropout at time *t* can depend on elements of the covariates and the observed *Y* before *t* or in the case of the covariates, at *t*.

By including a rich set of predictors, the MAR assumption can be made more plausible (Collins, Schafer, & Kam, 2001). If all the predictors of dropout are known and used in a model, then a MAR model is adequate. In the design of a study it is thus important to identify the likely causes for dropout and collecting variables that would measure this. Baseline characteristics and response variables over time could be useful. A study can also be designed so that withdrawal is triggered by a response variable deteriorating beyond a specific level (Carpenter et al., 2002). This is the case in the Hypericum perforatum study discussed in Chapter 6 (Hypericum Depression Trial Study Group, 2002).

### 2.2.3 Missing not at random (MNAR)

MNAR has several synonyms and is also called not missing at random (NMAR). Data are MNAR if the probability of an observation being missing depends on unobserved measurements, conditional on the observed data. Even given the information about the missingness mechanism in the fully observed data, the reason for an observation being missing depends on the unseen value of that observation. The distribution of future observations conditional on past observations differs between those who have missing data and those who do not (National Research Council, 2010). MNAR is often defined by exclusion; when neither MCAR nor MAR assumptions are valid, data are MNAR (Carpenter & Kenward, 2007). Missingness can depend on unobserved outcomes in addition to dependency on observed covariates and outcomes (Molenberghs & Kenward, 2007).

MNAR data in a clinical trial could arise, for example, when a participant had been doing well until midway in a trial and was then lost to follow-up, because, after the last observed visit the participant relapsed into a worsened condition. In such an instance dropout was related to the outcome of interest, but the observed data did not predict the dropout. The unobserved data held information not foreseen by the observed data. Missingness due to adverse events are difficult to classify as MNAR versus MAR because the relationship to the observed outcome may vary from situation to situation. Missingness due to adverse events are not MNAR, if it was observed that the

participant's condition worsened prior to the dropout but would be MNAR if the participant's condition did not worsen prior to the dropout (Mallinckrodt, Sanger, et al., 2003).

Future observations cannot be predicted without bias by the model. It is impossible to be certain whether there is a relationship between missing values and the observed outcome variable or to judge whether the missing data can be adequately predicted from the observed data (CHMP, 2010). Ignoring non-random missing data when it is present, means choosing not to model the relationship between the unobserved response and dropout. This means ignoring the fact that dropout indicates some deviation in the response. This effect is then ignored in the estimates of parameters (Carpenter et al., 2002).

Valid inference under MNAR requires explicit or implicit use of the missing value mechanism. In practice, we often do not know what it is. Because MNAR methods require assumptions that cannot be validated from the data at hand, a definitive MNAR analysis does not exist (Carpenter & Kenward, 2007; Mallinckrodt, Sanger, et al., 2003). Any analysis of data in the presence of MNAR data is assumption driven. The panel on handling missing data in clinical trials (National Research Council, 2010) called this the "*central problem of missing data analysis in clinical trials*". Since no definitive analysis exists, sensitivity analysis plays a large role in the analysis of missing data assumed to be MNAR (Mallinckrodt, Clark, et al., 2003; Molenberghs & Kenward, 2007).

The observed data can be used to distinguish between MCAR and MAR, but the observed data cannot be used to distinguish between MNAR and MAR (Carpenter et al., 2002; Horton & Kleinman, 2007). This means that one can never rule out MNAR, because in order to do so one needs to observe the missing measurements. The MNAR models in a sensitivity analysis need to be selected to generate conclusions bounded by the results of the MAR model and the worst case MNAR model. The sensitivity of conclusions to non-random dropout can be assessed by modifying the MAR model to allow for various non-random dropout scenarios and seeing whether conclusions vary (Carpenter et al., 2002). An overall assessment of MAR versus MNAR is not possible, because every MNAR model has a unique MAR counterpart, with the same fit as the original MNAR model. This MAR model would also lead to the same predictions of the observed data as the MNAR model (Molenberghs et al., 2008).

Most analytic approaches under MNAR are based on models for the joint distribution of the outcome variable and the missing data mechanism. Classifications of different MNAR models are based on the factorisation of these joint models. These missing data models fit in the broader context of joint models for repeated measures and time to event (Hogan et al., 2004).

The analysis of MNAR data proceeds in two steps. During the first step the statistical relationship between the chance of seeing a variable and its unseen value is described. During the second step we describe how the distribution of the data differs among participants with missing observations (Carpenter & Kenward, 2007).

The only way forward in the face of MNAR data is to use auxiliary information and knowledge about possible data mechanisms to describe distributions for the missing data or to conduct sensitivity analyses to gauge whether conclusions will change when varying assumptions are made about the missing data mechanism while assuming MNAR.

### 2.2.4 Summary of MCAR, MAR and MNAR

It is implausible that clinical trial data would be MCAR. MNAR analyses are difficult to implement and interpret and it is difficult to know whether the assumptions made are correct. MAR likelihood-based methods seem to be most suited to longitudinal clinical trials data especially if covariate information predictive of missingness is collected and conditional analyses can be done. However, this depends on the estimand and the MAR analysis generally answers the per protocol question and not the ITT question. Since it is difficult to rule out the possibility of MNAR data, it is suggested that when an MAR analysis is done, MNAR analysis is used to assess robustness of the results from the likelihood-based MAR analysis (Mallinckrodt, Sanger, et al., 2003; Molenberghs & Kenward, 2007). With missing data it is important to remember that no definitive correct analysis exists.

Rubin (1987) argued that, even if data are MNAR, after accounting for the information about the missingness mechanism in the observed data, there is relatively little information remaining in the unseen data. This is a further argument for the use of MAR methods.

## 2.3 Monotone and non-monotone missing data

The pattern of missing data can be classified as monotone or non-monotone missing data. If the data matrix can be rearranged so that there is a hierarchy of missingness so that observing a particular variable $Y_t$ for a participant implies that $Y_{t-1}$ is observed then missingness is said to be monotone. Simple methods can be used if the pattern is monotone (Horton & Kleinman, 2007). In longitudinal data monotone missing data are created when missingness is caused by dropout (Molenberghs & Kenward, 2007). When missingness is caused by participants randomly missing some visits and arriving for other visits missingness is non-monotone. When missingness is non-

monotone, models for the missingness of one variable may include covariates or previous outcome data which also have missing values.

Some of the methods discussed in the chapters that follow can only be applied when missing data are monotone, for example weighted GEE and pattern-mixture models under identifying restrictions. This will be discussed in more detail where these methods are described. It is also important to know whether missingness is monotone or non-monotone when choosing the multiple imputation method.

## 2.4 A taxonomy of missing data

To appropriately describe missing data issues, it is convenient to adopt a number of interconnected but logically independent taxonomic dimensions. Several have been touched upon already in what preceded.

**Missing data mechanism**

The missing data mechanism refers to whether data are MCAR, MAR or MNAR, in other words it refers to why data are missing. These were defined and discussed in detail in Section 2.2

**Missing data pattern**

This refers to whether data are complete or missing and whether missing data are monotone or non-monotone. This has been discussed in Section 2.3.

**Missing data frameworks**

Will the data be analysed using a pattern-mixture model, a selection model, or employing a shared-parameter model? A selection model factors the joint distribution of the measurement and response mechanism into the marginal measurement distribution and the distribution of the missing data, conditional on the outcomes. A pattern-mixture model does the reverse (Kenward & Molenberghs, 2009). These are discussed in Chapter 3.

**Inferential paradigm**

Three commonly used paradigms exist for statistical inference, namely likelihood, Bayesian and frequentist inference.

**Ignorability**

Rubin (1976) defined the concept of ignorability in missing data. This refers to the fact that under certain conditions the mechanism generating missing data can be ignored when the interest is in inference about the measurement process without biasing the analysis.

In order for missing data to be ignorable in the likelihood setting, the assumption is made that the parameters of the missing data mechanism are distinct from the parameters of the sampling model, the separability condition (Ibrahim et al., 2005). When the separability condition holds in a likelihood framework or under Bayesian inference, ignorability is equivalent to the union of MAR and MCAR. Frequentist methods such as GEE are generally unbiased only under MCAR, even though there are some exceptions (such as restricted maximum likelihood (REML), certain forms of profile likelihood, protective estimation methods, etc.). However, if GEE are weighted using weights that depend on the missingness probability, GEE can be used under MAR. This is what is called inverse probability weighting (Carpenter & Kenward, 2007). When ignorability holds, one need not specify the missing data mechanism, but only the full data response model (Daniels & Hogan, 2008).

**Study population**

Study population defines the study population to which the analysis refers. These include the ITT, per protocol or "as treated" populations. Some implications of missing data for the ITT population were discussed in Section 1.4.

### 2.4.1 Summary

Several approaches to analysing missing data are possible by making selections on the taxonomy dimensions discussed in the previous section. In the case of continuous data, if missingness is non-monotone, but is dominated by dropout, the question translates to a selection model question. This implies we are interested in the overall treatment effect change over time, not specific to a dropout group. This analysis would make use of all the data; both from complete and incomplete sequences.

We return to the conundrum mentioned in Section 1.4: clinical trials generally require an analysis by ITT, which is complicated by missing data. Some methods would be valid if the MAR assumption can be made, and the missing data mechanism is ignorable, such as direct likelihood, direct Bayesian inference, multiple imputation and inverse probability weighting methods (such as weighted estimating equations). However, these methods address the per protocol hypothesis and not the ITT hypothesis.

In Chapter 3 various methods to deal with missing data are discussed. Most of these methods can be implemented in standard software.

## 2.5 The estimand

The choice of an estimand involves the outcome measure, the population of interest and the time over which the analysis is conducted (National Research Council, 2010). The estimand is that quantitative feature which one would like to know from the population, but is only available through the inferences drawn from a selected sample from this population. It is important that the estimand be clearly defined, by specifying the outcome measure and the target population of interest clearly.

The full analysis set is all participants who met the inclusion and exclusion criteria and were randomised. Ideally, inference for the target population of eligible participants would be drawn from the full analysis set (Carpenter & Kenward, 2013).

The estimand one chooses is influenced by whether one wants to determine the efficacy or the effectiveness of the intervention. The efficacy is defined as the effects of the intervention if taken as specified in the protocol, in other words the benefit of the drug expected at the endpoint of the trial if all participants adhered to the treatment. Effectiveness, on the other hand, is the effects of the treatment observed in the trial, given imperfect adherence and other deviations from the protocol (Mallinckrodt et al., 2012).

(Carpenter, Roger, & Kenward, 2013) preferred to introduce the less ambiguous terms ‘de jure’ and ‘de facto’. In essence, the de jure hypothesis is the efficacy hypothesis. The de jure question asks what the expected treatment effect would be in the target population if the treatment and control were taken as specified in the protocol. This is also what is sometimes meant under the per protocol effect. Within the context of missing data, this question becomes “If after participants were not observed, they continued with the treatment randomised to and adhered to the protocol, what would the treatment effect be at the end of the study?”

The de facto question refers to the effectiveness question, namely what effect would be seen in practice if this treatment was assigned to the population of eligible patients. This is what is sometimes meant under the ITT question. In the context of missing data, the de facto question is “If we had observed all the participants at the end of the trial, what would the treatment effect have been?” (Carpenter et al., 2013).

The following five estimands were listed by the National Research Council (2010) in the context of a symptom relief trial with two treatment arms.

**Estimand 1: Differences in outcome improvement at the planned endpoint for all randomised participants.**

This estimand compares mean outcomes for participants randomised to the treatment and control arms, regardless of what treatment they actually received; the traditional ITT framework. Data after participants stopped using the randomised medication or started using rescue medication are included in this analysis. This answers the de facto question, or the effectiveness hypothesis regarding treatment policies relative to control. The observed difference between the treatment and control includes the effect of the treatment randomised to as well as any changes made to the participants' treatment over time, either due to lack of effect or side-effects. During the drug development process causal effects are usually the focus, not treatment policies. A parallel-group randomised trial in which outcome data are collected on all participants, those who adhere to the protocol and treatment regimen and those who do not adhere, supports the use of this estimand. If outcome data are not collected after participants drop out or switch from the assigned treatment, this estimand is not supported (Mallinckrodt et al., 2012; National Research Council, 2010).

**Estimand 2: Difference in outcome improvement in tolerators.**

This estimand compares the mean outcomes for treatment versus control in the subset of participants who tolerated and adhered to the treatment. A design that supports this estimand is a study with an active run-in phase used to identify participants who meet tolerability and efficacy criteria to continue. After the run-in phase, participants are randomised to one of the two treatment arms. Drug benefit is assessed only in participants who were selected because they responded favourably during the run-in phase (Mallinckrodt et al., 2012; National Research Council, 2010).

**Estimand 3: Difference in outcome improvement if all participants tolerated or adhered.**

This estimand assesses the outcome if all participants adhered to the treatment regimen for the study duration and did not drop out of the study. It addresses the de jure (efficacy) hypotheses about the causal effects of the treatment randomised to. Although this is the question one would like to answer in a clinical trial, it is unlikely that all participants will adhere. In real trials when using Estimand 3, one should asses the degree of non-adherence and how this may influence the treatment benefit observed. When data are missing, this estimand requires imputation of what would have been the outcome if individuals who did not comply had complied (Mallinckrodt et al., 2012; National Research Council, 2010).

**Estimand 4: Difference in areas under the outcome curve during adherence to treatment**

This estimand compares the means of the area under the curve over the duration of protocol adherence between the two arms. This estimand simultaneously quantifies the effect of treatment

on both the outcome measure and the duration of tolerability or adherence in all participants. A parallel group randomised trial supports the use of this estimand (National Research Council, 2010).

**Estimand 5: Difference in outcome improvement during adherence to treatment**

This estimand is the difference in mean outcomes as long as participants adhere to the protocol and treatment. This estimand includes both the duration of tolerability or adherence and outcome improvement in all participants. A parallel-group randomised trial supports the use of this estimand. This estimand assesses the de facto hypothesis regarding the randomised treatment. Assessing effects only while the drug is taken overestimates effectiveness at the endpoint, if the drug is not long acting or does not alter the disease permanently (Mallinckrodt et al., 2012; National Research Council, 2010).

Mallinckrodt et al. (2012) also defined a sixth estimand.

**Estimand 6: Difference in outcome improvement in all randomised participants at the planned endpoint of the trial attributable to the initially randomised medication.**

This estimand assesses effectiveness or the de facto hypothesis and requires data after withdrawal of the randomised study medication until the planned endpoint of the study. This estimand needs to be free from the confounding effects of rescue medication, because inference is to be made on the study medication and not the treatment regimen. While Estimand 3 addresses the de jure (efficacy) hypothesis, this estimand addresses the de facto (effectiveness) hypothesis. Both make causal inference regarding the initially randomised treatment in all participants at the planned endpoint of the trial.

Many more estimands are possible in specific situations and different trials.

# Chapter 3

# Approaches to dealing with missing data

When data are incomplete all analyses make assumptions in addition to the assumptions already made with complete data. Although simple analyses are always appealing, simple methods do not necessarily make plausible assumptions (Molenberghs & Kenward, 2007). With high rates of missingness results may be problematic to interpret regardless of the analytic methodology used (Mallinckrodt, Sanger, et al., 2003). In order to interpret the results correctly one needs to be clear about the assumptions made in the data analysis. In the discussion of various approaches to analysing missing data that follows, the assumptions underlying each of the approaches are discussed.

The missing data process could either be of direct interest in the modelling of the data or merely a nuisance that needs to be accounted for. The research question determines whether the missing data process is of interest or merely a nuisance. Dropout, or missing data for other reasons, can reflect a clinically meaningful response to treatment which one might want to take into account when making inference about the treatment.

A critical discussion of the number, timing, pattern, reason for and possible implications of missing values should be included in the final report. Graphical summaries, like Kaplan-Meier plots, of the dropout patterns should be provided so that it can be seen whether there is a differential dropout pattern between treatment groups. Whether participants with and without missing values have different characteristics at baseline can contain important information. The final report should have

a discussion on whether the pre-defined analysis is still sensible, with additional sensitivity analyses (CHMP, 2010).

A systematic approach for analysing incomplete data could follow the following steps.

1. Consider the reason, or mechanism, which caused the data to be missing.

Take account of all the information about the missingness mechanism in the observed data. Determine the probability distribution of the missing data, since this informs the analysis. Consider how the reason for missing a visit depended on the previous visits and the baseline data. Next consider how the missingness mechanism depends on unseen measurements. This affects the probability distribution of the missing data and the estimated intervention effect at the final visit (Carpenter & Kenward, 2007).

2. Some assumptions need to be made.

A missingness mechanism and the distribution of the missing data given observed data need to be assumed. Possible distributions of the missing data given the observed data can be identified. Focus on whether the distribution of participants' unseen observations at later visits, given their observations at previous visits and baseline, is different from that seen among the participants who have no missing data. The available data do not have any information about these distributions. The validity of the assumption about the distribution of the missing data cannot be verified. It is important to see how different assumptions made about the missing data distribution change the conclusions drawn. The final conclusions are valid if the assumptions made about the missing data mechanism and the distribution of missing data was correct. If these assumptions are wrong, the model is wrong and the conclusions drawn using this model are also wrong.

Various methods have been proposed throughout the years for dealing with missing data. In a review paper Horton and Kleinman (2007) list various approaches of dealing with missing data. They do not specifically refer to the longitudinal case. They list the following: complete case method, ad hoc methods like LOCF, multiple imputation, likelihood-based approaches, weighting methods and Bayesian methods. We do not discuss ad hoc methods in the rest of this thesis.

## 3.1 Complete case analysis

In a complete case analysis only those participants with complete data are analysed. All participants who withdrew or have missing data are excluded. In a clinical trial context this implies an analysis of completers (Horton & Kleinman, 2007; Molenberghs & Kenward, 2007). The advantage of this approach is that it is easy to implement and easy to describe. Although it seems appealing, this approach is inefficient, because not all collected data are used, and has the potential

to be biased if the missing data mechanism is MAR or MNAR (Horton & Kleinman, 2007; Molenberghs & Kenward, 2007).

When the true missingness mechanism is MCAR this analysis is unbiased and sensible, even if inefficient (Carpenter & Kenward, 2007; Molenberghs & Kenward, 2007). A complete case analysis violates the ITT principle, since it does not include all randomised participants, but only includes those who have completed the study and have no missing data (CHMP, 2010). Any analysis that violates the ITT principle is open to bias relative to the ITT question, since randomisation is not preserved and the comparability of randomised groups, guaranteed by randomisation, is lost. Complete case analysis is done on a subset of participants, those who completed, and a different type of participant may be retained in the different treatment arms, for example participants not benefiting could drop out of the placebo arm, while participants experiencing side effects could drop out of the active arm. When an ITT question is not the focus of the analysis, such as when an as treated or per protocol analysis is done, the above mentioned bias is less important. There are distortions in the mean structure, the variance structure and the correlation structure when a complete case analysis is done (Molenberghs & Kenward, 2007). The complete case analysis requires the assumption that the complete cases are a random subsample of all cases (Little, 1993).

The bias introduced by complete case analysis depends on the degree of deviation from MCAR, how much data are missing and the specific analysis. Thus, the potential for bias increases if more data are missing (Little & Rubin, 2002).

The European Medicines Agency (EMEA) suggested that complete case analysis be used in confirmatory trials as a secondary supportive analysis (sensitivity analysis) to illustrate the robustness of conclusions (CHMP, 2010). However the National Research Council (2010) stated that these analyses do not have a place in regulatory submissions. In practice, it is often incorrectly used as an easy starting point in any analysis, before more complicated models are fitted.

In a simulation example Ibrahim et al. (2005) showed that if the missing data mechanism depended on the outcome a complete case analysis was inefficient and outperformed by maximum likelihood, multiple imputation, Bayesian analysis and weighted estimating equation estimates, whether or not the distribution of the missing data was correctly specified in these methods. It seems plausible to use the likelihood approach as the easiest approach to handle the missing data problem. Technically it is a strong and robust approach.

## 3.2 Single imputation and multiple imputation

We discussed the inefficient and potentially incorrect way of analysing missing data by throwing away the missing data and ignoring missing data and the missing data mechanism in the analysis of data, the complete case analysis. The other extreme is to replace the missing data by creating data. This is called imputation.

Longitudinal imputation is done by using data from the same participant at different times, often without using data from other participants. This is in contrast to cross-sectional imputation methods where data at a particular time point is used to impute missing data at that time point. Different single imputation techniques used with longitudinal data include interpolation (if a value at t = 2 is missing, this value is imputed as the value between t = 1 and t = 3), LOCF, ratio imputation and regression imputation. Longitudinal regression imputation is done by fitting a linear or other regression model for each participant with a missing value. This model includes the outcome variable as dependent variable and time and other covariates as independent variables. The predicted value of the outcome for the time point with a missing value is then imputed. It is believed that these longitudinal imputation methods use more data and provide better imputations than cross-sectional methods (Twisk, 2013).

There are various problems with single imputation of data. In the analysis of observed data we allow for the fact that measurements are made with error. To assume that if data are missing we can impute the missing value without error (single value) is illogical. Another problem is that with conditional mean imputation, the imputed data are much less variable than the observed data would have been. Thus analysing the imputed data as observed data leads to an underestimation of standard errors and p-values. The confidence intervals for treatment effects calculated with imputed data are thus too narrow and artificially create an impression of precision (Carpenter & Kenward, 2007; CHMP, 2010).

Multiple imputation, first suggested by Rubin (1987) overcomes the limitations of single imputation. It involves producing several imputed data sets, each with different imputed values from the posterior predictive distribution of the missing data given the observed data. The analyst applies conventional estimation methods to each of these imputed data sets. In multiple imputation an extra step is added after data were imputed. This additional step is needed to correctly estimate the variability of quantities estimated from a completed data set. Parameter estimates are averaged across the analyses of the imputed data sets. Standard errors are calculated using Rubin's (1987) formula. Multiple imputation provides an approach for accounting for the variability of the estimated distribution of the missing data given the observed data. Multiple imputation does not

treat any one set of imputations as the true unobserved values of the missing data (Carpenter & Kenward, 2007).

Multiple imputation is done in three steps to estimate incomplete data regression models.

Step 1. Plausible values for missing observations are created that reflect uncertainty about the missing data models. These values are used to fill in or impute the missing values, generally assuming the missing data process is MAR. This process is repeated, resulting in the creation of a number of, say *m*, completed data sets. Taking into account the uncertainty in estimating both the relationship between the variables and the residual variability several complete data sets are imputed. These provide a representation of the distribution of the missing data given the observed.

Step 2. Each of these data sets is analysed using complete data methods. The data analysis method that would have been appropriate had there been no missing data is used for this analysis.

Step 3. The results are combined, which allows the uncertainty regarding the imputations to be taken into account. Typically five to ten imputations are created. These results are unbiased and have approximately the correct standard error (Horton & Kleinman, 2007; Molenberghs & Kenward, 2007).

In short, imputation processes similar to stochastic regression are run on the same data set multiple times. Each imputed data set is analysed separately and the results are averaged except for the standard error term. The standard error is constructed by the within imputation variance of each data set as well as the variance between imputed items on each data set. These two variances are added together and their square root determines the standard error. The noise due to residual variation, as well as the additional noise due to imputation, is introduced to the regression model.

The repeated imputations in Step 1 are draws from the posterior predictive distribution of the missing values under a specific Bayesian model for both the data and the missing data mechanism (Rubin, 1996). Multiple imputation is at heart a Bayesian procedure, but Rubin (1987) provides technical conditions under which multiple imputation can be interpreted validly under the frequentist paradigm.

For each of the *m* imputations a point estimate, $\widehat{Q}_i$, is computed for the parameter ($Q$) of interest, for $i = 1,\ldots,m$. The combined point estimate of this parameter ($Q$) for multiple imputations is the average of the *m* estimates; each calculated using a complete (imputed) data set. In a similar manner a within-imputation variance, $\overline{U}$, can be calculated as the average of the *m* variances from each of the imputed complete data sets. A between imputation variance is also calculated as

$$B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{Q}_i - \bar{Q})^2.$$

The variance estimate of $\bar{Q}$ is

$$T = \bar{U} + \left(1 + \frac{1}{m}\right)B,$$

the sum of the within-imputation component and the between-imputation component (Rubin, 1987). The imputation model can include baseline and other variables which are not included in the eventual trial analysis (Ibrahim et al., 2005). Including more variables in the imputation model can improve the accuracy of the imputation (Carpenter & Kenward, 2007).

For a single parameter, the ratio of the between-imputation component of variance to the within-imputation component, defined as

$$R = \frac{(1 + m^{-1})B}{\bar{U}}$$

gives the relative increase in variance due to the missing data. This indicates how the missing data increase the uncertainty of the estimates. One can view this quantity as the cost due to missing data (Rubin, 1987). Interval estimates and significance levels can be obtained using a *t* distribution with center $\bar{Q}$ and scale $T^{1/2}$ and degrees of freedom

$$v = (m - 1)(1 + r^{-1})^2.$$

The two-sided p-value for the null hypothesis

$$H_0: Q = 0$$

is computed by comparing

$$\frac{\bar{Q}}{T^{\frac{1}{2}}}$$

with a t-distribution with *v* degrees of freedom (Little & Yau, 1996).

If the missingness pattern is monotone, parametric regression methods assuming multivariate normality or nonparametric methods that uses propensity scores are appropriate (Rubin, 1987). If the missingness pattern is not monotone a Markov Chain Monte Carlo (MCMC) method is used (Schafer, 1997). Instead of imputing all missing values using the MCMC method, just enough missing values can also be imputed to make the imputed data sets monotone and methods appropriate for monotone data sets are then used on these imputed monotone data sets.

In the parametric regression method, a regression model is fitted for each variable with missing values, with the previous variables as covariates. Based on the fitted regression coefficients, a new regression model is simulated from the posterior predictive distribution of the parameters and is used to impute the missing values. For variable $Y_{ij}$ with missing values a model

$$Y_{ij} = \beta_0 + \beta_1 Y_{i1} + \beta_2 Y_{i2} + \ldots + \beta_{j-1} Y_{i,j-1} + \varepsilon_{ij}$$

is fitted using observations with observed values for $Y_{i1}$ to $Y_{i,j-1}$.

This model included the regression parameter estimates

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{j-1})'$$

And the covariance matrix

$$\hat{\sigma}_j^2 V_j$$

where $V_j$ is the matrix derived from the intercept and observed variables. For each imputation, new parameters

$$\beta_* = (\beta_{*0}, \beta_{*1}, \ldots, \beta_{*(j-1)})' \text{ and } \sigma_{*j}^2$$

are drawn or simulated from the posterior predictive distribution of the parameters. The missing value is then replaced by the value calculated using the parameters and observed or simulated previous observations of the outcome variable and a simulated error term (Rubin, 1987; SAS Institute Inc., 2004).

The propensity score method generates a propensity score for each missing variable to estimate the probability that the observation is missing. The observations are grouped based on these propensity scores and an approximate Bayesian bootstrap imputation is applied to each group (Rubin, 1987).

Under the MCMC method one generate multiple imputations using MCMC sampling with a single chain that is long enough for the distribution to reach a stationary distribution. The initial estimates for the algorithm are obtained using the expectation maximisation (EM) algorithm and the initial estimates for the EM algorithm are generated using a complete case analysis (Molenberghs & Verbeke, 2005). Routinely, a multivariate normal distribution is assumed for continuous data. A general contingency table model is assumed for categorical data. Schafer (1997) considered a specific distribution for a mixture of continuous and discrete outcomes. Through MCMC one can simulate the Bayesian joint posterior distribution of the unknown quantities and obtain estimates of the posterior parameters. During the imputation step the missing values for each observation is simulated independently. The imputation step draws values for $\boldsymbol{Y}_i^{mis}$ from a conditional distribution for $\boldsymbol{Y}_i^{mis}|\boldsymbol{Y}_i^{obs}$. Given a complete sample the posterior step simulates the posterior population mean vector and covariance matrix. These new estimates are then used in the next imputation step. The two steps are iterated long enough for the results to be stable, thus creating a Markov chain which converges in distribution to $p(\boldsymbol{y}_i^{mis}, \theta|\boldsymbol{y}_i^{obs})$. This simulates independent draws of the missing values from the posterior distribution. Various prior distributions can be

used, either non-informative priors when no prior information exist, or informative priors where appropriate.

Multiple imputation can also be done through multivariate imputation by chained equations (MICE), also known as fully conditional specification and sequential regression multivariate imputation. MICE assumes the missing data are MAR given the variables used in the imputation process. Many multiple imputation procedures assume a large joint model for all the variables, for example a joint normal distribution. In the MICE procedure a series of regression models are run where each variable with missing data is modelled conditional on the other variables in the imputer's model. Each variable can be modelled according to its own distribution. Binary variables can be modelled using logistic regression and continuous variables can be modelled using linear regression (Azur et al., 2011; White, Royston, & Wood, 2011).

Initially, a single imputation, using mean imputation or simple random sampling, is performed for every missing value in the data set (called a place-holder imputation). These placeholder imputations for one variable (variable A) are then set back to missing. A regression model is then fitted with variable A as the dependent variable and the other variables in the imputation model as independent variables. The missing values for variable A are replaced by simulated draws from the corresponding posterior predictive distribution of variable A. When variable A is subsequently used as an independent variable in the regression models for other variables, both observed and imputed values are used. This process is repeated for all variables that have missing data and all missing values are imputed. These steps are repeated for a number of cycles, generally 10 to 20, with the imputations being updated at each cycle. At the end of these cycles the final imputations are retained, resulting in one imputed data set. By the end of the cycles, it is important that convergence is achieved and the distribution of the parameters governing the imputations become stable. This will avoid dependence on the order in which the variables are imputed. The convergence can be checked by comparing the regression models at subsequent cycles. This process is then repeated to create the *m* multiple imputed data sets (Azur et al., 2011; White et al., 2011).

MICE is flexible and can impute many different types of data (binary, ordinal, unordered categorical and continuous) because each variable is imputed using its own imputation model. Justification of the MICE procedure rested on empirical studies rather than clear theoretical rationale. Fitting a series of conditional distributions may not be consistent with a proper joint distribution. A consequence of incompatible conditional regressions is that the distribution of imputed values may depend on the order in which the variables were imputed (Azur et al., 2011; White et al., 2011).

When there are relatively few variables needing imputation, the variables are continuous and approximately normally distributed (in which case a multivariate normal model would be appropriate) a multivariate normal model may be preferable. Results based on MCMC sampling assuming that the data are multivariate normal agree asymptotically with those from MICE when all imputation models are linear (Azur et al., 2011; White et al., 2011).

Multiple imputation is said to be proper if it leads to consistent asymptotically normal estimators, correct variance estimators and valid tests. Proper multiple imputation will be multiple imputations for which Rubin's rule yields a consistent asymptotically normal estimator of the unknown parameter and a weakly unbiased estimator of the asymptotic variance in sufficiently regular models (Rubin, 1987). Proper multiple imputations are imputations where the values of the complete data statistics $Q$ and $U$ created through the multiple imputation are approximately unbiased for the complete-data analogues for large $m$; thus

$$E(\bar{Q}_\infty|X,Y) = \hat{Q} \text{ and}$$
$$E(\bar{U}_\infty|X,Y) = U \text{ and}$$
$$E(B_\infty|X,Y) = var(\bar{Q}_\infty|X,Y)$$

where $B$ is the variance-covariance matrix (Rubin, 1996). Rubin (1987) concluded that if imputations are drawn from a Bayesian posterior distribution of $\mathbf{Y}^{mis}$ under the response mechanism and an appropriate model for the data, then in large samples the imputation is proper.

Imputations drawn from a Bayesian predictive distribution are proper when the model used for the imputations and the model used for the analysis are compatible. When $\hat{Q}$ or $U$ involves some variable, $X$, then leaving $X$ out of the imputation scheme would result in an improper imputation, which would lead to biased estimation and invalid inference. Variables correlated with the imputed outcome and not included in the set of predictors will lead to bias. At a minimum clustering and stratification indicators and sample design weights should be included in the imputation model. Therefore the biggest problem with the imputer's model is excluding variables associated with the outcome. Including too many variables and including unimportant variables can lead to a small loss in precision but this is unimportant compared to the increased validity achieved when relevant variables are included. It is therefore recommended that as many variables as possible are included in the imputer's model (Rubin, 1996). Any variables that will be included in subsequent analyses, including interactions, should be included in the imputer's model. The imputer's model can include variables that will not be included in the analysis model. Generally the imputation will be proper if all sources of variability and uncertainty are included in the imputed values, including

prediction errors of the individual values and errors of estimation in the fitted coefficients of the imputation model (White et al., 2011).

Proper multiple imputation requires that the imputed data are drawn from the Bayesian posterior distribution where uncertainty about the parameters in the imputation model is properly represented, using uninformative priors for the parameters. Each set of imputed data will be based on a different set of model parameters, which are drawn for each imputation from the Bayesian posterior (Carpenter & Kenward, 2007).

Multiple imputation can be used with various analysis methods. It can be used with a cross-sectional analysis at the endpoint, a selection model, a pattern-mixture model or a shared-parameter model (Carpenter & Kenward, 2007; Janssens, Molenberghs, & Kerstens, 2012). Step 2 in the methods above is varied to use the method most appropriate to the problem at hand.

Little and Yau (1996) proposed a method where multiple imputation is used to analyse the data according to the ITT principle. Missing values of the outcome are imputed with multiple imputation using a model that conditions on the actual, or assumed, treatments received. The imputed data are then analysed based on the treatment as randomised. Observed variables that improve the imputation should be included in the imputation model even if they are not included in the analysis model; in this example actual treatment received is such a variable.

Multiple imputation can be used either in a longitudinal or cross-sectional setting. Longitudinal outcome data can be imputed by regarding the different time points as different variables. For example, with three time points, the outcomes at $t = 1$, $t = 2$ and $t = 3$ are included in the imputer's model. However, the time dependent property of longitudinal data is not taken into account during imputation, for example that the observation at $t = 1$ is before the observation at $t = 2$.

Several researchers have developed methods to undertake multiple imputation in a longitudinal setting. For example Liu, Taylor, and Belin (2000) implemented a random coefficients model to impute incomplete multivariate continuous longitudinal data. Multivariate repeated measures were jointly modelled, a normal model was assumed for time dependent variables in a regression model conditional on the time independent variables and time. Gibbs sampling in which the parameters and missing values are drawn iteratively from conditional distributions was used to draw model parameters and impute missing observations. Li, Mehrotra, and Barnard (2006) used a propensity score-based multiple imputation method for longitudinal data with binary responses. The imputations are done sequentially over time. Missing responses at time 2 were imputed first given data at time 1, then missing responses at time 3 were imputed given the observed plus imputed data

at times 1 and 2, etc. A propensity score was calculated at each time point with missing data and the propensity scores at each time point were used to group participants. Missing values were then imputed separately for each propensity score group to produce a complete data set. This imputation process was then repeated multiple times to create *m* complete data sets.

Several authors also describe methods for multiple imputation of time dependent covariates in a longitudinal setting. Carrigan et al. (2007) introduced a random intercept into the imputation component of the model to incorporate within-participant correlation and take into account the longitudinal study design. Nevalainen, Kenward, and Virtanen (2009) extended an iterative procedure, fully conditional specification, to generate values of a time-dependent covariate in the repeated measurement setting by being doubly iterative over the follow-up time of individuals. These methods were developed to impute covariates but can also be adapted to impute repeated outcome variables.

Rubin (1987) argued that only three to ten imputations are needed and five imputations is often used as default. Rubin showed that the efficiency of an estimate based on *m* imputations is approximately

$$(1 + \frac{\gamma}{m})^{-1},$$

where $\gamma$ is the rate of missing information. If the rate of missing information is 30% or lower, five imputations provide a 94% efficient estimator. With 50% missing information, five imputed data sets produce a point estimate that is 91% as efficient as one based on an infinite number of imputations, and 10 data sets produce a point estimate that is 96% efficient. With higher rates of missing information, more imputations had larger added benefit. For example, if the rate of missing information was 90% then five imputations had 85% efficiency and 20 imputations had 96% efficiency.

Rubin's (1987) formula focused on relative efficiency but did not take precision of standard errors and other estimates into account. Both efficiency of the point estimate and precision of estimation play a role in the number of imputations needed. With few imputations the standard error and therefore p-values and confidence intervals can be unstable. Looking at precision of standard errors instead of relative efficiency authors have recalculated the number of imputations needed and came up with larger numbers than Rubin did.

Graham, Olchowski, and Gilreath (2007) did simulations and found that 20 imputations led to a small loss of power with 10-30% missing information. If missing information was 50%, 40 imputations were recommended. Bodner (2008) also illustrated that less than ten imputations led

to substantial imprecision in important quantities based on the standard error such as p-values, confidence intervals and fractions of missing information. For example, with 50% missing information they recommended that 50 imputations were needed to achieve specified precision at 95% confidence levels. White et al. (2011) went further than considering statistical efficiency and power. They were interested in the repeatability of results and considered the Monte Carlo error (standard deviation across repeated runs of the same imputation procedure with the same data) of the results. Monte Carlo errors will be smaller with a larger number of imputations. They derived an easy rule of thumb, namely that the number of imputations should at least be larger than the percentage of missing data.

With easily available computing resources, there is no reason for a large number of imputations not to be used in relatively small data sets. White et al. (2011) recommend 100 to 500 imputations. Taking efficiency, power and precision into account it is recommended that 50 to 100 imputations are done.

Multiple imputation gives similar results to likelihood analyses when the imputation model is congenial, especially as the number of imputations increases. Multiple imputation offers advantages if covariates are missing, because likelihood analyses in these instances might be impracticable. Where responses only are missing and likelihood analysis is possible, multiple imputation adds little advantage (Molenberghs & Kenward, 2007).

Multiple imputation does not aim to create information by simulating values, but it rather tries to represent the observed information in a way that enables valid analysis with complete data tools, while taking account of increased variability created by missing data. The simulation is only used to handle the missing information. The rest of the information is handled by the complete data method (Rubin, 1996). Over the past decade multiple imputation has become one of the most used methods to handle missing data, probably due to its ease of use and versatility.

## 3.3 Modelling frameworks

Any analysis should be based on explicit and understandable assumptions. Three main frameworks exist where approaches to missing data can be developed. These are selection models, pattern-mixture models and shared-parameter models. Each is discussed briefly.

### 3.3.1 Selection models

Selection models require the specification of the full data model and a selection model that characterises the probability of data being missing as a function of covariates and the full data

(Hogan et al., 2004). In a selection model, the joint distribution of the $i^{th}$ participant's outcomes, $\boldsymbol{Y}_i$, and missingness indicators, $\boldsymbol{R}_i$, is factored as the marginal density of the measurement process and the conditional distribution of the missingness process given the measurement model $f(\boldsymbol{r}_i|\boldsymbol{y}_i)$; thus

$$f\ (\boldsymbol{y}_i, \boldsymbol{r}_i|X_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i|X_i, \boldsymbol{\theta}) f(\boldsymbol{r}_i|\boldsymbol{y}_i, W_i, \boldsymbol{\psi})$$

where $X_i$ is a design matrix for the measurement process and $W_i$ is a design matrix for the missingness mechanism.

It is called a selection model, since $f(\boldsymbol{r}_i|\boldsymbol{y}_i, W_i, \boldsymbol{\psi})$ can be seen as a participant's selection process to continue or leave the study. Participants are thus selected for dropout by their response. A participant's missing values are selected through a probability model, given their measurements, whether observed or not (Molenberghs & Kenward, 2007).

If we assume that the missing data process is MCAR this translates to

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i|X_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f\left(\boldsymbol{y}_i^{obs}|X_i, \boldsymbol{\theta}\right) f(\boldsymbol{r}_i|W_i, \boldsymbol{\psi}).$$

If we assume that the missing data process is MAR this factorisation translates to

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i|X_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f\left(\boldsymbol{y}_i^{obs}|X_i, \boldsymbol{\theta}\right) f\left(\boldsymbol{r}_i|\boldsymbol{y}_i^{obs}, W_i, \boldsymbol{\psi}\right).$$

This implies that the marginal model for the observed data only, is required. The joint likelihood is factored into a term including the observed data and $\boldsymbol{\theta}$ and a term including the missingness indicator and $\boldsymbol{\psi}$. The measurement model, $f\left(\boldsymbol{y}_i^{obs}|X_i, \boldsymbol{\theta}\right)$, and the missingness model, $f(\boldsymbol{r}_i|\boldsymbol{\psi})$ or $f\left(\boldsymbol{r}_i|\boldsymbol{y}_i^{obs}, W_i, \boldsymbol{\psi}\right)$ can be fitted separately, provided that the separability requirement holds. This implies that the parameters, $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$, are functionally independent or distinct of each other (Molenberghs & Kenward, 2007). In that case inference for $\boldsymbol{\theta}$ can be drawn from the observed data alone. In the MAR case covariates associated with missingness should be included in the model (Carpenter et al., 2002).

If the missingness is assumed to be MNAR, then the joint likelihood cannot be simplified. Thus, in order to find the maximum likelihood estimates, integration is required. Since this is not possible analytically, numerical integration is required to find the maximum. This can be difficult and time consuming. An EM algorithm or MCMC approach could be considered (Carpenter et al., 2002). In a Bayesian framework the missing observations are regarded as parameters and if vague priors are used the posterior means could be good approximations of the maximum likelihood estimates. The MNAR model has the same structure as the MAR model, with terms relating the missing data indicator to the individual's value of the outcome variable at time *j*, earlier time points and baseline covariates being added (Carpenter et al., 2002).

In the special case of dropout, denote the time at which dropout occurs as $D_i$, if no dropout occurred, $D_i = n_i + 1$. Diggle and Kenward (1994) combined a multivariate normal model for the measurement process, $f(\boldsymbol{y}_i|X_{\boldsymbol{i}}, \boldsymbol{\theta})$, and a logistic regression model for the dropout process, $f(\boldsymbol{r}_i|\boldsymbol{y}_{\boldsymbol{i}}, W_i, \boldsymbol{\psi})$. Define $\boldsymbol{h}_{ij} = (y_{ij}, \ldots, y_{i,j-1})'$ as the observed history of participant $i$ up to time $t_{i,j-1}$. The logistic dropout model could be written as

$$logit\left[P\left(D_i = j \middle| D_i \geq j, \boldsymbol{h}_{ij}, y_{ij}, \boldsymbol{\psi}\right)\right] = \psi_0 + \psi_1 y_{ij} + \psi_2 y_{i,j-1}.$$

If $\psi_1 = 0$ then this model refers to the MAR case, since dropout does not depend on the current measurement and when $\psi_1 = \psi_2 = 0$ this model refers to the MCAR case, since dropout does not depend on the outcome at all. The parameter and precision estimates are obtained using maximum likelihood. This involves computationally demanding integration, making this model difficult to use. Two key features of this model determine the identification of parameters: normality of the response distribution and linear dependence between

$$logit\left[P\left(D_i = j \middle| D_i \geq j, \boldsymbol{h}_{ij}, y_{ij}, \boldsymbol{\psi}\right)\right]$$

and $Y_{ij}$. It is not possible to evaluate whether the normality assumption holds if data are missing. Knowledge of the subject matter is often used to determine whether it is plausible that the data are normally distributed (Hogan et al., 2004).

Parametric selection models in a MNAR context raise problems around the identification of parameters and sensitivity to assumptions. These problems can be somewhat alleviated or at least illuminated by the use of semi-parametric selection models (National Research Council, 2010). Semi-parametric selection models can be constructed by using a parametric model for the missing data mechanism and a semi- or non-parametric model for the observed data or the complete data distribution. For example Rotnitzky, Robins, and Scharfstein (1998) considered such a non-likelihood approach and proposed a class of augmented inverse probability weighting estimators, an extension of GEE. When joint estimation of the non-response parameter and the outcome measure parameter is difficult, they proposed regarding the nonresponse model parameter as known and performing sensitivity analysis to determine how the outcome measure parameter changes. Scharfstein, Rotnitzky, and Robins (1999) extended this by allowing semi-parametric nonresponse mechanisms.

Conceptually selection models involve two stages. These stages are usually performed simultaneously. The first stage is to develop a predictive model for whether or not a participant has missing data with variables obtained prior to the missing data (such as at baseline). This model of missingness provides a predicted probability of missingness or propensity for each participant. These missingness propensity scores are then used in the second stage longitudinal data model as a

covariate to adjust for the potential influence of missing data. By modelling the missingness process, selection models provide valuable information regarding the predictors of missingness.

Different models may be appropriate for the missingness model. One model could assume that observations are missing at certain visits, but the participant may be observed again at the next visit. In this instance the simple model for the response at visit *j* can be given as:

$$logit\ P(R_{ij} = 1) = \alpha_j + \delta y_{ij}.$$

In other words the log odds of observing participant *i* at visit *j* depends on the visit and on the response. This model cannot be fitted alone. Using numerical integration over the unobserved responses, it can be fitted in conjunction with a mixed model for the response. $\delta$ depends on the distributional assumptions about the missing data. If $\delta = 0$, the MAR assumption holds and the probability of response does not depend on the missing observations. A more plausible model could also include other covariates and the outcome variable observed at previous visits. The more complicated model could be given as

$$logit\ P(R_{ij} = 1) = \alpha_j + \boldsymbol{\beta}'\boldsymbol{x_i} + \gamma Y_{i,j-1} + \delta y_{ij}$$

(Carpenter & Kenward, 2007).

The above model can accommodate withdrawal by adding

$$P(R_{ij} = 1 | R_{i,j-1} = 0) = 0 \text{ where } j > 1.$$

If withdrawal does not depend on the last observation, but on the pattern of observations throughout the trial, one can replace $Y_{ij}$ in the above model with the slope for participant *i* (Carpenter & Kenward, 2007).

Selection models require integration over the missing data. This can be done using numerical integration. Selection models could have limited practical use, since they require specialised numerical routines for maximising the likelihood. The likelihood could also be flat with respect to parameters that characterise the non-MAR selection, leading to numerical instability (Hogan et al., 2004).

Bayesian models with vague priors can also be used to fit these models, using MCMC methods (Carpenter & Kenward, 2007). Carpenter et al. (2002) suggested that selection models be fitted using a Bayesian framework since it is easier to program, quicker to estimate and flexible. They fitted it in BUGS (Spiegelhalter et al., 1995) using non-informative or vague priors. Using vague priors implies that the parameter estimates approximate maximum likelihood estimates. Within BUGS the sensitivity of the model to non-random missingness can be assessed.

### 3.3.2 Pattern-mixture models

Little (1993, 1994) described a class of models to model the dependence between missing response and dropout; calling these pattern-mixture models. He gave a statistically rigorous treatment of random-effects pattern-mixture models for longitudinal data with dropout. According to Little MAR assumptions need not be made; a model can be specified that does not require the missing data mechanism to be ignorable. Participants are divided into groups according to their missing data pattern. This method enables one to assess the degree to which important model terms depend on the missing data pattern (Hedeker & Gibbons, 1997).

Pattern-mixture models provide a flexible class of models for data that are not MCAR (Little, 1993). Pattern-mixture models are based on the factorisation

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i|\boldsymbol{x}_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i|\boldsymbol{r}_i, \boldsymbol{x}_i, \boldsymbol{\theta}) f(\boldsymbol{r}_i|, \boldsymbol{x}_i, \boldsymbol{\psi}) \qquad \text{(Equation 1)}$$

where the distribution of $\boldsymbol{Y}_i$ is conditioned on $\boldsymbol{R}_i$. The pattern-mixture model is based on a marginal model for the dropout process and a conditional model for the observed outcome given the dropout pattern (Verbeke, Lesaffre, & Spiessens, 2001). This allows for a different response model for each pattern of missing values. The observed data are a mixture of these weighted by the probability of each missing value or dropout pattern (Molenberghs & Kenward, 2007). The term pattern-mixture is derived from this and reflects that the marginal distribution of Y is a mixture of distributions (Little, 1993). If the missing data mechanism is ignorable then the component $f(\boldsymbol{r}_i|, \boldsymbol{x}_i, \psi)$ gives no information about $\theta$ and can be ignored for the likelihood inference. It can be based on the likelihood obtained by integrating missing values out of the density $f(\boldsymbol{y}_i|\boldsymbol{r}_i, \boldsymbol{x}_i, \boldsymbol{\theta})$ (Little, 1993; Little, 1994).

Define $t_i$, the last occasion at which a measurement was obtained, as $t_i = D_i - 1$; and $t_i$ is the realisation of the dropout index $T_i$. Then the factorisation of Equation 1 becomes

$$f(\boldsymbol{y}_i, t_i|\boldsymbol{x}_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i|t_i, \boldsymbol{x}_i, \boldsymbol{\theta}) f(t_i|, \boldsymbol{x}_i, \boldsymbol{\psi}).$$

Consider a trivariate normal outcome where $T_i$ can take values 1 and 2 for dropouts and 3 for completers. A pattern-mixture model implies a different distribution for each time of dropout, with $\boldsymbol{Y}_i|t_i \sim N\,[\,\boldsymbol{\mu}\,(t_i)\,, \Sigma\,(\,t_i\,)]$ (Molenberghs & Kenward, 2007).

Consider the case of a two-arm randomised trial. For simplification of notation, two dropout categories are considered only; either a participant completed the trial or did not complete the trial. This can be extended to more patterns. Let

$Y_{iC}$: Response for the $i^{th}$ participant in the control arm

$R_{iC}$ : Indicator whether participant in the control arm completed the trial. $R_{iC} = 1$ if completed, $0$ otherwise

$\pi_C$ : Probability of withdrawal in the control arm, $\pi_C = P(R_{iC} = 0)$

$Y_{iI}$, $R_{iI}$ and $\pi_I$ is defined similarly for the intervention arm

In the control arm, observed responses come from a distribution with mean $\mu_C$ and variance $\sigma^2$, while unobserved responses, in participants who dropped out, come from a shifted distribution with mean $\mu_C + \delta_C$ and variance $\sigma_M^2$, where $\delta_C$ is the difference in the control arm between the true mean of the unobserved data and the true mean of the observed data. Thus

$$Y_{iC} | R_{iC} = 1 \sim (\mu_C, \sigma^2)$$

$$Y_{iC} | R_{iC} = 0 \sim (\mu_C + \delta_C, \sigma_M^2)$$

For the intervention arm, $\mu_I$ and $\delta_I$ are defined similarly. The variances are assumed to be equal in the control and intervention arms. Under this model the average response in the control arm is

$$(1-\pi_C)\,\mu_C + \pi_C(\mu_C + \delta_C).$$

The average response in the intervention arm is

$$(1-\pi_I)\,\mu_I + \pi_I(\mu_I + \delta_I).$$

If $\delta_C = \delta_I = 0$ then the MAR assumption holds. The average effect of the intervention is

$$\Delta = (1-\pi_I)\mu_I + \pi_I(\mu_I + \delta_I) - [(1-\pi_C)\mu_C + \pi_C(\mu_C + \delta_C)] = (\mu_I - \mu_C) + (\delta_I\pi_I - \delta_C\pi_C)$$

The average treatment effect amongst completers, $(\mu_I - \mu_C)$, can be estimated using the complete case analysis. Therefore the total treatment effect is equal to the treatment effect in completers plus the bias due to informative withdrawal (Carpenter & Kenward, 2007).

The assumptions made by the pattern-mixture model are the following:

- In this example, normality is assumed within pattern. Non-Gaussian outcomes can also be accommodated.
- The conditional distribution of the outcome is assumed to depend on the dropout time only through the dropout pattern. Outcome is assumed to be independent of dropout within a pattern.
- Intermittent missingness is assumed to be MAR.
- Variances are assumed to be constant across patterns, but this assumption can be relaxed.
- Covariate effects are the same for missing and observed data within a dropout pattern (Hogan et al., 2004).

White et al. (2007) took a Bayesian approach in a non-longitudinal setting and assumed a prior distribution for the informative missingness parameters. In the control arm the distribution of $\delta_C \sim N(m_C, s_C^2)$ and in the intervention arm the distribution of $\delta_I \sim N(m_I, s_C^2)$. White et al. then assume non-informative priors, independent of $\delta$. The probability of dropout, $\pi_I$ and $\pi_C$, are

estimated by the fraction of dropouts in each of the arms ($p_I$ and $p_C$). The posterior mean and variance of the treatment effect is estimated by correcting the complete case estimate taking account of the informative dropout. Let $\Delta^{CC} = (\mu_1 - \mu_C)$ and $se(\hat{\Delta}^{CC})$ be the standard error of the complete case treatment estimate. Then the posterior mean is calculated as:

$$\hat{\Delta}^{CC} + m_I p_I - m_C p_C,$$

while the posterior variance is given by:

$$se(\hat{\Delta}^{CC})^2 + p_I^2 s_I^2 - 2cs_C s_I p_C p_I + p_C^2 s_C^2 + (m_I^2 + s_I^2)\frac{p_I(1-p_I)}{n_I} + (m_C^2 + s_C^2)\frac{p_C(1-p_C)}{n_C}$$

The posterior mean is the mean of the complete case analysis corrected for the prior distribution of the means and the observed proportion of drop outs. The estimate for the posterior variance includes the variance for the complete case estimate corrected for uncertainty about $\delta_C$, $\delta_I$, $p_C$ and $p_I$ and using the prior variances $s_C$ and $s_I$. The latter terms decrease with sample size. This method inflates the standard error to reflect uncertainty due to missing information. It is also possible to include covariates on which dropout and outcome depend (White et al., 2007).

White et al. (2007) elicited the distribution of possible values of the parameters in the prior distributions from experts in the field by obtaining opinions from experts on the difference in true means between the observed and unobserved data. The concept of a correlation was too unfamiliar to the experts and information on the correlation between $\delta_C$ and $\delta_I$ could not be elicited. If no prior information is available, a working value of the correlation can be adopted and the value for $\delta$ which causes the treatment effect to be non-significant is then found. Zero is a good choice for the correlation since this gives the widest confidence intervals. The plausibility of this value of $\delta$ is then assessed. If the reasons why outcomes are missing are recorded it may be appropriate to have a different $\delta$ parameter for each reason. Separate $\delta$ parameters should be perfectly correlated within trial arms; introducing imperfectly correlated $\delta$ parameters would erode the systematic element and artificially reduce the correction for dropout.

These pattern-mixture models could be implemented with random-effects mixed models. The steps in implementing these models are as follows:

1. Assign a variable or a set of dummy variables that identify the pattern of the missing data. Some patterns of missing data can be grouped together if some groups are small.
2. Fit a mixed model with the pattern of missing data variable as a main effect and as an interaction with other effects in the model.
3. Derive an overall averaged estimate for the model parameters, averaging over the missing data patterns. Estimates are obtained for the fixed effects separately for completers ($\hat{\beta}^c$) and dropouts ($\hat{\beta}^d$). Averaged estimates for these parameters are then equal to

$$\hat{\beta} = \pi^c \hat{\beta}^c + \pi^d \hat{\beta}^d$$

where $\pi^c$ and $\pi^d$ represent the population weights for completers and dropouts. These weights can be estimated by the sample proportions. Estimates of the standard errors can be obtained using the delta method (Hogan and Laird 1997) as described in Section 3.3.2.1.

Pattern-mixture models can be used with random-effects models, but can also be used with other longitudinal methods that allow for missing data across time, such as structural equation models and GEE (Hedeker & Gibbons, 1997). Hedeker and Gibbons only discussed the implementation with mixed models.

Carpenter and Kenward (2007) suggested the use of a pattern-mixture approach implementing multiple imputation. The analysis proposed by White et al. (2007) can be extended to longitudinal data by modifying the conditional distribution of the missing data given the observed data under MAR after a participant dropped out. The assumption is made that participants who drop out have on average a poorer (or different) response than predicted by MAR. Let the change in rate of decline be denoted by $\delta$. Then the conditional mean for the first response after withdrawal is reduced by $\delta$, the second by $2\delta$, etc (see Figure 3.1). The mean and variance of $\delta_l$ for all treatment groups $l$ are elicited from experts. It is assumed to be normally distributed. The correlation between any two treatment arms should also be elicited.

This approach uses multiple imputation as follows: The MAR assumption is made and $m$ imputations are created. With two treatment groups the following is sampled for each of the $k$ imputations:

$$\begin{pmatrix} d_{1k} \\ d_{2k} \end{pmatrix} \sim N\left( \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \right)$$

For each imputation the first MAR imputed value is decreased by $d_{lk}$, the second by $2d_{lk}$, etc. The data sets are then analysed as discussed in Section 3.2 for multiple imputation. If the time between observations is not equal, the multipliers of $d$ can be chosen to reflect the spacing between observations. Interim missing observations can be decreased by $d_l$ or can be left with the MAR imputed values. This is consistent with assuming that interim missing data are different from drop out and is truly random, whereas participant drop out is not random. In the absence of prior information one can set $\delta_1 = \delta_2$.

Pattern-mixture models are under-identified and hence over-specified. If data are missing then there are no data to identify the $t^{\text{th}}$component of $\boldsymbol{\mu}(t_i)$ or the $t^{\text{th}}$column of $\Sigma(t_i)$. The data supply no information about the parameters of the distribution with missing data (Little, 1993). Little

(1993, 1994) solves this problem by placing restrictions. The restrictions depend on the missing data mechanism assumed and reflect the contextual knowledge of the nature of the mechanism creating the missing data. Inestimable parameters of the incomplete patterns are set equal to functions of the parameters describing the distribution of the completers. Alternative approaches for overcoming the under-identification exist (Molenberghs & Kenward, 2007). Molenberghs and Kenward (2007) argue that the under-identification can be seen as an advantage, because it forces one to make assumptions transparent. This serves as a useful starting point for sensitivity analyses.

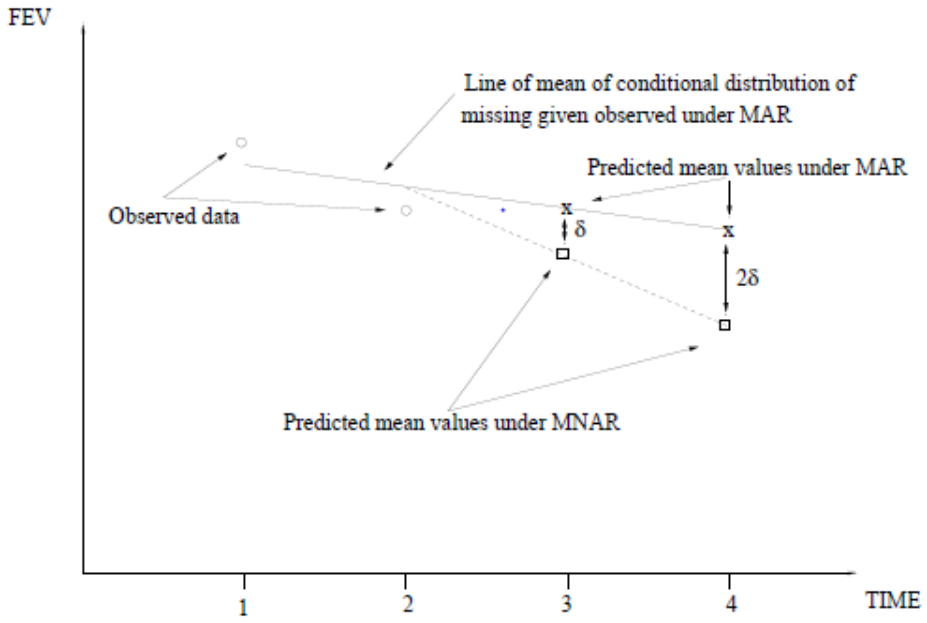



**Figure 3.1: Schematic illustration of increasing the rate of decline by δ after withdrawal (Carpenter & Kenward, 2007)**

Little (1994) advocated the use of identifying restrictions and listed some. In the case of monotone missing data, with $t = 1, \ldots, n$ dropout patterns, the complete data density for pattern $t$ is given by

$$f_t(y_1,\ldots,y_n) = f_t(y_1,\ldots,y_t) f_t(y_{t+1},\ldots,y_n \mid y_1,\ldots,y_t).$$

The first factor can be identified or modelled from the observed data. The second factor is not identified from the observed data. Identifying restrictions are applied in order to identify the second component. One can base identification on all patterns for which a given component is identified. This is given by

$$f_t(y_s \mid y_1,\ldots,y_{s-1}) = \sum_{j=s}^{n} \omega_{sj} f_j(y_s \mid y_1,\ldots,y_{s-1})$$

with $s = t+1,\ldots,n$. Thus

$$f_t(y_1,\ldots,y_n) = f_t(y_1,\ldots,y_t) \prod_{s=0}^{T-n-1} \sum_{j=n-s}^{n} \omega_{n-s,j} f_j(y_{n-s} \mid y_1,\ldots,y_{n-s-1})$$

(Molenberghs & Kenward, 2007).

Different choices can be made for $\omega_s$, leading to different special cases of the above equation. The only restriction is that the $\omega_s$ are non-negative and sum to one (Thijs et al., 2002).

1. Complete case missing values (CCMV).

   In the complete case missing value approach $\omega_{sn} = 1$ and all other $\omega$ = 0. This implies that missing information is borrowed from the completers only. If the model effect for a dropout pattern cannot be estimated, the estimated model effect of the fully observed pattern will be used. This method requires a reasonable number of complete cases and can be inefficient if the fraction of complete cases is small (Little, 1993).

2. Neighbouring case missing values (NCMV).

   In this instance $\omega_{ss} = 1$ and all other $\omega = 0$ are missing. Information is borrowed from the closest available pattern (Kenward, Molenberghs, & Thijs, 2003).

3. Available case missing values (ACMV). In this case,

$$\omega_{sj} = \frac{\alpha_j f_j(y_1,...,y_{s-1})}{\sum_{m=s}^{n} \alpha_m f_m(y_1,...,y_{s-1})}$$

   where $\alpha_j$ is the fraction of observations in pattern *j*. In the absence of an estimate ACMV will average over all patterns where the model effect could be estimated. The average is weighted proportional to the number of participants in the pattern. This is the counterpart of MAR in the pattern-mixture model context. The other two identifying restrictions (complete case missing values and neighbouring case missing values) lead to MNAR models (Molenberghs & Kenward, 2007; Molenberghs et al., 1998).

Previously three steps in fitting pattern-mixture models with random-effects mixed models were given. Here we give the steps to fit pattern-mixture models using identifying restrictions (Kenward et al., 2003; Thijs et al., 2002):

1. Specify and fit a model to the pattern specific identifiable densities $f_t(y_1,...,y_n)$.
2. Choose a model for $f_t(y_{t+1} \mid y_1,...,y_t)$ or select an identification method. This can be done in a data independent way by placing a prior on the parameters. If this is done in a data dependent way procedures described in number 4 below can be followed.
3. Using this model or identification method, the conditional distribution of the unobserved outcomes, given the observed ones, $f_t(y_{t+2},...,y_n \mid y_1,...,y_{t+1})$, is determined.
4. Inference is now based on the observed quantities using the full distribution specified. This could imply integration over the distribution of the unknown quantities. However,

simulation-based approaches such as multiple imputation is often implemented. The multiple imputation is done for the unobserved components, given the observed component and the pattern-specific densities.

5. The multiple imputed data sets are then analysed using proper methods as described in Section 3.2, a pattern-mixture model, selection model or any other model would be appropriate. The analysis model does not need to be the same as the imputation model.

Instead of identifying restrictions model simplification can be done to identify the parameters (Thijs et al., 2002). Two types of simplification can be done in fitting pattern-mixture models. The trend can be restricted to functional forms supported by the information available in a pattern. These models are fitted by creating a model within each of the patterns separately. Secondly, the parameters can vary across the patterns in a parametric way; for example, it could be assumed that the time trend is parallel across patterns. These models are fitted by treating the pattern as a covariate (Hogan & Laird, 1997; Michiels et al., 2002; Thijs et al., 2002). The available data can be used to assess whether such simplifications are supported where data are available. It is argued that these solutions are assumption rich. If identifying restrictions are used, the assumptions are more clearly specified.

Estimation relies on linearity and homoscedasticity of the regression and the missing data mechanism is assumed to depend on $\mathbf{Y}_i$ in an additive manner (Little, 1994). Maximum likelihood methods for selection models require numerical methods, whereas the pattern-mixture models lead to maximum likelihood estimates with explicit forms (Little, 1993).

Pattern-mixture models can be used under ITT in a two arm trial of active treatment versus placebo by making two assumptions. The first is that participants who withdraw discontinue treatment and the second is that future outcomes given past outcomes are the same for participants who withdraw in the active arm as for participants in the placebo group with the same past. The appropriate pattern-mixture model can be constructed by using the estimated model from the placebo group to represent the future behaviour of withdrawals from the active group. This model for the future behaviour is the MAR model for all observed data and can be consistently estimated from the observed data. This pattern-mixture model differs only from a likelihood model fitted under MAR in its implications for the unobserved behaviour of withdrawals from the active group. Imputations for future outcomes differ between the two models, and this affects the estimated final treatment comparisons. A variety of alternative pattern-mixture models can be constructed to examine the impact of different behaviours of the withdrawals on the results from the ITT analyses.

Little and Yau (1996) call this model the zero dose model. They also suggested other applications of pattern-mixture models for example the continuing dose and the nearest dose model. They proposed that missing values of the outcome after drop out be imputed with multiple imputation using a model that conditions on the actual, or assumed, dose of treatments received after dropout; data are thus imputed according to different patterns depending on the actual treatment received, leading to sensitivity analyses. In their study they collected information on the doses received after dropout. In the nearest dose model, cases in the control group are assigned a zero dose after drop out and cases in the treatment group are assigned a treatment dose group closest to the actual recorded dose after drop out. The data imputed using the data observed after drop out are then analysed based on the treatment as randomised, according to the classical ITT principle.

Pattern-mixture models can be factorised as follows:

$$f\ (y_1,...,y_n,d=t) = f_t(y_1,...,y_t) f_t(y_{t+1} \mid y_1,...,y_t) f_t(y_{t+2},...,y_n \mid y_1,...,y_{t+1}) f_t(d=t)$$

(Kenward et al., 2003).

Fitting a pattern-mixture model creates some complications. One pair of treatment contrasts is created for each pattern fitted. One can either fit a stratified analysis, where the null hypothesis addresses all the pairwise contrasts simultaneously or one can analyse the marginal effects, for example a single marginal treatment contrast (Hogan & Laird, 1997). Pattern-mixture models do not automatically provide estimates and standard errors of marginal quantities of interest, such as the overall treatment effect (Thijs et al., 2002). The marginal contrasts are obtained by weighting each of the treatment estimates by the number of participants in the pattern. The marginal within-imputed and between-imputed variances are obtained using the delta method (Hogan & Laird, 1997). The primary analysis usually is the marginal analysis (Kenward et al., 2003).

Pattern-mixture models can also be used to identify which patterns are responsible for a treatment effect. The stratified analysis allows a more detailed consideration of the treatment effect (Kenward et al., 2003).

When the marginal distribution of the outcomes is of interest, the mixing over the different dropout patterns is needed. These models often include many parameters and some may be estimated inefficiently (Verbeke et al., 2001). An advantage of pattern-mixture models is that it is regarded as more honest than other methods since the untestable assumptions are stated and not implicit. The need for assumptions and their implications are more obvious in pattern-mixture models. These models could also be computationally easier to execute (Michiels et al., 2002). With pattern-mixture modelling one can decide whether or not to model the data beyond the moment of dropout,

whereas with selection models this always happens (Thijs et al., 2002). However, in order to estimate the large number of parameters in most pattern-mixture models, each dropout pattern needs to occur sufficiently often (Hogan & Laird, 1997).

Pattern-mixture models play a role in sensitivity analyses, because they separate the observed data distribution and the predictive distribution of missing data given observed data (National Research Council, 2010).

#### 3.3.2.1 The Delta Method

The derivation of most of the theory discussed in this overview chapter is available in referenced texts and was not repeated here. However, the formulas for using the delta method to calculate the standard errors are not available when there are more than two categories of missing data. This section therefore gives the complete derivation using the delta method where there are three or four groups in the pattern-mixture model. These derivations are done so that variances and parameter estimates obtained from standard procedures in standard statistical software can be used to calculate the standard error, thus increasing the likelihood that it can be used by practicing applied statisticians. Generalisation to any number of groups is relatively straightforward and can be done conveniently using vector and matrix algebra.

The delta method can be applied to calculate the variance when a pattern-mixture model is used. This is described for the two and three group example. The theorem and proof for the univariate version of the delta method is given in Casella and Berger (2002), page 243. The theorem states that if $\hat{\theta}_n$ is a sequence of random variables that satisfies

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \rightarrow N(0, \sigma^2)$$

in distribution, then for a given function $g$,

$$\sqrt{n}\left[g\left(\hat{\theta}_n\right) - g(\theta)\right] \rightarrow N(0, \sigma^2[g'(\theta)]^2);$$

provided that $g'(\theta)$ exists and is not *0*.

A multivariate version of the delta method was also specified (Casella & Berger, 2002). Assume that $\widehat{\boldsymbol{\theta}}_n$ is a *p x 1* vector with an asymptotic normal distribution:

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \rightarrow N(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta})),$$

where $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is a $p \times p$ asymptotic covariance matrix of $\sqrt{n}\widehat{\boldsymbol{\theta}}_n$. For large *n*, $\widehat{\boldsymbol{\theta}}_n$ is distributed $N(\boldsymbol{\theta}, \frac{1}{n}\boldsymbol{\Sigma}(\boldsymbol{\theta}))$ or $N(\boldsymbol{\theta}, \boldsymbol{V}(\boldsymbol{\theta}))$, where $\boldsymbol{V}(\boldsymbol{\theta}) = \frac{1}{n}\boldsymbol{\Sigma}(\boldsymbol{\theta})$. Let $\boldsymbol{g}(\boldsymbol{\theta})$ be a vector function of $\boldsymbol{\theta}$, i.e.

$$\boldsymbol{g}(\boldsymbol{\theta}) = (g_1(\boldsymbol{\theta}), g_2(\boldsymbol{\theta}), \ldots, g_r(\boldsymbol{\theta}))',$$

which has a continuous first partial derivative. Then

$$\sqrt{n}\,[\left(g(\widehat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta})\right] \rightarrow N(0, \left(\frac{\partial g(\theta)}{\partial \theta}\right)\boldsymbol{V}(\boldsymbol{\theta})\left(\frac{\partial g(\theta)}{\partial \theta}\right)').$$

Hedeker and Gibbons (1997) calculated an overall estimate per treatment arm combining the pattern observed in the completers and the pattern observed in the dropouts. This is done by letting $X_c$ denote the number of completers and $X_d$ denote the number of dropouts. The proportion of completers and dropouts are given by $\pi_c$ and $\pi_d$, respectively, and the parameter estimates in the completers and dropouts are denoted by $\beta_c$ and $\beta_d$. Furthermore, $\pi_d = 1 - \pi_c$, since $\pi_c + \pi_d = 1$. Now let

$$\boldsymbol{\theta} = [\pi_c, \beta_c, \beta_d\,]' \text{ and } g(\boldsymbol{\theta}) = \pi_c\beta_c + \pi_d\beta_d.$$

The averaged overall estimate is then calculated by replacing the $\pi$'s with the proportion of participants in the sample in the appropriate group and replacing the $\beta$'s with the estimate of the parameter in each of the groups. The authors give a formula for calculating the variance of $g(\boldsymbol{\theta})$ but do not show the derivation of this formula. It is derived as follows for each of the fixed effects. The superscript denoting the fixed effect is excluded to simplify the formulas.

It follows that

$$X_c \sim BIN(n, \pi_c\,), \text{ and } \hat{\pi}_c = \frac{X_c}{n} \simeq N(\pi_c, \frac{\pi_c(1-\pi_c)}{n})].$$

The variance-covariance matrix of $\widehat{\boldsymbol{\theta}}$ is given by

$$V(\widehat{\boldsymbol{\theta}}) = \begin{bmatrix} \frac{\pi_c(1-\pi_c)}{n} & 0 & 0 \\ 0 & var(\hat{\beta}_c) & 0 \\ 0 & 0 & var(\hat{\beta}_d) \end{bmatrix} = \begin{bmatrix} v_1 & 0 & 0 \\ 0 & v_2 & 0 \\ 0 & 0 & v_3 \end{bmatrix}.$$

Completers and dropouts are assumed to be independent groups. The function $g(\boldsymbol{\theta})$ can also be written as

$$g(\boldsymbol{\theta}) = \pi_c\beta_c + (1 - \pi_c)\beta_d,$$

$$\text{giving } \frac{\partial g(\boldsymbol{\theta})}{\partial \pi_c} = \beta_c - \beta_d;$$

$$\frac{\partial g(\boldsymbol{\theta})}{\partial \beta_c} = \pi_c \text{ and}$$

$$\frac{\partial g(\boldsymbol{\theta})}{\partial \beta_d} = 1 - \pi_c\,.$$

$$\mathrm{V}[\mathrm{g}(\widehat{\boldsymbol{\theta}})] = [\beta_c - \beta_d, \pi_c, 1 - \pi_c] \begin{bmatrix} v_1 & 0 & 0 \\ 0 & v_2 & 0 \\ 0 & 0 & v_3 \end{bmatrix} \begin{bmatrix} \beta_c - \beta_d \\ \pi_c \\ 1 - \pi_c \end{bmatrix}$$

$$= [v_1(\beta_c - \beta_d), \pi_c v_2, (1 - \pi_c)v_3] \begin{bmatrix} \beta_c - \beta_d \\ \pi_c \\ 1 - \pi_c \end{bmatrix}$$

$$= v_1(\beta_c - \beta_d)^2 + \pi_c{}^2\, v_2 + (1 - \pi_c)^2\, v_3$$

$$= (\beta_c - \beta_d)^2 \quad \frac{\pi_c(1-\pi_c)}{N} + \pi_c{}^2 \ \mathrm{var}(\beta_c) + \ (1-\pi_c)^2 \ \mathrm{var}(\beta_d).$$

The estimate of the variance is

$$\widehat{\mathrm{var}}\left[\mathrm{g}(\hat{\boldsymbol{\theta}})\right] = (\hat{\beta}_c - \hat{\beta}_d)^2 \ \frac{\hat{\pi}_c(1-\hat{\pi}_c)}{N} + \ (\hat{\pi}_c)^2\widehat{\mathrm{var}}(\hat{\beta}_c) + (1\text{-}\hat{\pi}_c)^2\widehat{\mathrm{var}}(\hat{\beta}_d)$$

$$= (\hat{\beta}_c - \hat{\beta}_d)^2 \ \frac{\hat{\pi}_c\hat{\pi}_d}{N} + (\hat{\pi}_c)^2\widehat{\mathrm{var}}(\hat{\beta}_c) + (\hat{\pi}_d)^2\widehat{\mathrm{var}}(\hat{\beta}_d).$$

**The extension of the above to three groups is as follows:**

For three groups, say A, B and C, let the population weights be $\pi_a$, $\pi_b$ and $\pi_c$ and

$$\pi_a + \ \pi_b + \pi_c = 1,$$

so that

$$\pi_c = 1 - (\pi_a + \pi_b).$$

The vectors of the parameter estimates from the individual models for each of the patterns are $\boldsymbol{\beta}_A$, $\boldsymbol{\beta}_B$ and $\boldsymbol{\beta}_C$. Let $\boldsymbol{\theta} = (\pi_a, \pi_b, \beta_A, \beta_B, \beta_C)$ and as in the two group case,

$$\text{let } g(\boldsymbol{\theta}) = \pi_a\beta_A + \pi_b\beta_B + \pi_c\beta_C$$

which can also be written as

$$g(\boldsymbol{\theta}) = \pi_a\beta_A + \pi_b\beta_B + [1 - (\pi_a + \pi_b)]\beta_C \text{ and}$$

$$\mathrm{V}(\hat{\boldsymbol{\theta}}) = \begin{bmatrix} \mathrm{var}(\hat{\pi}_a) & \mathrm{cov}_{AB} & 0 & 0 & 0 \\ \mathrm{cov}_{AB} & \mathrm{var}(\hat{\pi}_b) & 0 & 0 & 0 \\ 0 & 0 & \mathrm{var}(\hat{\beta}_a) & 0 & 0 \\ 0 & 0 & 0 & \mathrm{var}(\hat{\beta}_b) & 0 \\ 0 & 0 & 0 & 0 & \mathrm{var}(\hat{\beta}_c) \end{bmatrix},$$

where $\mathrm{cov}_{\mathrm{AB}} = \mathrm{cov}(\hat{\pi}_\mathrm{a}, \hat{\pi}_\mathrm{b})$.

The plausible multinomial assumption is made that the three dropout status groups are independent, therefore the covariances between the β's are 0. However, the proportions $\hat{\pi}_a$ and $\hat{\pi}_b$ are not independent and the covariances are not 0 in that case. The required derivatives are:

$$\frac{\partial g(\boldsymbol{\theta})}{\partial \pi_\mathrm{A}} = \beta_A - \beta_C$$

$$\frac{\partial g(\boldsymbol{\theta})}{\partial \pi_\mathrm{B}} = \beta_B - \beta_C$$

$$\frac{\partial g(\boldsymbol{\theta})}{\partial \beta_\mathrm{A}} = \pi_a$$

$$\frac{\partial g(\boldsymbol{\theta})}{\partial \beta_\mathrm{B}} = \pi_b$$

$$\frac{\partial g(\boldsymbol{\theta})}{\partial \beta_\mathrm{C}} = 1 - (\pi_a \ + \pi_b \ ).$$

$$\mathrm{V}\left[\mathrm{g}(\hat{\boldsymbol{\theta}})\right] = \ \left[\hat{\beta}_A - \hat{\beta}_C, \ \hat{\beta}_B - \hat{\beta}_C, \ \pi_\mathrm{a}, \ \pi_\mathrm{b}, \{1 - (\pi_\mathrm{a} + \pi_\mathrm{b})\}\right]\mathrm{V}(\hat{\boldsymbol{\theta}}) \left[\frac{\partial g(\boldsymbol{\theta})}{\partial \theta}\right]$$

$$= \left[(\hat{\beta}_A - \hat{\beta}_C)var(\hat{\pi}_a) + cov_{AB}(\hat{\beta}_B - \hat{\beta}_C), \quad (\hat{\beta}_A - \hat{\beta}_C)cov_{AB} + (\hat{\beta}_B - \hat{\beta}_C)var(\hat{\pi}_b), \right.$$

$$\left. \pi_a var(\hat{\beta}_A), \pi_b var(\hat{\beta}_B), \quad \{1 - (\pi_a + \pi_b)\}var(\hat{\beta}_C)\right]\left[\frac{\partial g(\boldsymbol{\theta})}{\partial \theta}\right]$$

$$= (\hat{\beta}_A - \hat{\beta}_C)^2 var(\hat{\pi}_a) + 2cov_{AB}(\hat{\beta}_A - \hat{\beta}_C)(\hat{\beta}_B - \hat{\beta}_C) + (\hat{\beta}_B - \hat{\beta}_C)^2 var(\hat{\pi}_b) + \pi_a{}^2 var(\hat{\beta}_A) + \pi_b{}^2 var(\hat{\beta}_B) + \{1 - (\pi_a + \pi_b)\}^2 var(\hat{\beta}_C) \quad \text{(Equation 2)}$$

The variances and covariances in this equation above are equal to

$$\text{cov}_{AB} = \text{cov}_{BA} = \text{cov}(\hat{\pi}_a, \hat{\pi}_b) = \frac{-\pi_a \pi_b}{N}$$

$$\text{var}(\hat{\pi}_a) = \frac{\pi_a(1 - \pi_a)}{N}$$

$$\text{var}(\hat{\pi}_b) = \frac{\pi_b(1 - \pi_b)}{N}$$

If $cov_{AB}$, $\text{var}(\hat{\pi}_a)$, $\text{var}(\hat{\pi}_b)$ are replaced in Equation 2 the variance is estimated by the following:

$$\widehat{var}[g(\hat{\boldsymbol{\theta}})] = (\hat{\beta}_A - \hat{\beta}_C)^2 \frac{\hat{\pi}_a(1 - \hat{\pi}_a)}{N} - 2\frac{\hat{\pi}_a \hat{\pi}_b}{N}(\hat{\beta}_A - \hat{\beta}_C)(\hat{\beta}_B - \hat{\beta}_C) + (\hat{\beta}_B - \hat{\beta}_C)^2 *$$

$$\frac{\hat{\pi}_b(1 - \hat{\pi}_b)}{N} + \hat{\pi}_a^2 var(\hat{\beta}_A) + \hat{\pi}_b^2 var(\hat{\beta}_B) + [1 - (\hat{\pi}_a + \hat{\pi}_b)]^2 var(\hat{\beta}_C).$$

This formulation of the equation is not in the most natural form to fit the model. The following changes can be made to write the formula in a form that is easier to use.

As above, it follows that:

$$\pi_a + \pi_b + \pi_c = 1,$$

$$\text{so that } \pi_c = 1 - (\pi_a + \pi_b)$$

$$\text{and } \overline{\boldsymbol{\beta}} = g(\boldsymbol{\theta}) = \pi_a \beta_A + \pi_b \beta_B + \pi_c \beta_C$$

which can also be written as

$$g(\boldsymbol{\theta}) = \pi_a \beta_A + \pi_b \beta_B + [1 - (\pi_a + \pi_B)]\beta_C.$$

One change is made by defining

$$\beta_B = \beta_A + \beta_{\Delta 1} \text{ and } \beta_C = \beta_A + \beta_{\Delta 2}$$

with $\beta_{\Delta 1}$ and $\beta_{\Delta 2}$ indicating how groups B and C differ from group A then

$$\begin{aligned} g(\boldsymbol{\theta}) = \overline{\boldsymbol{\beta}} &= \pi_a \beta_A + \pi_b \beta_B + \pi_c \beta_C \\ &= \pi_a \beta_A + \pi_b(\beta_A + \beta_{\Delta 1}) + [1 - (\pi_a + \pi_b)](\beta_A + \beta_{\Delta 2}) \\ &= \pi_a \beta_A + \pi_b \beta_A + \pi_b \beta_{\Delta 1} + \beta_A + \beta_{\Delta 2} - \pi_a \beta_A - \pi_a \beta_{\Delta 2} - \pi_b \beta_A - \pi_b \beta_{\Delta 2} \\ &= \beta_A + \pi_b \beta_{\Delta 1} + [1 - (\pi_a + \pi_b)]\beta_{\Delta 2} \\ &= \beta_A + \pi_b \beta_{\Delta 1} + \pi_c \beta_{\Delta 2}. \end{aligned}$$

Define $\boldsymbol{\theta} = (\beta_A, \beta_{\Delta 1}, \beta_{\Delta 2}, \pi_b, \pi_c)$. Then

$$\frac{\partial g(\boldsymbol{\theta})}{\partial \beta_{\mathrm{A}}} = 1$$

$$\frac{\partial g(\boldsymbol{\theta})}{\partial \pi_{\mathrm{b}}} = \beta_{\Delta 1}$$

$$\frac{\partial g(\boldsymbol{\theta})}{\partial \beta_{\Delta 1}} = \pi_b$$

$$\frac{\partial g(\boldsymbol{\theta})}{\partial \beta_{\Delta 2}} = \pi_c$$

$$\frac{\partial g(\boldsymbol{\theta})}{\partial \pi_{\mathrm{C}}} = \beta_{\Delta 2}.$$

$$\mathrm{V}(\widehat{\boldsymbol{\theta}}) = \begin{bmatrix} \mathrm{var}(\hat{\beta}_A) & \mathrm{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 1}) & \mathrm{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 2}) & 0 & 0 \\ \mathrm{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 1}) & \mathrm{var}(\hat{\beta}_{\Delta 1}) & 0 & 0 & 0 \\ \mathrm{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 2}) & 0 & \mathrm{var}(\hat{\beta}_{\Delta 2}) & 0 & 0 \\ 0 & 0 & 0 & \mathrm{var}(\hat{\pi}_b) & \mathrm{cov}(\hat{\pi}_b, \hat{\pi}_c) \\ 0 & 0 & 0 & \mathrm{cov}(\hat{\pi}_b, \hat{\pi}_c) & \mathrm{var}(\hat{\pi}_c) \end{bmatrix}$$

$$\left(\frac{\partial g(\boldsymbol{\theta})}{\partial \theta}\right) \mathrm{V}(\widehat{\boldsymbol{\theta}}) \left(\frac{\partial g(\boldsymbol{\theta})}{\partial \theta}\right)' =$$

$$[1, \pi_b, \pi_c, \beta_{\Delta 1}, \beta_{\Delta 2}] \begin{bmatrix} \mathrm{var}(\hat{\beta}_A) & \mathrm{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 1}) & \mathrm{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 2}) & 0 & 0 \\ \mathrm{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 1}) & \mathrm{var}(\hat{\beta}_{\Delta 1}) & 0 & 0 & 0 \\ \mathrm{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 2}) & 0 & \mathrm{var}(\hat{\beta}_{\Delta 2}) & 0 & 0 \\ 0 & 0 & 0 & \mathrm{var}(\hat{\pi}_b) & \mathrm{cov}(\hat{\pi}_b, \hat{\pi}_c) \\ 0 & 0 & 0 & \mathrm{cov}(\hat{\pi}_b, \hat{\pi}_c) & \mathrm{var}(\hat{\pi}_c) \end{bmatrix} \begin{bmatrix} 1 \\ \pi_b \\ \pi_c \\ \beta_{\Delta 1} \\ \beta_{\Delta 2} \end{bmatrix}$$

$$= \big[\mathrm{var}(\hat{\beta}_A) + \pi_b\,\mathrm{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 1}) + \pi_c\,\mathrm{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 2}), \mathrm{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 1}) + \pi_b\,\mathrm{var}(\hat{\beta}_{\Delta 1}), \mathrm{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 2}) + \pi_c\,\mathrm{var}(\hat{\beta}_{\Delta 2}), \beta_{\Delta 1}\,\mathrm{var}(\hat{\pi}_b) + \beta_{\Delta 2}\,\mathrm{cov}(\hat{\pi}_b, \hat{\pi}_c), \beta_{\Delta 1}\mathrm{cov}(\hat{\pi}_b, \hat{\pi}_c) + \beta_{\Delta 2}\,\mathrm{var}(\hat{\pi}_c)\big]\ [1, \pi_b, \pi_c, \beta_{\Delta 1}, \beta_{\Delta 2}]'$$

$$= \mathrm{var}(\hat{\beta}_A) + \pi_b\,\mathrm{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 1}) + \pi_c\,\mathrm{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 2}) + \pi_b\mathrm{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 1}) + \pi_{\mathrm{b}}^2\mathrm{var}(\hat{\beta}_{\Delta 1}) + \pi_c\,\mathrm{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 2}) + \pi_{\mathrm{c}}^2\mathrm{var}(\hat{\beta}_{\Delta 2}) + {\beta_{\Delta 1}}^2\mathrm{var}(\hat{\pi}_b) + 2\beta_{\Delta 2}\,\beta_{\Delta 1}\mathrm{cov}(\hat{\pi}_b, \hat{\pi}_c) + {\beta_{\Delta 2}}^2\mathrm{var}(\hat{\pi}_c)$$

$$\mathrm{var}(\hat{\beta}) = \mathrm{var}(\hat{\beta}_A + \pi_b\hat{\beta}_{\Delta 1} + \pi_c\hat{\beta}_{\Delta 2})$$

$$= \mathrm{var}(\hat{\beta}_A) + \mathrm{var}(\pi_{\mathrm{b}}\,\hat{\beta}_{\Delta 1}) + \mathrm{var}(\pi_c\hat{\beta}_{\Delta 2}) + 2\mathrm{cov}(\hat{\beta}_A, \pi_b\hat{\beta}_{\Delta 1}) + 2\mathrm{cov}(\pi_b\hat{\beta}_{\Delta 1}, \pi_c\hat{\beta}_{\Delta 2}) + 2\mathrm{cov}(\hat{\beta}_A, \pi_c\hat{\beta}_{\Delta 2})$$

$$= \mathrm{var}(\hat{\beta}_A) + \pi_{\mathrm{b}}^2\mathrm{var}(\hat{\beta}_{\Delta 1}) + \pi_{\mathrm{c}}^2\mathrm{var}(\hat{\beta}_{\Delta 2}) + 2\pi_b\mathrm{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 1}) + 2\pi_b\pi_c\mathrm{cov}(\hat{\beta}_{\Delta 1}, \hat{\beta}_{\Delta 2}) + 2\pi_c\mathrm{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 2})$$

Since the covariance between $\hat{\beta}_{\Delta 1}$ and $\hat{\beta}_{\Delta 2}$ is *0*, it follows that

$$\mathrm{var}(\hat{\beta}) = \mathrm{var}(\hat{\beta}_A) + \pi_{\mathrm{b}}^2\mathrm{var}(\hat{\beta}_{\Delta 1}) + \pi_{\mathrm{c}}^2\mathrm{var}(\hat{\beta}_{\Delta 2}) + 2\pi_b\mathrm{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 1}) + 2\pi_c\mathrm{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 2})$$

Combining $\mathrm{var}(\hat{\beta})$ and $\mathrm{V}(\widehat{\boldsymbol{\theta}})$ gives the following:

$$\left(\frac{\partial g(\boldsymbol{\theta})}{\partial \theta}\right) \mathbf{V}(\widehat{\boldsymbol{\theta}}) \left(\frac{\partial g(\boldsymbol{\theta})}{\partial \theta}\right)' = \mathrm{var}(\hat{\beta}) + {\beta_{\Delta 1}}^2\mathrm{var}(\hat{\pi}_b) + {\beta_{\Delta 2}}^2\mathrm{var}(\hat{\pi}_c) + 2\beta_{\Delta 2}\,\beta_{\Delta 1}\mathrm{cov}(\hat{\pi}_b, \hat{\pi}_c)$$

$$= \text{var}(\hat{\beta}) + \beta_{\Delta 1}{}^2 \frac{\pi_b(1-\pi_b)}{N} + \beta_{\Delta 2}{}^2 \frac{\pi_c(1-\pi_c)}{N} - 2\beta_{\Delta 2}\,\beta_{\Delta 1}\frac{\pi_b \pi_c}{N}.$$

This means that the variance can be calculated in standard statistical software such as SAS by calculating the variance for $\hat{\beta}$ using the estimate statement and then adding the contribution to the variance made by the other four terms; $\hat{\beta}_{\Delta 1}$ and $\hat{\beta}_{\Delta 2}$ are available as standard SAS output and $\pi_b$ and $\pi_c$ can be estimated from the sample. A Wald test statistic for the null hypothesis $H_0: \beta_1 = \beta_2 = \cdots = 0$ is given by

$$\beta_*' \left(\frac{\partial g(\boldsymbol{\theta})}{\partial \theta}\right) \mathbf{V}(\widehat{\boldsymbol{\theta}}) \left(\frac{\partial g(\boldsymbol{\theta})}{\partial \theta}\right)' \beta_*$$

where $\beta_* = (\beta_1, \ldots, \beta_n)'$ (Thijs et al., 2002).

Extending these methods to four groups follows the same general ideas and is given in Appendix 1.

### 3.3.3 Shared-parameter models

In shared-parameter models, a vector of latent variables or random effects, $\boldsymbol{b}_i$, is included in the joint model, where one or more components are shared between both factors in the joint distribution. The latent term or random effects capture dependence between dropout ($R$) and the response process ($Y$). The missing data process and the observed measures are independent, conditional on a common set of latent variables (random effects).

The joint distribution can be augmented with random effects

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i, \boldsymbol{b}_i | X_i, W_i, M_i, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\xi}),$$

where $\xi$ is the parameter vector describing the random effects and M is the covariates describing the random effects distribution.

The joint model can be factored either as a selection model or a pattern-mixture model. The selection model factorization is:

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i, \boldsymbol{b}_i | X_i, W_i, M_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i | X_i, \boldsymbol{b}_i, \boldsymbol{\theta}) f(\boldsymbol{r}_i | \boldsymbol{y}_i, \boldsymbol{b}_i, W_i, \boldsymbol{\psi})\, f(\boldsymbol{b}_i | M_i, \boldsymbol{\xi})$$

The pattern-mixture model factorization is:

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i, \boldsymbol{b}_i | X_i, W_i, \boldsymbol{M}_i, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\xi}) = f(\boldsymbol{y}_i | \boldsymbol{r}_i, \boldsymbol{b}_i, X_i, \boldsymbol{\theta}) f(\boldsymbol{r}_i | \boldsymbol{b}_i, W_i, \boldsymbol{\psi}) f(\boldsymbol{b}_i | M_i, \xi)$$

(Molenberghs & Kenward, 2007). A key feature of these models is that they are specified conditional on the latent term.

The shared-parameter model assumes MNAR. MAR implies that the conditional density of $R_j$ conditional on the complete data, $\boldsymbol{Y}^{obs}$, does not depend on the missing data, while MNAR implies that the conditional density of $R_j$ conditional on the complete data, $\boldsymbol{Y}^{obs}$, depends on the

missing data, $\boldsymbol{Y}^{mis}$. The following argument shows that the typical shared-parameter models are MNAR. The conditional density of $R_j$ given Y = ($\boldsymbol{Y}^{obs}, \boldsymbol{Y}^{mis}$) is

$$f\left(r_i|\boldsymbol{y}^{mis}, \boldsymbol{y}^{obs}\right) = \frac{\int g(r_i|b)g\left(\boldsymbol{y}^{mis}, \boldsymbol{y}^{obs}|b\right)h(b)db}{\int g(\boldsymbol{y}^{mis}, \boldsymbol{y}^{obs}|b)h(b)db}$$

$$= \int g\left(r_j|b\right)h\left(b|\boldsymbol{y}^{mis}, \boldsymbol{y}^{obs}\right)db$$

The conditional density depends on $\boldsymbol{Y}^{mis}$ since h(**b**|$\boldsymbol{y}^{mis}, \boldsymbol{y}^{obs}$) depends on $\boldsymbol{Y}^{mis}$. h(b|$\boldsymbol{y}^{mis}, \boldsymbol{y}^{obs}$)db can be seen as an empirical Bayes posterior distribution, which depends on the entire likelihood, including contributions due to both $\boldsymbol{Y}^{mis}$ and $\boldsymbol{Y}^{obs}$ (Albert & Follmann, 2009).

Wu and Carroll (1988) proposed latent variable models in cases such as informative right censoring, where a joint model of the continuous response variable and the time to dropout, modelled by for example a proportional hazards regression, is developed. The missingness and outcome models are linked because the same random effects are in both models. The correlations between these variables induce dependence between dropout and response. The latent variable or random effect helps to capture or account for individual to individual heterogeneity which is a significant feature in clinical trial data.

Wu and Carroll (1988) used a Gaussian random coefficient model combined with a model for time to drop out such as logistic or probit regression. They assumed the full data distribution follows a linear random effects model with random intercepts and slopes; where conditional on random effects

$$Y_{it} = b_{0i} + b_{1i}t + \varepsilon_{it}$$

where the random effects ($b_{0i}, b_{1i}$) follow a bivariate normal distribution and $\varepsilon_{it} \sim N(0, \sigma^2)$. The hazard of drop-out did not depend directly on the Y's but on the random effects

$$\text{logit}(R_{it}) = \upsilon_0 + \upsilon_1 b_{0i} + \upsilon_2 b_{1i}$$

An individual's random slope was included as a covariate in the model for the censoring process. When the regression coefficient for the random slope is not 0 there is dependence between the response and the missing data process. The conditional joint distribution of Y and R given b is structured as f(y|b)f(r|b). The marginal joint distribution is obtained by integrating against f(b). Data from both Y and R are used to model the distribution of b. The combination of probit and Gaussian response allows the integral to be solved.

A key feature of these models is that they are specified conditionally on the random effect term with the assumption that repeated measures are independent of dropout times conditional on the random effect (Hogan et al., 2004). This assumption is untestable because the outcome measurement is always missing when data are incomplete.

In shared-parameter models parameters can be estimated using different methods. One of these is maximum likelihood estimation. This can be computationally intensive, since it involves integrating over the random effects distribution. In cases where the direct integral cannot be calculated, approaches such as Monte Carlo EM or Laplace approximations of the likelihood can be used (Albert & Follmann, 2009).

Alternatively non-weighted analyses can be done, where regression models are fitted separately for each participant, and the parameter estimates are then averaged over participants. For example, for the linear mixed effects model

$$Y_{ij} = \beta_0 + \beta_1 t_j + \beta_2 x_i + \beta_3 x_i t_j + b_{0i} + b_{1i} t_j + \varepsilon_{ij}$$

we can fit a least squares regression for Y for each participant. The within group average of the slopes provide an unbiased estimate of the mean slope for each group. To compare two groups one can perform a t-test using the individually estimated slopes. This is unbiased but is inefficient compared to shared-parameter models where individual estimates are weighted according to their precision.

Lastly shared-parameter models can be approximated using conditional approaches. This has been proposed by Wu and Bailey (1989), who approximated the parameter estimates conditional on dropout with constant variance and mean given as a polynomial expression of dropout. They modelled the joint distribution of $\boldsymbol{Y_i}$ and $b_i$ conditional on , $d_{i,}$ and $\boldsymbol{x}_i$.

$$f(\boldsymbol{y}_i, \boldsymbol{b}_i | d_{i,} x_i) = g(\boldsymbol{y}_i | \boldsymbol{b}_i, d_{i,} \boldsymbol{x}_i) m(\boldsymbol{b}_i | d_{i,} \boldsymbol{x}_i).$$

Follmann and Wu (1995) and Albert and Follmann (2000) proposed similar methodology for the analysis of binary longitudinal data and repeated count data.

In all three methods, estimated variances of the parameter estimates can be calculated using bootstrap methods. In the maximum likelihood approach asymptotic variances can be obtained by the matrix of observed Fisher information (Albert & Follmann, 2009).

Shared-parameter models differ from selection models in how they relate the probability of a missing observation and the response process. Shared-random effects models link the two by relating an individual's propensity to response with propensity to miss a visit, while selection models directly model the probability of missing a visit as a function of the response. These models are particularly appropriate when the response is variable over time (Albert & Follmann, 2009).

### 3.3.4 Joint modelling of longitudinal data and time to missingness

Joint modelling of time to event and longitudinal outcomes has been done in various contexts (Henderson, Diggle, & Dobson, 2000; Tsiatis & Davidian, 2004; Tsiatis, Degruttola, & Wulfsohn, 1995). If the interest of inference is in the association between the endogenous time dependent response variable and the survival mechanism the longitudinal and survival processes have to be modelled jointly, including parameters that represent their correlation. Such models couple the survival or time to event model with a model for the repeated measurements and allow one to incorporate measurement error in the longitudinal variable into the model, which is not possible if one simply models the time to event. Joint modelling has also been used to model longitudinal markers as surrogates for survival. In this context Tsiatis et al. (1995) explored whether CD4+ count can be used as a surrogate marker for survival in AIDS patients.

Time to event or survival data is generated by observing participants until an event occurs. A complication with time to event data is that the event does not occur for some participants during the follow-up period of the study. For these participants only the maximum waiting time up to which it is known that the event has not occurred is collected. For all participants a right censored time, which is the maximum of follow-up time or the time to occurrence of the event, is recorded (Kalbfleisch & Prentice, 2002).

A joint model consists of a survival sub-model and a longitudinal sub-model. The survival sub-model is constructed as follows

$$h_i(t|\mathcal{M}_i(t), \boldsymbol{w}_i) = h_0(t)\exp[\Upsilon'\boldsymbol{w}_i + \alpha\boldsymbol{m}_i(t)]$$

where

$\mathcal{M}_i(t)$ denotes the history of the true unobserved longitudinal process up to time t

$\boldsymbol{w}_i$ is a vector of baseline covariates with a corresponding vector of regression coefficients $\Upsilon$

$h_0$(t) is the baseline risk function

α quantifies the effect of the underlying longitudinal outcome to the risk of an event

$\boldsymbol{m}_i(t)$ denotes the true and unobserved value of the longitudinal outcome at time t. This is different from $\boldsymbol{y}_i(t)$ which is the observed value of the outcome, contaminated with measurement error (Rizopoulos, 2012).

The survival function, which is defined as the probability of event time being beyond some time point, t, can be obtained as

$$S_i(t|\mathcal{M}_i(t), \boldsymbol{w}_i) = \exp(-\int_0^t h_0(s)\exp[\Upsilon'\boldsymbol{w}_i + \alpha\boldsymbol{m}_i(s)ds].$$

The survival function depends on the whole covariate history $\mathcal{M}_i(t)$ and not only on the current value of the time-dependent true value, $\boldsymbol{m}_i(t)$ (Rizopoulos, 2012).

The hazard function is the probability of an event occurring in the next short period of time, given that the event had not occurred up to that time and all the past history. In the widely used semi-parametric Cox proportional hazards model the baseline covariates are modelled parametrically while the baseline hazard function is modelled non-parametrically with no specific form and is considered a nuisance parameter. It is not possible to simultaneously estimate the baseline hazard function and the parameters of interest (Cox, 1972). Cox (1975) suggested an estimation method based on conditional probabilities at the event times, based on maximising the partial likelihood which does not depend on the baseline proportional hazard function and only the parameters of explanatory variables are estimated. Thus in a standard survival function the baseline risk function is not specified. However in the joint modelling framework the baseline hazard function, $h_0$(t), has to be specified. Common choices for $h_0$(t) are parametric distributions such as the Weibull or Gamma distributions or cubic splines or a piecewise constant model. These models can be made flexible by increasing the number of internal knots and the estimation of standard errors follows from asymptotic maximum likelihood theory (Rizopoulos, 2012).

The second component of the joint model is the participant specific longitudinal model that is specified using, for example, linear mixed models. The goal of this model is to reconstruct the complete longitudinal history, $\mathcal{M}_i(t)$, of each participant. Since **Y** is the observed outcome, which is equal to the true unobserved outcome plus error:

$$y_{ij} = m_{ij} + \varepsilon_{ij}$$

$$m_{ij} = \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{v}_i'\boldsymbol{b}_i$$

where $\varepsilon_{ij} \sim N(0, \sigma^2)$, $\boldsymbol{b}_i \sim N(0, D)$ and $\boldsymbol{x}_i$ is the design vector for the fixed effects β and $\boldsymbol{v}_i$ is the design vector for the random effects $\boldsymbol{b}_i$ (Rizopoulos, 2012). $Y_{ij}$ is only collected intermittently and the modelling is complicated by the fact that the longitudinal measurements are recorded with error. Previous attempts at incorporating the longitudinal model with the time to event model did not take this into account (Tsiatis & Davidian, 2004). Good estimates of $\mathcal{M}_i(t)$ can only be obtained if the time structure in $\boldsymbol{x}_i$ and $\boldsymbol{v}_i$ is specified correctly. If participants show non-linear longitudinal trajectories, high-dimensional functions of time can be considered, either higher-order polynomials or splines (Rizopoulos, 2012).

The assumption is made that the random effects account for both the association between the longitudinal process and the survival process, and the correlation between the repeated measurements in the longitudinal process. Thus, we assume conditional independence. This is given as

$$p(D_i, \delta_i, \boldsymbol{y}_i | \boldsymbol{b}_i, \theta) = p(D_i, \delta_i | \boldsymbol{b}_i, \theta) p(\boldsymbol{y_i} | \boldsymbol{b}_i, \theta)$$

where $D_i$ is the observed time to event, and $\delta_i$ is the event indicator (1 if the event was observed and 0 if the event was not observed) and $\theta = (\theta_t', \theta_y', \theta_b')'$ where $\theta_t$ denote the parameters for the event time outcome, $\theta_y$ denote the parameters for the longitudinal outcome and $\theta_b$ denote the unique parameters for the random-effects covariance matrix. The assumption is made that given the observed history, the censoring mechanism and the mechanism that generates times of visits are independent of the true event times and future measurements. This means that whether a participant arrives for a visit or withdraws depends on observed past longitudinal measurements, but not on additional latent participant characteristics associated with disease stage or prognosis. The observed data cannot be used to test this assumption (Rizopoulos, 2012; Tsiatis & Davidian, 2004).

The main estimation method proposed for joint models is maximum likelihood (Henderson et al., 2000). An early formulation of joint models included a longitudinal model of participant-specific random effects expressed as a linear or more general spline function of time with a parametric lognormal time to event model. They developed an EM algorithm to maximize the log-likelihood, which involved intractable integrals over the distribution of $\alpha_i$, assumed to be normally distributed (Degruttola & Tu, 1994). In general, the maximum likelihood estimates are derived as the modes of the log-likelihood function corresponding to the joint distribution of the observed outcome. Fitting joint models for longitudinal and survival data requires a combination of double numerical integration and optimization. The integral with respect to time in the survival function and the integral with respect to the random effects in the score vector do not generally have an analytic solution and a numerical approach is usually employed to approximate these integrals in the calculations of the log-likelihood and the score vector. Maximization of the log-likelihood function can be achieved using the EM algorithm, Fisher scoring, or the Newton-Raphson algorithm. The observed data score vector can be used to calculate the Hessian matrix and standard errors using an approximate observed information matrix. Fitting joint models are computationally intensive. Convergence problems are not uncommon (Rizopoulos, 2012). Bayesian estimation using MCMC techniques has been considered (Chi & Ibrahim, 2006; Wang & Taylor, 2001). A conditional score approach in which the random effects are treated as nuisance parameters was used to develop estimating equations that yield asymptotically normal estimators. In this approach no assumption was made regarding the distribution of the random effects (Tsiatis & Davidian, 2001).

The joint models discussed to this point apply mostly to the modelling of time to event taking the longitudinal covariates into account. Joint models can also be used to model incomplete longitudinal data. In this case, the occurrence of an event is defined as the time of dropout, or the time when the longitudinal process is discontinued. Intermittent missing values are treated as ignorable and assumed to be MCAR and can be ignored in the likelihood function. The

longitudinal response vector can be divided in an observed and missing part. $D_i^*$ is the true dropout time for participant $i$, while $D_i$ is the observed dropout time. The dropout mechanism is derived by the conditional distribution of the time to dropout given the complete vector of longitudinal responses

$$p\left(D_i^* \middle| \boldsymbol{y}_i^{obs}, \boldsymbol{y}_i^{mis}, \theta\right) = \int p\left(D_i^*, \boldsymbol{b}_i \middle| \boldsymbol{y}_i^{obs}, \boldsymbol{y}_i^{mis}, \theta\right) db_i$$

$$= \int p\left(D_i^* \middle| \boldsymbol{b}_i, \boldsymbol{y}_i^{obs}, \boldsymbol{y}_i^{mis}, \theta\right) p\left(\boldsymbol{b}_i \middle| \boldsymbol{y}_i^{obs}, \boldsymbol{y}_i^{mis}, \theta\right) db_i$$

Applying the conditional independence assumption, this becomes

$$\int p\left(D_i^* | \boldsymbol{b}_i, \theta\right) p\left(\boldsymbol{b}_i \middle| \boldsymbol{y}_i^{obs}, \boldsymbol{y}_i^{mis}, \theta\right) db_i$$

The time to dropout depends on $\boldsymbol{Y}_i^{mis}$ through the posterior distribution of the random effects $p\left(\boldsymbol{b}_i \middle| \boldsymbol{y}_i^{obs}, \boldsymbol{y}_i^{mis}, \theta\right)$. Under the joint model the survival and longitudinal sub-models share the same random effects. Therefore joint models fall under the shared-parameter model discussed in Section 3.3.3. Joint models correspond to a MNAR missing data mechanism. Under a simple random effects structure this missing data mechanism implies, for example, that participants who show steep increases in their longitudinal profiles may be more or less likely to drop out (Rizopoulos, 2012).

The joint model of the survival and longitudinal sub-models is

$$h_i(t) = h_0(t)\exp[\Upsilon' \boldsymbol{w}_i + \alpha[\boldsymbol{x}_i'(t)\beta + \boldsymbol{v}_i'(t)\boldsymbol{b}_i]].$$

There is a connection between the association parameter α and the missing data mechanism. If α = 0 then the missing data mechanism is MCAR. In this instance $h_i(t) = h_0(t)\exp[\Upsilon' \boldsymbol{w}_i]$ and the dropout process does not depend on either the missing or observed longitudinal responses. Censoring corresponds to a MAR missing data mechanism because in the formulation of the likelihood function of joint models it was assumed that the censoring mechanism depends on the observed history but is independent of future unobserved outcomes (Rizopoulos, 2012).

If α = 0 the parameters in the two sub-models are distinct and their corresponding parameters can be estimated separately. Under a full likelihood approach, the estimated parameters derived from maximizing the log-likelihood of the longitudinal process will also be valid under a MAR missing data mechanism where the dropout depends on the observed responses only. An advantage of shared-parameter models is that these models can handle both intermittent missingness and monotone missingness. Many of the pattern-mixture and selection models have difficulty handling non-monotone missing data (Rizopoulos, 2012).

Sensitivity analysis within this framework consists of fitting several joint models where different alterations are made in the survival sub-model. Alternative parameterization of the longitudinal marker can also be considered.

## 3.4 Likelihood-based approach

Likelihood-based analysis is a viable approach when analysing incomplete longitudinal data when the MAR assumption is plausible. All available cases are used in a likelihood-based way using ignorability theory. Ignorability has been defined in Section 2.4; and means that the mild separability assumption is made that the parameters of the missing data mechanism are distinct from the parameters of the sampling model. This implies that the mechanism generating missing data can be ignored when the interest is in inference about the measurement process, without biasing the analysis. Ignoring the missing data mechanism assumes there is no scientific interest in the missing data mechanism.

The following desirable properties are associated with maximum likelihood analyses when the assumptions are met: consistency, asymptotic efficiency and asymptotic normality, under broadly applicable regularity conditions. Consistency implies that the estimates are progressively less biased and less variable given a large sample. Asymptotic efficiency implies that the estimates have minimum standard errors and asymptotic normality enables the user to use normal approximations when calculating confidence intervals and p-values. These desirable properties are all large sample approximations (Allison, 2009).

Likelihood-based approaches use a parametric model to formulate a statistical mechanism for the missing data and base inference on the likelihood function of the incomplete data. The objective is to draw inference about a parameter, $\theta$, in a model $f\ (\boldsymbol{y}|\theta)$ for the response data that is not fully observed. Under the MAR assumption, $\theta$ and the missing data model are functionally independent and missing data can be treated as ignorable. In this case inference is drawn about $\theta$ without having to specify a model that relates the missing data process to the *Y* or *X* (National Research Council, 2010). Information from the non-missing data is used to provide information about the missing data. Missing data are not imputed (Mallinckrodt, Clark, et al., 2003).

In the absence of missing data, likelihood-based methods entail the maximization of the full data likelihood. When data are incomplete this likelihood is replaced by the observed data likelihood, where the individual likelihoods are integrated over the missing values,

$$\prod_{i=1}^{N} \int f\ (\boldsymbol{y}_i^{\text{obs}}, \boldsymbol{y}_i^{\text{mis}}, \boldsymbol{r}_i|\boldsymbol{\theta}, \boldsymbol{\psi})\text{d}\boldsymbol{y}_i^{mis}.$$

Under ignorability and MCAR or MAR missingness, the integral can be rewritten as an integral over the missing values and the distribution of the missing data mechanism (under a selection model). This simplifies the integral tremendously.

If the MAR assumption holds, ignorability is assumed and the sample is relatively large, maximum likelihood theory can be used to estimate θ. Under regularity conditions $\hat{\theta}$ has a normal distribution with mean $\theta$ and variance estimated by the inverse of the observed information matrix $I^{-1}(\hat{\theta})$ or using bootstrap methods (Kenward & Molenberghs, 1998; National Research Council, 2010).

Sometimes numerical approximation is needed to maximize the likelihood. The EM algorithm, a general-purpose iterative algorithm, can be used for calculating maximum likelihood estimates in these instances (Dempster, Laird, & Rubin, 1977; Horton & Kleinman, 2007; Little & Rubin, 2002; Molenberghs & Kenward, 2007).

Likelihood-based estimation adjusts data in terms of the conditional expectation of the unobserved measurements given the observed measurements. Thus a likelihood-based ignorable analysis accommodates information on a participant with post-randomisation outcomes, even with missing data. An ignorable likelihood is therefore consistent with the ITT analysis, provided that the treatment compliance is the same for those who drop out and those who remain in the study (Molenberghs & Kenward, 2007; Molenberghs et al., 2004). Likelihood-based methods are easy to implement because no data manipulation is required to accommodate the missing data.

If data sets have no missing data, the estimate of the mean in a saturated means multivariate normal model does not depend on the specification of the variance matrix. However, with missing data, the inference of the mean depends on the specification of the variance matrix (National Research Council, 2010).

Likelihood-based methods treat longitudinal data as clustered data that are temporally aligned, assuming that the missing data are orthogonal to the missingness process given observed data, regardless of where the dropout occurs. Under MAR the likelihood-based approach requires correct specification of the full data model (Hogan et al., 2004).

One application of direct likelihood methods is through mixed models. A mixed model is a model that includes both fixed and random effects. The most common random effects model is the generalised linear mixed model which combines concepts from generalised linear models and linear mixed models (Molenberghs & Kenward, 2007). Generalised linear mixed models assume a

random component, the components of *Y*, which is independent from each other and follows some distribution with constant variance of errors; a systematic component, a linear prediction using the covariates and $\boldsymbol{\beta}$; and a link function between the random and systematic components. This generalises and extends linear models by allowing the distribution of the systematic component to be from an exponential family other than the normal distribution and the link function can become any monotonic differentiable function. Classical linear models assume a Gaussian distribution for the components of Y and the identity function as a link function (McCullagh & Nelder, 1989).

If $\boldsymbol{Y}_i$ is the $n_i$-dimensional vector of all measurements for participant $i$, then $Y_{ij}$ can be modelled using, for example, a random intercept and slope model as

$$Y_{ij} = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})t_{ij} + \beta_2 z_i + \beta_3 t_{ij} z_i + \varepsilon_{ij} \qquad \text{(Equation 3)}$$

$$\boldsymbol{b}_i = (b_{i0}, b_{i1})' \sim N(0, \boldsymbol{G}),\ \boldsymbol{\varepsilon}_i \sim N(0, \Sigma_i),$$

$\boldsymbol{b}_1, \ldots, \boldsymbol{b}_N$ and $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_N$ are independent, where $z_i$ is the treatment group and $t_{ij}$ is a time variable.

The components in $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)'$ are fixed effects and the components in $\boldsymbol{b}_i$ are random effects (Molenberghs & Kenward, 2007). The fixed effect of greatest interest in clinical trials is generally treatment allocation. Additional fixed effects such as baseline covariates, demographic data or site can also be included (Mallinckrodt, Sanger, et al., 2003). The participants in a study are thought of as representative of a larger population of participants, therefore, the effects specific to an individual, $\boldsymbol{b}_i$, are treated as random effects (Hedeker & Mermelstein, 2007).

The random effects approach is based on the assumption that for every participant the response can be modelled by a linear regression model with participant specific regression coefficients. These analyses use likelihood-based estimates and participant specific effects and correlations between the repeated measurements are modelled via the within participant error correlation structure (Mallinckrodt, Sanger, et al., 2003). The random participant effects included in the model account for the correlation between the repeated measures of a participant. These random effects reflect the change of each participant over time and explain the correlational structure of the longitudinal data. They indicate the degree of variation per participant that exist in the population (Hedeker & Mermelstein, 2007).

The error terms are assumed to be normally distributed and independent conditional on the random individual-specific effects, $\boldsymbol{b}_i$. This model represents the measurement of Y as a function of time, at the individual and population levels. The overall population intercept is given by $\beta_0$, the random slope for participant *i* is given by $b_{i0}$, and $\beta_1$ is the overall population slope (the effect of time). The intercept parameters indicate the starting point and the slope parameters indicate the degree of

change over time. The population intercept and slope parameters represent the trend for the population and the individual parameters show how the individual differs from the population trend. Equation 3 includes a time by treatment interaction and can also be expanded to include other interactions, higher order polynomials or time varying coefficients (Hedeker & Mermelstein, 2007).

Random-effects models are useful in the longitudinal setting because participants are not required to be measured on the same number of time points (the number of data points per individual is not assumed to be balanced) and time is treated as a continuous variable. Participants with incomplete data or designs where follow-up times are not uniform across participants can thus be included. The $\boldsymbol{Y}_i$ vector and the $X_i$ and $Z_i$ matrices carry the $i$ subscript, thus no assumption of complete data on the response across time points is made (Hedeker & Gibbons, 1997; Hedeker & Mermelstein, 2007). The change in the response variable for each participant can be estimated. The advantage of these mixed models is that the predictors of both intra-individual (within-participant) and inter-individual (between-participants) variation can be assessed (Hedeker & Mermelstein, 2007).

The model assumes that the available data for a participant represents that participant's deviation from the average trend observed for the whole sample. The model estimates the participant's trend across time on the basis of the observed data for the participant and the time trend estimated for the whole sample, adjusted for covariates (Hedeker & Gibbons, 1997).

Mixed effects models consider the unique attributes of each participant. Responses of individual participants are allowed to vary and to be correlated over time. Information from the observed outcomes of a participant can be used to provide information about the unobserved outcomes. All available data are used to compensate for the data missing on a particular participant (Mallinckrodt, Sanger, et al., 2003).

Estimation and inference is based on maximum likelihood. To determine the significance of model parameters, large sample variances and covariances of the maximum likelihood estimates are obtained. This is used to construct confidence intervals and conduct Wald tests for the model parameters (Hedeker & Gibbons, 1997). In general no analytic solution exists and numerical maximisation routines are used. In practice often REML estimation is used (Laird & Ware, 1982). This leads to smaller bias in the estimation of the variance component than maximum likelihood (Molenberghs & Kenward, 2007) because the method accounts for degrees of freedom lost in estimating the mean model.

When the interest is on inference on the fixed effects $\boldsymbol{\beta}$, *t* and *F* distributions can be used. The denominator degrees of freedom are estimated with the method of Kenward and Roger (1997). When interest lies in inference for some of the variance components classical Wald, likelihood ratio and score tests can be used. If the interest lies in the random effects empirical Bayesian estimation methods can be used (Molenberghs & Kenward, 2007). Inference is based on the conditional distribution for $\boldsymbol{Y}_i$ which is obtained by integrating out the random effects. The resulting likelihood for the joint distribution is maximised in the presence of N integrals over the q-dimensional random effects.

In order to fit a mixed effects linear model to a data set with missing data, a few assumptions are needed. The assumptions made include:

1. Because a normal outcome is assumed, multivariate normality ought to apply.
2. To avoid overly restrictive, unverifiable assumptions, it is sensible to select a saturated means model that include time, treatment and time by treatment interaction. In the same spirit, an unstructured variance-covariance matrix reduces the risk of model misspecification and, consequently, incorrect inferences.
3. The joint statistical behaviour of the unobserved measurements from an individual who drops out is assumed to be the same as that of an individual who does not drop out who shares the same history (previous measurements, including baseline) and the same covariates (including treatment group). If dropout means stopping treatment the assumption that these two groups behave the same is not true (Kenward, 2006).

The choice of the structure of the covariance matrix influences the point estimates when the data are incomplete. An unstructured covariance matrix is often the appropriate first choice. Carpenter and Kenward (2007) showed that using the unstructured covariance matrix does not lead to a noticeable loss of power if the sample size is not very small. A different covariance matrix can also be used in each treatment group.

Why does direct likelihood, via mixed models, provide a valid analysis when data are incomplete? It takes into account the expectation of the missing measurements, given the observed measurements and is valid and unbiased under MAR (Molenberghs & Kenward, 2007). It can be thought of as aiming to estimate the treatment effect that would have been observed if all participants had continued on treatment for the full study duration (CHMP, 2010). The CHMP also believe that these methods would overestimate the treatment effect of effective interventions, thus biasing results in favour of the treatment effect and should only be used when missing data are negligible. Mallinckrodt et al. (2003) showed that mixed model analysis was robust to the presence of MNAR data. Thus this method is appropriate to use with MCAR and MAR data, and could, in

some circumstances be appropriate with MNAR data. In some settings a biased MAR analysis will be less biased than a MCAR analysis.

Likelihood-based methods work in the context of missing data by adjusting the least square means for a participant at times after drop out. The magnitude of the adjustment is determined by the within participant correlation and by the participant's deviation from the group mean. The uncertainty in the adjustment is determined by the amount of data contributing to the group mean, the amount of data on the participant and the within-participant correlation (Mallinckrodt, Clark, et al., 2003).

Although mixed effects analysis uses all data collected on participants who completed and did not complete the study treatment, if data were only collected while participants were on treatment then the analysis estimates what would have happened if all participants with missing data had stayed in the study and completed treatment. This is not a strict ITT analysis, since it does not account for withdrawal from treatment. Continuing follow-up and data collection after discontinuation of treatment could alleviate this problem. However, as long as there are missing data off treatment, the likelihood-based methods will represent an analysis on treatment more closely than a strict ITT analysis (Keene, 2011).

Likelihood-based methods are the easiest methods to use and can be applied with most standard software, without much knowledge of missing data analysis. These methods should be encouraged, while understanding that they may not provide a strict ITT analysis, if treatment compliance of those who drop out is not the same as compliance of those who remain.

## 3.5 Weighting methods

If data are MAR a missingness weight can be assigned to the complete cases. This implies that the probability of being observed is a function of the variables measured (National Research Council, 2010). In this approach, also called inverse probability weighting, a model for the probability of missingness is fitted, and the inverse of these probabilities are used as weights for the complete case (Horton & Kleinman, 2007; Molenberghs & Kenward, 2007). Weighting by the inverse of the probability of being observed means that participants who were observed but had low probability of being observed, are given more weight and as such also represent the unobserved participants. This way, the complete cases are used to estimate parameters that are valid for both the complete cases and missing cases.

This is explained by starting with the univariate case. In the univariate case inverse probability weighting is done by assigning sampling weights to the complete cases. For example, in a survey,

if only some observations are observed a weighted average, where each observed observation is weighted by the inverse of the probability of response, would be unbiased, whereas the unweighted average would not be unbiased.

This idea is applied to regression with missing observations. If all values of the covariate are observed, then the true population parameter values will have an expected value of *0*. This is called a consistent parameter estimate. If some of the covariates are missing then the parameter estimates are no longer consistent and the expected value of the estimating equations is no longer *0*. If each covariate is observed with probability $\pi_i$, then weighting by $\frac{1}{\pi_i}$ creates estimating equations with expectation 0 and consistent parameter estimates. Often $\pi_i$ is estimated using logistic regression, after which a weighted regression will be performed with weights $\frac{1}{\pi_i}$ (Carpenter, Kenward, & Vansteelandt, 2006).

For example, an inverse probability weighting estimate of the mean is computed by first fitting a model for the probability of being observed;

$$\pi(\boldsymbol{y}_i, \boldsymbol{\theta}) = P(\mathrm{R}_i = 1 | \boldsymbol{x}_i, \boldsymbol{\theta}).$$

An indicator of whether an observation was observed is regressed against observed characteristics of the participant. The mean of *Y* is then estimated using the weighted average, the average of the observed *Y* weighted inversely by the probability of being observed.

$$\hat{\mu} = \frac{1}{n} \sum_i \frac{R_i \boldsymbol{y}_i}{\pi(\boldsymbol{x}_i, \hat{\theta})}$$

The correlation structure should ideally be the independence working correlation structure and standard errors can be estimated using bootstrap methods (National Research Council, 2010).

Considering longitudinal data with repeated measures, when there is no missing data repeated measures regression models can be fitted using GEE. For repeated measures with missing data GEE does not produce consistent and valid estimates unless the missingness is MCAR because regression estimation depend strongly on the assumed correlation structure. Weighted GEE can be used when the missingness mechanism is MAR and monotone. Echoing the ideas already discussed in the univariate case, the weight is obtained from the inverse probability of a participant being observed at that particular measurement occasion, given that the participant is still in the study, covariates and previous measurements on the response variable (Molenberghs & Kenward, 2007). The probability that participant *i,* who is still in the study at time *t – 1*, will be observed at time *t* is $\lambda_{it}$. The probability that participant *i* is still in the study is calculated using cumulative conditional probabilities;

$$\text{At t = 0: } \lambda_{i0} = 1$$

$$\text{at t = 1: } \pi_{i1} = \lambda_{i1}$$

$$\text{at t = 2, } \pi_{i2} = \lambda_{i1}\lambda_{i2}$$

$$\text{at t = 3, } \pi_{i3} = \lambda_{i1}\lambda_{i2}\lambda_{i3}$$

Additional assumptions underlying inverse probability weighting are that there are no covariate profiles where *Y* could not be observed. The possible values of missing responses are the same as the possible values of observed responses. Thus missing values cannot be imputed outside the range of observed values (National Research Council, 2010).

The procedure to use inverse probability weighting is as follows: Ensure that the missing data follow a monotone pattern. Additional information on covariates collected should be included when the probability of observing data is modelled. Calculate the probability that the outcome is observed. Fit the regression that would have been fitted had all data been fully observed, but weight individual contributions to the model by the inverse of the probability of being observed (Carpenter et al., 2006; National Research Council, 2010).

Given a large sample and MAR missingness, the inverse probability weighted GEE yields unbiased estimates when the probability of observing data is correctly specified. A disadvantage of this method is that the model for the probability of being observed has to be correctly specified for the estimator to be unbiased. If the probabilities are calculated incorrectly the observations are not weighted correctly and will not represent the missing data. The parameter estimates are inefficient relative to likelihood-based estimates (National Research Council, 2010).

Observations with very small probability of being observed will have large inverse weights. These observations will then dominate the method and lead to instabilities and high variance in small samples. The augmented inverse probability weighting estimator improves on this, since the variance is not as high when individual weights are high. Strata can be created based on the predicted probability of being complete. Respondents are then weighted by the inverse of the response rate in these strata. This can be controlled by choosing strata that limit the size of the weights, thereby controlling the variance. The theoretical justification of this method is based on large-sample arguments and it may not hold in small samples (National Research Council, 2010).

Weighted estimating equations are robust, since it does not depend on knowledge of the distribution of the unobserved data, and the estimates are consistent and asymptotically normal. It can be applied in settings where other approaches are difficult to apply. Other methods may be optimal when the distributional assumptions are correct, but they may not be desirable when the assumptions are violated. The estimates are also potentially inefficient, since it uses only data from

completers (Ibrahim et al., 2005). In addition to the methods that use data from completers, Robins et al (1994; Robins, Rotnitzky, & Zhao, 1995) and others (Carpenter et al., 2006; Vansteelandt, Carpenter, & Kenward, 2010) extended the methods by including incomplete cases in the calculation of the weights. This improved efficiency.

Augmented inverse probability weighting estimators are obtained by adding to the inverse probability weighting estimating functions an augmentation term that depends on unknown functions of the observed data. In this way the augmented inverse probability weighting GEE procedure does not use information from complete cases only. If these functions are appropriately chosen, it leads to efficiency.

The augmented inverse probability weighting estimator of $\mu = E(Y_k)$ is:

$$\hat{\mu} = \frac{1}{n}\sum_i \frac{R_{ik} y_{ik}}{\pi_i(z_{ik};\hat{\theta})} + \frac{1}{n}\sum_i \left\{\frac{R_{ik}}{\pi_i(z_{ik};\hat{\theta})} - 1\right\} g(z_{ik}^{-}, x_i),$$

where $R_{ik} = 1$ if $Y_{ik}$ is observed and 0 if $Y_{ik}$ is not observed. The observed history at time k is denoted as $z_k^-$, $\pi(Z_k;\theta) = P(R_k = 1|z_k,\theta)$ is the probability that a participant remained in the study to time k and $g(z_{ik}^-, x_i)$ is a function of the observed history up to k-1 (National Research Council, 2010).

The first term in this estimator is the inverse probability weighting estimator of μ, which weights the observed measurements. The second term is the augmented term added and has mean 0 because the expectation of $R_{ik}$ is $\pi_i(z_{ik};\hat{\theta})$, so that this is still a consistent estimator. This term includes information on both fully observed participants and participants with missing values. The variance will be determined by $g(Z_{ik}^-, X_i)$ which can be chosen to minimise the variance. The precise form for this expression is not known, but it can be estimated using participants with observed outcomes. If g is correct, then the augmented inverse probability weighting estimator is more efficient than the standard inverse probability weighting estimator (Molenberghs & Kenward, 2007; National Research Council, 2010). Robins and Rotnitzky (1995) described the calculation of this function. Thus augmented inverse probability weighting estimators are obtained by augmenting the estimating function with a term that depends on a function of the observed data. If these functions are chosen correctly the efficiency of the standard weighted GEE can be improved.

Double robust estimation combines inverse probability weighting with regression modelling. Each observation is weighted equal to the inverse of the probability of being observed to create pseudo-populations of complete and incomplete participants that represent what would have happened to the population under those two conditions. Maximum likelihood regression or GEE is conducted

within these pseudo-populations adjusting for confounders and variables of interest. An estimator is doubly robust if it is consistent when either a model for the probability of being observed or a model for the distribution of the complete data is correctly specified or if both are correctly specified. Because one cannot know whether the model for missingness is correctly specified, this is a highly desirable property (Molenberghs & Kenward, 2007; Vansteelandt et al., 2010).

Such a model can be obtained by replacing the observed value in the substantive model with the fitted value of a generalised linear model analysis of the outcome on background characteristics fitted to the responders using the weights $1/\pi_i$. The doubly robust nature of the estimator can be seen as follows. When the imputation model is correctly specified, the misspecification of the weights does not influence the expected value of the estimator. The estimator can be written as a set of equations including a product of the model residuals and the weights, and since the expected value of the model residuals is 0, the term would be 0, regardless of the weights and therefore gives an unbiased estimator. If the weights are correctly calculated (the response model is correctly specified), then misspecification of the imputation model does not affect the validity of the doubly robust estimator because the weighted least squares estimator satisfies

$$\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi_i}Y_i = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi_i}m^*(X_i)$$

where $m^*(X_i)$ is the fitted value for participant $i$. If the weights are correctly specified then $\frac{R_i}{\pi_i}$ has an expected value of 1. This implies that $m^*(X_i) = Y_i$ in expectation (Vansteelandt et al., 2010).

A doubly robust estimator offers the advantage over standard inverse probability estimators or likelihood-based estimators that there are two chances to make correct inference. These methods do assume that the substantive model is correctly specified. The doubly robust estimators are less efficient than maximum likelihood estimators when the likelihood is correctly specified, but they are more robust to incorrect specification of the model. Doubly robust estimators can be difficult to construct, or might not even exist in some instances (Bang & Robins, 2005; Robins et al., 1994). These methods assume a MAR missingness mechanism (Vansteelandt et al., 2010).

The standard errors cannot be calculated using standard methods, because with inverse probability weighting the standard errors ignore the imprecision of the estimated weights. One solution is to calculate the standard errors using bootstrap techniques. Bootstrap may not perform well when some individuals are oversampled in some of the bootstrap samples. An alternative is to consider sandwich estimators. These may not perform well with larger weights because they are based on large sample approximations that might not hold when some individuals have large weights (Vansteelandt et al., 2010).

## 3.6 Bayesian approaches

Ibrahim et al. (2005) argued that maximum likelihood and multiple imputation have Bayesian connections. Multiple imputation was originally derived from a Bayesian framework where the sampling in the imputation step is done from the posterior distribution of interest. Maximum likelihood can be viewed as a large-sample Bayesian method, with non-informative priors.

Bayesian approaches have also been applied more generally in the analysis of missing data. This is done by specifying priors on all the parameters and specifying distributions for the missing covariates (Horton & Kleinman, 2007). The missing data are then sampled from their conditional distribution via the Gibbs sampler. The Gibbs sampler is an algorithm that involves sampling a Markov chain where the kernel is the product of the sequentially updated full conditional distributions of the parameters and the stationary distribution is the posterior distribution (Geman & Geman, 1984). Fully Bayesian methods are powerful and general methods for dealing with missing data since existing Bayesian methods can be adapted for dealing with missing data.

In classical, frequentist statistical analysis parameters are regarded to be fixed non-random quantities. Probability statements made concern the data observed. In Bayesian analysis the parameters are treated as realisations of random variables. Probability statements are then made about the model parameters and not about the data. Bayesian analyses have three components; the prior distribution, the likelihood and the posterior distribution. The prior distribution reflects the distribution of the parameters before the data are seen; thus expressing uncertainty about the parameters prior to seeing the data. The likelihood gives the distribution of the observed data. This is the same distribution that would be used in classical frequentist inference. The posterior distribution uses Bayes' Theorem to combine information from the prior distribution and the likelihood. The posterior distribution expresses uncertainty about the unknown parameters after seeing the data. In ignorable methods the posterior distribution is $p(\theta|D) \propto p(\theta)L(\theta|D)$, with $p(\theta)$ the prior distribution and $L(\theta|D)$ the full data likelihood. Thus the Bayesian inference is done by specifying a model, specifying prior distributions for the parameters of the model, and then updating the prior information on the parameters using the model specified and the data observed to obtain the posterior distribution of the parameters. Our assumptions about the missing data can thus be made explicit through the prior distributions (Daniels & Hogan, 2008).

Bayesian inference distinguishes between observable quantities, the data, and unobservable quantities, such as the statistical parameters and the missing data. The unobserved quantities have an associated probability distribution. Thus Bayesian methods are a natural way of handling

missing data because a probability distribution is estimated for each missing value. This also allows for uncertainty to be captured. Missing data are treated as unknown parameters.

Prior distributions quantify *a priori* knowledge about the parameter(s). An informative prior is a prior that contains information through a probability distribution. The prior distribution can come either from previous studies, historical information or expert opinion. If this is not available a vague prior can be used if no information is available. These are called non-informative priors. Specifying the prior is important and can influence the results. The Bayesian approach allows assumptions about non-identified parameters to be formalised through prior distributions (Daniels & Hogan, 2008). The sensitivity of the results to the prior used should always be explored.

To conduct the fully Bayesian analysis on the observed data a posteriori the following steps are followed:

1. The distribution of the data containing the missing data is specified.
2. The joint prior distribution is specified.
3. Sample from the posterior using simulations methods like the Gibbs sampler (Ibrahim et al., 2005).

With complete data, when informative priors are used, information from the prior is combined with information from the data to create the posterior. For example in the case of the normal linear regression model, the posterior mean of $\beta$ is a weighted average of the ordinary least square estimator and the prior mean. The weights depend on the data and the prior variance. The strength and informativeness of the prior is determined by the prior variance. With incomplete data some information only derives from the prior (Daniels & Hogan, 2008).

It is not easy to derive the parameters of the prior distribution from experts. One suggestion is to rather ask experts to predict future data or data patterns and to use this to model the data and derive parameters from these models. This is done because it is easier for experts to visualise data than to visualise parameters (Daniels & Hogan, 2008).

When there are incomplete data, a prior belief is assigned to the distribution of the missing data points. The prior is assumed to be independent of the data. If covariates are observed and only the response has missing data, one need not assign a missingness model, provided the missing data mechanism is ignorable. There is no fundamental distinction in the handling of missing data and unknown parameters within the Bayesian framework. The missing data are then treated as additional unknown parameters and the imputation of the missing outcome variables is unnecessary for a valid inference about the model parameters. If imputation of the missing data is needed,

values can be generated from the posterior predictive distribution. However, if the assumption that the missing data mechanism is ignorable cannot be made, then we need to specify a response model of missingness. If covariates are missing an imputation model for the missing data is required. All that is needed is the specification of the appropriate joint model for the observed and missing data and model parameters. Posterior samples of the model parameters and missing values are then generated in the usual way using MCMC (Mason et al., 2012).

The posterior distribution is the basis of Bayesian inference, but for complex models the integration is not possible analytically. MCMC is done by sampling from the joint posterior distribution in cases where analytical integration is not possible (Gilks, Richardson, & Spiegelhalter, 1996). Sound inference based on MCMC methods requires that the Markov chain being simulated has achieved steady state or converged. Convergence is checked by running multiple chains with different starting points and checking whether the chains converge to the same point.

Inference based on serially correlated MCMC draws is less precise than if the draws were independent. In Bayesian analysis assessing the fit of the model includes checking for sensitivity of the analysis to the prior specified, or conflict between the prior and the data observed. Posterior predictive model checking is also done, in which the simulated data from the posterior predictive distribution is compared to the observed data. If the model fits well the simulated data should be similar to the observed data (Gelman et al., 2004). The Deviance Information Criterion (DIC) can be used to compare models in a similar way as the Akaike Information Criterion, with smaller values indicating better model fit. It gives a measure of model fit, that is penalised for model complexity (Spiegelhalter et al., 2002). When data are incomplete the DIC is based on the fit of the joint model $f(\boldsymbol{y}, \boldsymbol{r}|\theta, \psi)$ to the observed data ($\mathbf{Y}^{\text{obs}}$ and $\boldsymbol{R}$). The DIC indicates how consistent the modelling assumptions are with the observed data, but cannot provide information about the fit to the data that was not observed (Daniels & Hogan, 2008).

Bayesian analysis provides a flexible way to model MNAR data, using a selection model factorisation of a joint model, consisting of a model of interest and a model of missingness. The results are not robust to incorrect specification of the various parts of the joint model.

The advantage of Bayesian modelling is that it provides a model based approach to missing data, which is theoretically sound. Uncertainty can be modelled. These models also have the ability to incorporate realistic assumptions about the reasons for missing data.

## 3.7 Death as a cause of missing data

One of the more difficult issues in the analysis of longitudinal data is how to handle death as a cause of dropout and thus a cause of missing data. Little has been written about addressing missing data when deaths occur during follow-up and is related to the outcome of interest. The National Research Council (2010) recommended that one define the estimand carefully; for example as outcomes in those who would remain alive on either treatment. This estimand applies to a sub-group that cannot necessarily be identified.

Most methods for the analysis of data after dropout assume that participants who drop out could have been measured after their dropout time. Random-effects models or multiple imputation may impute data beyond dropout; this is not plausible if drop out was caused by death (Hogan et al., 2004). A possible way to approach this is to draw inferences about the sub-population of participants who would survive (Robins, Greenland, & Hu, 1999; Rubin & Frangakis, 1999).

Regression models can be used to describe the relationship between predictors and the longitudinal response, but survival influences whether the longitudinal response will be observed at later visits. Thus if one modelled longitudinal data truncated by death, the model will explicitly or implicitly also include survival. Kurland et al. (2009) discuss several modelling options for longitudinal data truncated by death. These methods answer different scientific questions and also have different estimands.

The first type of methods models the outcome unconditional on survival. These models are appropriate if deaths do not occur or are independent of the response process or do not result in truncation. When these conditions are not met, the unconditional distribution averages $f(\boldsymbol{y}_i|s_i)$ over the survival function $f(s_i)$, where $S_i$ represents the survival time for participant $i$. These unconditional models are usually not the correct model to use when there is imbalance due to death (Kurland et al., 2009). Mixed effects models are an example of these and may impute data beyond death.

A second direction one could take is through models that are fully conditional on death. Only participants who survived are analysed, or those who survived are analysed separately from those who did not survive. One way to do a fully conditional analysis is with pattern-mixture models, where outcome is modelled separately for groups defined by the survival time. These models can model the individual trajectories correctly, but they use future survival time to model earlier outcomes. Principal stratification as described by Frangakis and Rubin (2002) can also be used, where causal effects are estimated for principal strata as defined by the potential survival outcome.

These models describe the causal treatment effect for each stratum. Terminal decline models are a third fully conditional option. These models include only the participants who died during the study and the analysis counts backwards from death (Kurland et al., 2009).

The third direction one could take is to use partly conditional models. These can be fitted by conditioning the expected value of the response at time $t_{ij}$ on the participant being alive at time $t_{ij}$. Partly conditional regression models correct the shortcoming that data are imputed after death by methods that model the correlation structure of longitudinal data, by assuming independence among the longitudinal responses. These models then fit regression models conditional on being alive and describe the outcome variables among the dynamic cohort of surviving participants (Kurland et al., 2009).

The last method to handle missing data created by death is to fit a joint model of both the outcome and survival. A joint model of the probability of being alive and healthy is created. Joint models were discussed in detail in Section 3.3.4.

Dufouil, Brayne, and Clayton (2004) argued that death and dropout refer to two different kinds of loss to follow-up and should not simply be combined in analysis. They also distinguished between a mortal and immortal cohort. In an immortal cohort, all participants are observed at all time points. Analyses based on likelihood give correct results when the missing observations are MAR, but full probability modelling of participant-specific response profiles does not readily accommodate different treatment of death and drop-out. Marginal modelling approaches, such as GEE, can accommodate this; however GEE is only valid under MCAR. Inverse probability weighting methods can solve this problem, and weighted GEE is valid under MAR.

Dufouil et al. (2004) suggested imputing values for drop outs and allowing participants who died to be removed from the study. They assumed that the mortality rate following drop-out was the same as the mortality rate for participants remaining in the study. They defined $\pi_{it}$ as the probability that participant *I,* seen at time $t - 1$ and still alive at time *t,* will drop out at time *t*. Using ideas from inverse probability weighting they fit a logistic regression using data from the survivors at each wave to model $\pi_{it}$ and then use double decrement life table models to calculate the weights for inverse probability weighting. They then simulate imputation of missing data for participants who dropped out, but not for those who died. This analysis assumes that following drop out the distribution of outcome in surviving participants mirrors that in participants remaining in the study, conditional on observed data. If the interest is to predict the longitudinal outcome in living participants, then the mortal cohort analysis is appropriate. This method can only be used with monotone missing data.

## 3.8 Concluding points

Several methods to analyse longitudinal data when data are incomplete were discussed in this chapter. These methods require assumptions made about the distribution of the missing data, often the conditional distribution of the missing data given the observed data, or the mechanism that created the missing data. The resulting conclusions can be sensitive to the assumptions made (Ibrahim et al., 2005; Molenberghs & Kenward, 2007; National Research Council, 2010). The validity of these assumptions can, however, not be tested using the available data.

The various methods discussed all have advantages and disadvantages. The mechanism creating the missing data (MCAR, MAR, MNAR) determines to a large extent which of the analysis methods are appropriate in any specific instance. If the missing data mechanism is MCAR almost any method would be appropriate, LOCF being a counterexample. The complete case analysis is only unbiased if the data are MCAR. Under MAR, inverse probability weighting, likelihood-based approaches, Bayesian methods and multiple data imputation are used and valid, given some assumptions. Under MNAR conditions selection models, pattern-mixture models and shared-parameter models are used.

The parameters estimated using a pattern-mixture model, selection model or shared-parameter model cannot be compared directly. This is because $\theta$ in a selection model represents a marginal effect, in a pattern-mixture model it represents an effect conditional on missing data and in a shared-parameter model it measures an effect conditional on a latent variable. However, the estimate obtained from a pattern-mixture model can be averaged to obtain a marginal estimate, which is comparable to a selection model estimate.

Methods that do not assume MAR widen confidence or credible intervals because they inflate the standard error to reflect the uncertainty created by the missing data (White et al., 2007). Certain methods are almost never appropriate. These include single imputation methods, such as LOCF. The analysis of a clinical trial should account for the uncertainty created by missing data. The significant tests should have valid type I error rates and standard errors should be calculated appropriately. This is accomplished through multiple imputing, inverse probability weighting, Bayesian methods and maximum likelihood methods.

Likelihood-based estimation in an incomplete multivariate setting involves adjustment in terms of the conditional expectation of the unobserved measurements given the observed ones. This is therefore a proper way to use the available post randomisation outcomes, even when some data are missing (Molenberghs & Kenward, 2007). Mallinckrodt et al. (2003) proposed that likelihood-

based mixed effects models are the appropriate choice for the primary analysis; since these resolved the problems posed to the greatest extent. However, the mean structure and the variance-covariance structure need to be correctly modelled.

Regulatory agencies used to ask for a single, pre-specified primary analysis. This is difficult to implement with incomplete data and several sensitivity analyses. MNAR methods are not simple, and in a sensitivity analysis context do not lead to a single analysis, nor is it easy to pre-specify. Ignorable likelihood-based mixed-effects analyses are consistent with the need for a simple, pre-specified analysis, based on the ITT principle, provided that the treatment compliance is the same for those who drop out and those who remain in the study and data are collected after drop out. It may not be feasible to collect data after drop out and information on treatment compliance may not be available in those who dropped out (Mallinckrodt, Clark, et al., 2003). Any particular MNAR model leads to a set of conclusions if the assumptions about the dropout mechanism were true. Since these assumptions cannot be tested, any particular model could not be the definitive conclusion of a trial (Molenberghs & Kenward, 2007). This has changed over the years and now regulators may want reassurance that inference is robust to departures from the assumptions of the primary analysis. This can be done with sensitivity analysis (Carpenter et al., 2013).

We did not consider incomplete time-to-event data. The standard approach to dealing with missing data in these studies is to censor the participant at the last time where data were observed. The assumption underlying this is that censoring is non-informative, or MAR. This assumption is probably valid for participants who reach the scheduled end of study without experiencing the event under study. This assumption is not appropriate for participants who leave the study early, either due to a competing event or withdrawal from the study. Including a withdrawn participant as censored in a time to event analysis is regarded as following the ITT principle. In reality in order to follow the ITT principle, one might need to assume that participants who withdraw early from the study are not MAR, but have informative censoring. These methods are not as well developed as in the longitudinal case with continuous outcome data (Keene, 2011).

# Chapter 4
# Sensitivity analyses

With incomplete data there is no definitive correct analysis of the data; especially not when missing data are MNAR. In addition, the data cannot be used to verify any of the assumptions made about the missing data mechanism, since this is not contained in the observed data. This point was made throughout the previous chapters and is often highlighted by various authors (Carpenter & Kenward, 2007; Mallinckrodt, Clark, et al., 2003; Molenberghs & Kenward, 2007). Because the data include no information about the missing data process and the non-identified parameters, a sensitivity analysis should be done over a range of different values of the parameters (Daniels & Hogan, 2000). Sensitivity analyses should play a central role in the analysis of incomplete data.

Recommendation 15 of the National Research Council (2010) states: "Sensitivity analyses should be part of the primary reporting of findings from clinical trials. Examining sensitivity to the assumptions about the missing data mechanism should be a mandatory component of reporting."

There are many approaches to sensitivity analyses. These include adding sensitivity parameters to models, investigating local influence, incorporating prior belief, fitting different models, such as pattern-mixture models and selection models.

## 4.1 Definition of sensitivity analysis

The CHMP (2010) defined sensitivity analyses as a set of analyses where the missing data are handled in a different way in each analysis. They argued that it should be presented to support the

main analysis. Beunckens et al. (2009) adapted a pragmatic definition of a sensitivity analysis as an analysis in which several statistical models are considered simultaneously and/or where a statistical model is further scrutinised using specialised tools, such as diagnostic measures. White et al. (2007) described sensitivity analyses as follows: Instead of using the complete case analysis at the final endpoint (assuming MCAR within randomised groups) the intervention effect should be estimated assuming MAR (using an appropriate mixed model or multiple imputation). The sensitivity of the MAR analysis to informative missingness should then be investigated rather than exploring the sensitivity of the MCAR analysis to informative missingness. During this process a range of conclusions, rather than a single conclusion, is obtained. This provides insight into the sensitivity to the assumptions made (Thijs et al., 2002). Daniels and Hogan (2008) defined sensitivity analyses as the assessment of sensitivity of model-based inferences to assumptions that cannot be verified or checked with the data. In the missing data setting, this refers to inferences around the full data distribution.

Janssens et al. (2012) noted that sensitivity could refer to different concepts. It could refer to varying the parameter of interest, ways of data analysis, assumptions about the data, assumptions about the missing data or different analysis populations. In the sensitivity analysis they performed the way in which the data were analysed, the model specification, the class of model (selection, parameter or multiple imputation), assumptions about the data and assumptions about the missing data (MAR versus MNAR) were varied.

Thus although there seems to be consensus about the need for and importance of sensitivity analysis, there is not much consensus about exactly what is defined as a sensitivity analysis.

## 4.2 Guiding principles for sensitivity analysis

Several guidance for sensitivity analyses were suggested by various authors. Carpenter and Kenward (2007) gave the following guiding principles for sensitivity analyses.

1. The sensitivity analyses should be pre-defined, and address the impact of clinically plausible departures from MAR. The specification of the sensitivity analyses should be part of the trial planning process. Opinions should be collected on the differences between responders and non-responders from experts in the field.
2. The sensitivity analyses should be transparent to non-statisticians, most notably the investigators and regulators.
3. The statistical methods employed should be applicable to a wide range of settings.
4. Technically simple methods are often used because it is argued that these methods are less complicated. The assumptions behind these simple methods could sometimes be unrealistic or complicated to understand. It should therefore not be assumed that a

technically simple method is transparent. More complicated methods with simple and transparent assumptions are favoured over technically simple methods with opaque assumptions.

5. It is important to vary the assumptions about the missing data mechanism rather than to simply vary the statistical methods or models used. Different methods that make the same assumption about the underlying and unobserved missing data mechanism could lead to the same conclusion, without providing a true sensitivity analysis. The appropriate techniques should be used to analyse the data under each of the alternative assumptions.

The CHMP (2010) suggested the following:

1. Certain types of missing data should be treated as MNAR.
2. Multiple imputation methods should be used to incorporate a pattern-mixture approach.
3. The impact different model settings have on the results should be compared.
4. Retrieved dropout data could be utilised.
5. A worst case analysis should be done.
6. A responder analysis should be done where all missing data are treated as failures and another analysis should be done where missing data due to certain reasons are treated as failures and missing data due to other reasons are treated as successes.

Some of the suggestions by the CHMP seem to violate the fifth guiding principle of Carpenter and Kenward (2007) as given above, since these methods focus on changing the methodology or model, rather than varying the assumptions made in the analysis.

A distinction can be made in sensitivity analysis between two different analyses; one tests assumption sensitivity and one tests parameter sensitivity. For assumption sensitivity, alternative models are fitted by changing the key assumptions; including assumptions about the model of the missingness and the distributional assumptions of the full data model. Parameter sensitivity involves running the base model with the parameters controlling the extent of the departure from MAR in a plausible range. Within the selection model paradigm, a strategy is to consider various dependencies of the missing data process on the outcomes and/or covariates.

## 4.3 The role of MAR and MNAR models in sensitivity analyses

Various plausible MNAR models could be fitted in a sensitivity analysis (Molenberghs & Kenward, 2007). Or one can assess how a collection of MNAR models differ from the set of models with equal fit to the observed data but that are of a MAR nature.

Molenberghs et al. (2004) suggested that a sensitivity analysis is a good compromise between fitting only a MNAR model and ignoring MNAR models completely. Such an analysis could explore the dependence of the results on departures from MAR assumptions. The aim of sensitivity analysis is to explore plausible departures from MAR and whether this changes the conclusions, rather than to confirm that a specific MNAR model is correct (Carpenter & Kenward, 2007).

Carpenter et al. (2002) suggested that a sensitivity analysis should consist of building an '*envelope of conclusions*' bounded by the results of the MAR model and the worst case MNAR model. The selection of a worst case model should depend on scientific judgement. There is no use in assuming worst case implausible outcomes. The sensitivity of the conclusions to non-random missingness should be assessed by modifying the MAR model to allow for various plausible MNAR scenarios. The conclusions are then examined to see whether they vary.

In a similar vein Beunckens et al. (2009) suggested that a selected number of plausible MAR models be fitted, or that a preferred (primary) analysis is supplemented with a number of variations. If the conclusions are stable across all the models it provides an indication that the results are robust to inherently untestable assumptions about the missingness mechanism.

## 4.4 Pattern-mixture models in sensitivity analysis

Daniels and Hogan (2000) parameterised the pattern-mixture model in a manner that allowed sensitivity analyses to be formulated in terms of between-pattern differences in means and variances. The sensitivity analysis then follows by varying non-identified parameters directly, allowing examination of all specifications along a continuum. They considered the special case where

$$y_i = (y_{i1}, \dots, y_{in})'$$

and only $y_{in}$ is missing (Little & Wang, 1996). An identifying restriction is

$$\mathrm{p}\left(y_{in}^{mis} \left| Ly_i^{\star}, y_{in}\right.\right) = f(Ly_i^{\star} + \lambda y_{in})$$

where $L$ is a $1 \times n\text{-}1$ matrix and $y_i^{\star}=(y_{i1}, \dots, y_{i,n-1})$. Different values of $L$ and $\lambda$ characterise assumptions about the pattern specific regression of $y_{in}$ on $y_{i1}, \dots, y_{i,n-1}$.

The advantage of this method for sensitivity analysis is that it constrains marginal means, variances and covariances. It makes explicit the dependence of the missing data mechanism on $Y_i$. The non-identified parameters in a pattern are estimated using data from within the pattern. Parameters already identified are not dependent on $\lambda$. The data offer no information that can be used to estimate $\lambda$, but sensitivity analyses can be done by varying $\lambda$ through the range of plausible values and checking whether this changes the findings (Daniels & Hogan, 2000).

The pattern-mixture model is then rewritten in terms of between-pattern location and scale changes. This provides a useful framework for imposing model constraints and interpreting them and conducting sensitivity analyses. This is an easier way to communicate the sensitivity analyses to non-statisticians. It may be easier to formulate the constraints based on, for example, the differences in outcome between those who dropped out and those who completed the study than in terms of missing data mechanisms. Another advantage is that this method makes explicit all non-identified parameters. The non-identified parameters for patterns with incomplete data are functions of the fully identified parameters in the complete pattern through additive terms. Using these location-scale parameterisations reduces the sensitivity analyses to a series of complete data problems. The model is fully identified by fixing these components characterising the differences between completers and dropouts (Daniels & Hogan, 2000).

In order to draw inferences about the marginal mean over all patterns of missing data information about the marginal variance is needed. Sensitivity analyses based on the unidentified components of the variance and location parameters can be done based on the assumed differences between unobserved data within a pattern relative to patterns with more complete data. The non-identified components of these parameters can be varied without affecting the already identified components. The marginal distribution of the observed data can be held fixed while examining different non-ignorable missing data mechanisms. Sensitivity analyses are then done by comparing inference about the difference in treatment arms at different combinations of the unknown components in the parameters (Daniels & Hogan, 2000).

Pattern-mixture models offer an advantage over selection models when sensitivity analyses are done, since the assumptions are explicit (Verbeke et al., 2001). Special attention should be given to the ACMV restriction in pattern-mixture models since this model represents the MAR counterpart in other types of models. Complete case (CCMV) and neighbouring value missing (NCMV) value restrictions represent the extremes for the $\omega_s$ vector. It thus makes sense to regard the results found with these two methods as the ranges between which the results are likely to fall (Thijs et al., 2002).

Pattern-mixture models can be factorised as

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i|\boldsymbol{x}_i, \boldsymbol{\theta}, \psi) = f\left(\boldsymbol{y}_i^{obs}|\boldsymbol{r}_i, \boldsymbol{x}_i, \boldsymbol{\theta}\right) f\left(\boldsymbol{y}_i^{mis}|\boldsymbol{y}_i^{obs}, \boldsymbol{r}_i, \boldsymbol{x}_i, \boldsymbol{\theta}\right) f(\boldsymbol{r}_i|, \boldsymbol{x}_i, \boldsymbol{\psi})$$

This parameterisation splits the full data model into a factor with identified and factors with non-identified components. It is easy to identify the sensitivity parameter(s) from here. If a pattern-mixture model is used, the general model allows the missing data to be MNAR, with MAR as a special case. The sensitivity parameters are then defined such that the degree of departure from

MAR is quantified by the sensitivity parameters. These parameters are non-identified and the fit of the model to the observed data does not differ when different values for the sensitivity parameter are assumed. In most situations the models have a large number of sensitivity parameters. It is often necessary to reduce these parameters by making some assumptions. Care should be taken when choosing the prior distributions for the missing value mechanism and the parameter space over which the sensitivity parameters are varied (Daniels & Hogan, 2008).

Thijs et al (2002) also considered sensitivity analysis within the context of pattern-mixture models by using three distinct strategies to fit the models. The first strategy was to use identifying restrictions, the second was to fit a model for each pattern of missing data and the third was to fit a single model with pattern as covariate. The results obtained with these strategies were then contrasted to get a range of conclusions. The sensitivity analysis was therefore conducted by determining whether results change under the different assumptions made by these models.

If a pattern-mixture model is fitted, the choice of the different patterns could be subjective and it might be important to check the robustness of the findings if different patterns are chosen (Molenberghs & Kenward, 2007).

## 4.5 Bayesian methods in sensitivity analysis

In the context of Bayesian analysis, a sensitivity parameter is defined as a parameter created through a reparameterisation of the full data parameters such that the sensitivity parameter is a non-constant function of the parameters of the extrapolation distribution and the fit of the model to the observed data is not affected by the sensitivity parameter. When the sensitivity parameters are fixed, the full data model is identified. The sensitivity parameters are used to quantify beliefs about the missing data mechanism. This is done by fixing their values at constants, examining inference across a range of constants, or by assigning a prior distribution (Daniels & Hogan, 2008).

A challenge when using prior distributions or even parameters during sensitivity analyses is to make the priors understandable and transparent to subject matter experts interpreting the model-based inferences or to make them amenable to incorporation of external information (Daniels & Hogan, 2008; Scharfstein et al., 1999). Sensitivity analysis is done by looking at a specific model for the full-data distribution where some of the parameter(s) are sensitivity parameters as defined above. The sensitivity parameters are informed by prior distributions that characterise assumptions about the missing data mechanism or the missing data itself (Daniels & Hogan, 2008).

Some principles for the reparameterisation of the models for sensitivity analyses are:

1. Parameterise the model in terms of a sensitivity parameter that satisfies the definition of a sensitivity parameter given two paragraphs above. Thus assumptions about the missing data mechanism are fully encoded by prior distributions. One can move through all full-data models with MNAR missingness by varying the value of the sensitivity parameters. Changing the value of the sensitivity parameter does not affect the observed data likelihood and the fit to the observed data.
2. The prior distribution of the sensitivity parameter should reflect the uncertainty about the missing data assumptions and other non-identifiable aspects of the model. The prior distribution for the sensitivity parameter can be informed by external information, such as expert opinion. Uncertainty about these assumptions can be incorporated through Bayesian priors.
3. The full-data models should be centered at MAR, such that the MAR assumption coincides with a specific point of the sensitivity parameters. Then the effect of the missing data assumptions can be viewed in terms of departures from MAR (Daniels & Hogan, 2008).

## 4.6 Global and local sensitivity

Sensitivity analysis can also be performed on the level of individual observations. One tries to find observations that make the conclusions more in line with a MNAR model. Two techniques that allows for this are global and local influence. The method of local influence detects observations with a large influence on the model of interest or the missingness model parameters. Local influence changes, for example, the missingness process for one observation from MAR to MNAR. Global influence is based on case deletion, or occasionally a measurement deletion approach. It is based on the difference in log-likelihood between the model that is fitted to the entire data set and the data set where one participant is removed (Beunckens et al., 2009; Molenberghs & Kenward, 2007). Local and global influence are related to, but different from, each other.

## 4.7 Uncertainty region

Most sensitivity analyses are ad hoc. One attempt to create a systematic tool for sensitivity analysis is the interval of uncertainty, a region of possible values of a parameter that is consistent with the data (Beunckens et al., 2009; Molenberghs, Kenward, & Goetghebeur, 2001; Vansteelandt et al., 2006). This region depends on the data, both observed and missing, and on the model fitted. The region of ignorance, which is created due to the incompleteness of the data, is added to the measure of imprecision, created by the sampling uncertainty, to create the region of uncertainty. This region can be constructed in various ways, defined as strong coverage and weak coverage.

Strong 95% uncertainty intervals are designed to cover all values in the ignorance region itself simultaneously with 95% probability. This strong uncertainty region is calculated by adding the standard 100(1-α) % confidence limits to the estimated ignorance limits. That is

$$[C_L, C_U] = [\hat{\beta}_l - z_{\frac{\alpha}{2}} \text{se}(\hat{\beta}_l), \hat{\beta}_u - z_{\frac{\alpha}{2}} \text{se}(\hat{\beta}_u)]$$

where $C_L$ and $C_U$ are the lower and upper limits of the strong uncertainty region, $\hat{\beta}_l$ and $\hat{\beta}_u$ are the lower and upper limits of the estimated ignorance limits and $z_{\alpha/2}$ is the 100(1-α/2)th percentile of the standard normal distribution. Weak 95% uncertainty intervals are designed to have expected 95% overlap with the ignorance region and can be constructed as

$$[C_L, C_U] = [\hat{\beta}_l - z_{\frac{\alpha*}{2}} \text{se}(\hat{\beta}_l), \hat{\beta}_u - z_{\frac{\alpha*}{2}} \text{se}(\hat{\beta}_u)]$$

where $z_{\frac{\alpha*}{2}}$ solves the equation

$$\alpha = \frac{\text{se}(\hat{\beta}_l) + \text{se}(\hat{\beta}_u)}{\beta_u - \beta_l} \int_0^{+\infty} z\varphi(z + z_{\frac{\alpha*}{2}})dz + \epsilon$$

where $\varphi(.)$ is the standard normal density function. The correction term $\epsilon$ is so small that it can be set as 0 without affecting the accuracy. A solution for $z_{\frac{\alpha*}{2}}$ can only be found by substituting $\hat{\beta}_l$, $\hat{\beta}_u$, $\text{se}(\hat{\beta}_l)$ and $\text{se}(\hat{\beta}_u)$ with consistent estimators (Vansteelandt et al., 2006).

The largest possible set of identifiable parameters is selected. All remaining parameters are then regarded as sensitivity parameters. Values are chosen for the sensitivity parameters and the identifiable parameters are estimated using the most appropriate model. If this is done for all possible values of the sensitivity parameter, a region of estimates is obtained. Thus the combined effects of imprecision and ignorance are captured in the region of uncertainty. Details regarding construction and asymptotic properties can be found in Vansteelandt et al. (2006).

# Chapter 5

# SAPiT study data analysis

The study of missing data was motivated by the analysis of the secondary objectives of the SAPiT (**S**tarting **A**ntiretroviral therapy at three **P**oints **i**n **T**uberculosis) clinical trial (Abdool Karim et al., 2010). In Section 5.1 we describe the design and published results of the SAPiT trial. Section 5.2 continues with a description of the missing data in the trial. We then analyse the data using MCAR methods in Section 5.3, MAR methods in Section 5.4 and MNAR methods in Section 5.5. We conclude the chapter by contrasting these analyses and drawing a conclusion regarding the findings.

## 5.1 The SAPiT study

### 5.1.1 Background to the study

It is estimated that in 2010 there were about 34 million human immunodeficiency virus (HIV) infected people worldwide (Joint United Nations Programme on HIV/AIDS (UNAIDS), 2010) and 8.7 million new tuberculosis cases in 2011, 13% of these co-infected with HIV (World Health Organization, 2012). In Africa, 46% of tuberculosis patients tested for HIV, were found to be HIV-positive (World Health Organization, 2012). Globally, an estimated 380 000 HIV infected persons died due to tuberculosis in 2009 (Joint United Nations Programme on HIV/AIDS (UNAIDS), 2010). Tuberculosis is the most common presenting opportunistic infection in HIV infected individuals (Churchyard et al., 2000), and is the most common cause of mortality in acquired immune deficiency syndrome (AIDS) patients in developing countries (Mukadi, Maher, & Harries,

2001). In the presence of HIV, tuberculosis is associated with substantially higher case fatality rates, even with effective tuberculosis chemotherapy (Schluger, 1999).

Tuberculosis is the most common notified cause of death in South Africa (Health Systems Trust, 2008). In 2011, South Africa had an estimated 5.6 million people infected with HIV (Joint United Nations Programme on HIV/AIDS UNAIDS, 2012) and 401 048 reported tuberculosis cases in 2010 (Day, Gray, & Budgell, 2012), of whom approximately 53% were dually infected with tuberculosis and HIV (South African Department of Health, 2007). The World Health Organisation (WHO) estimates that 60% of tested tuberculosis patients in South Africa were HIV positive in 2010 (World Health Organisation, 2011). In this setting, routine HIV testing in tuberculosis patients is an important path for entry into HIV treatment programs.

Prior to the SAPiT study there was no clinical trial data to guide the timing of antiretroviral therapy initiation in tuberculosis patients. The timing of antiretroviral therapy initiation in tuberculosis patients was variable, depending on clinician judgement. Available guidelines were based on observational data and expert opinion (National Department of Health, 2004). The WHO guidelines (World Health Organisation, 2003) urged integrated treatment of HIV and tuberculosis, but clinicians often deferred HIV treatment in tuberculosis-HIV co-infected patients because of numerous concerns. These included drug interactions between rifampicin and some classes of antiretroviral therapy (Piscitelli & Gallicano, 2001), additive side effects and toxicities (Girardi et al., 2001) and a high pill burden.

The aim of the SAPiT trial was to determine the optimal time to initiate antiretroviral therapy in patients on tuberculosis treatment and the primary outcome was mortality (Abdool Karim et al., 2010).

### 5.1.2 Methods

The SAPiT trial (protocol number: CAPRISA 003) was an open label, randomised controlled trial comparing three treatment strategies of antiretroviral therapy initiation in HIV-tuberculosis co-infected patients. Participants were randomised to one of three arms:

1. Early integrated treatment arm: antiretroviral therapy initiated within 4 weeks of starting tuberculosis treatment
2. Late integrated treatment arm: antiretroviral therapy initiated within 4 weeks of completing the intensive phase of tuberculosis treatment, which happens approximately 2 months after the initiation of tuberculosis therapy
3. Sequential treatment arm: antiretroviral therapy initiated within 4 weeks of completing tuberculosis treatment, which happens from 6 months after initiation of tuberculosis therapy

The trial was conducted at the CAPRISA eThekwini tuberculosis-HIV clinic, which adjoins the Prince Cyril Zulu Communicable Disease Centre, one of the largest out-patient tuberculosis facilities in South Africa. From 28 June 2005 to 11 July 2008, HIV positive tuberculosis infected patients, who were 18 years or older were recruited from the Prince Cyril Zulu Communicable Disease Clinic. To qualify for inclusion in the trial, participants had to be initiated on the standard tuberculosis treatment regimens, as stipulated in the South African National Tuberculosis Control Programme guidelines (Department of Health, 2004), had to have a CD4+ count below 500 cells/mm$^3$ at screening and had to have no clinical contra-indications to initiation of antiretroviral therapy. Female participants were required to agree to use contraception while on efavirenz. After providing informed consent, qualifying tuberculosis-HIV co-infected patients were enrolled and randomised to one of the three treatment arms in a 1:1:1 ratio in permuted blocks of six or nine with no stratification, using sealed envelopes. Trial arm assignment was concealed until after randomisation, but thereafter both the participants and the clinic staff were aware of participant arm assignment.

According to the South African National Tuberculosis Control Programme Guidelines (Department of Health, 2004) the first episode of tuberculosis is treated with a fixed drug combination of rifampicin, isoniazid, ethambutol and pyrazinamide for a 2-month intensive phase. Thereafter, the 4-month continuation phase comprises a fixed-drug combination of isoniazid and rifampicin. Patients with a history of tuberculosis receive a 60-day intensive phase which includes streptomycin, followed by a 100-day continuation phase. The Prince Cyril Zulu Communicable Disease Centre offers clinic-based directly observed therapy. Some patients did not attend the clinic daily for directly observed therapy.

Participants were scheduled to visit the clinic monthly for 24 months. At these visits antiretroviral therapy was dispensed and clinical status was monitored. CD4+ count was measured using the FACS Calibur flow cytometer (Becton Dickinson, Franklin Lakes New Jersey, United States of America). HIV RNA was measured using the Roche Cobas Amplicor HIV-1 Monitor v1.5. CD4+ count and HIV RNA viral load were measured at screening, randomisation and every 6 months thereafter.

### 5.1.3 Primary and secondary endpoints of the SAPiT trial

The primary endpoint of the trial was all cause mortality. During an interim analysis of the data it was found that initiation of antiretroviral therapy during tuberculosis treatment (arms 1 and 2 combined) in patients with sputum tuberculosis and HIV co-infection with CD4+ counts <500 cells/mm$^3$ reduced mortality by 56% (95% confidence interval [CI]: 21% to 75%, p = 0.003).

Mortality in patients with integrated HIV and tuberculosis treatment was 5.4 per 100 person-years, while it was 12.1 per 100 person-years in patients when initiation of antiretroviral therapy was delayed until completion of tuberculosis treatment (Abdool Karim et al., 2010).

The population, demographic data and primary endpoint of mortality have been analysed and reported previously for interim data (Abdool Karim et al., 2010) and for complete data in the first two arms (Abdool Karim et al., 2011). It was shown that there was a significant difference in mortality between the treatment arms. At the time of the interim analysis, overall 8.1% of the participants in the study had died. In addition to the high mortality, loss to follow-up was also high with 18.4% of participants being terminated or lost to follow-up at the time of the analysis. This was fairly evenly spread between the study arms.

Secondary endpoints included tolerability, toxicity, CD4+ counts, HIV RNA viral load, tuberculosis outcomes and immune reconstitution inflammatory syndrome (IRIS). Tolerability was defined as study initiated treatment interruptions in the pharmacy records. Toxicity was assessed by a clinical checklist and standard laboratory assessments for haematological, hepatic and renal abnormalities. IRIS was defined as a paradoxical deterioration in clinical status after antiretroviral therapy initiation without another attributable cause. One of the secondary endpoints was to compare HIV-outcomes in terms of CD4+ count and HIV RNA viral load at 6, 12, 18 and 24 months post randomisation and to investigate the difference between the treatment arms in variations in longitudinal CD4+ count.

CD4+ cells are a type of white blood cells (lymphocyte) that fight infection. The CD4+ count measures the number of CD4+ cells in a certain volume of blood. HIV infects CD4+ T-cells and causes immunosuppression in the human host; this is measured as a low CD4+ count. The CD4+ count is an indication of the level of immunopathology caused by the HIV infection and is a crude representation of the damage done to the immune system by HIV infection. Thus CD4+ counts are used clinically as a measure of disease progression and effectiveness of antiretroviral therapy (Gray et al., 2010). CD4+ cell count is a useful surrogate marker of treatment effects and effectiveness (Daniels & Hughes, 1997). Normal CD4+ counts range between 600 and 1500 cells/$mm^3$ in adults (BARC laboratory reference ranges). This is much reduced during HIV infection. In 2010 the South African treatment guidelines required that antiretroviral treatment be started when a patient's CD4+ count was below 200 cells/$mm^3$

The objective of this chapter is to analyse the secondary outcome of CD4+ count profiles over time in the three arms. The goal is to characterise the changes in CD4+ count over time as a function of

treatment arm, in order to determine whether integrated HIV and tuberculosis treatment improved or worsened HIV outcomes.

The high mortality, combined with the even higher loss to follow-up rate suggested that the secondary objective of CD4+ count could not be analysed without taking missing data into account, since about a quarter of participants did not complete the study and did not have CD4+ counts past 12 months post randomisation. It is possible that dropout is related to CD4+ count. It is plausible that the unobserved (because they are missing) CD4+ counts among those who dropped out are lower than those who continue follow-up, even after adjusting for observed CD4+ counts and other covariates. The goal of this analysis was to analyse the secondary objective of CD4+ count in a valid way, while taking missing data into account. This analysis included the entire 24 months of follow-up per participant in the SAPiT trial. The primary endpoint was CD4+ count at 6, 12, 18 and 24 months.

### 5.1.4 Statistical notation

The statistical notation was introduced in Section 2.1. It was adapted for this data set as follows: We assume $N$ independent participants, indicated by $i = 1, \ldots, N$. The CD4+ count outcome for the $i^{th}$ participant at $j^{th}$ occasion is given by $Y_{ij}$. Treatment is indicated by the two variables $E_i$ and $L_i$. If the $i^{th}$ participant is in the early integrated treatment arm, then $E_i = 1$, else $E_i = 0$. If the $i^{th}$ participant is in the late integrated treatment arm, then $L_i = 1$, else $L_i = 0$. A participant $i$ belonging to the sequential treatment arm will thus have $E_i = L_i = 0$. The sequential arm is therefore the reference group. $t_{ij}$ is the time point for the $i^{th}$ participant at the $j^{th}$ occasion, and can take the values 0, 6, 12, 18 and 24 months. When dropout is defined as a binary variable, we indicate dropout by $D_i$. If participant $i$ dropped out, then $D_i = 1$, if participant $i$ did not drop out then $D_i = 0$. In some models dropout was not defined as a binary variable. In those instances it is described where the model is given.

## 5.2 Patterns of missing data

It is important to understand and try to explain why data are missing using the data collected. The number of missing data points and the reasons for having missing data are provided in Table 5.1. Most of the participants lost to follow-up were lost within the first year, and after 12 months only a few additional participants had missing values. Treatment arm seemed to be related to whether data were missing, with missing data being more prevalent as antiretroviral initiation was delayed (exact Cochran-Armitage trend test , p = 0.01). This test compares the number of participants in pattern 1 in Table 5.2 in the three treatment arms. The trend investigated went from early

integrated treatment arm (immediate antiretroviral therapy), to late integrated treatment arm (antiretroviral therapy at the end of the intensive phase) to sequential treatment arm (antiretroviral therapy after tuberculosis treatment). The early integrated treatment arm had fewer missing values than the other two treatment arms. Participants were also lost to follow-up earlier in the later treatment arms. More than one third of participants, 37.9%, were lost to follow-up by 24 months. Overall, 28% (69/243) of the missing data at 24 months were caused by deaths. Mortality had been shown to be related to treatment arm (Abdool Karim et al., 2010), thus any analysis of CD4+ count at 24 months could be biased if the issue of missing data were not properly taken into account.

**Table 5.1: Number of participants attending each 6-monthly evaluation**

| | Early integrated treatment arm N = 214 | | Late integrated treatment arm N = 215 | | Sequential treatment arm N = 213 | |
|---|---|---|---|---|---|---|
| | **N** | **Number missing (%)** | **N** | **Number missing (%)** | **N** | **Number missing (%)** |
| **Number with CD4+ count at** | | | | | | |
| 6 months | 173 | 41 (19.2%) | 159 | 56 (26.0%) | 154 | 59 (27.7%) |
| 12 months | 160 | 54 (25.2%) | 150 | 65 (30.2%) | 128 | 85 (39.9%) |
| 18 months | 152 | 62 (29.0%) | 143 | 72 (33.5%) | 123 | 90 (42.3%) |
| 24 months | 145 | 69 (32.2%) | 131 | 84 (39.1%) | 123 | 90 (42.3%) |
| **Number of participants dead at** | | | | | | |
| 6 months | 10 | | 5 | | 12 | |
| 12 months | 13 | | 12 | | 32 | |
| 18 months | 14 | | 15 | | 35 | |
| 24 months | 17 | | 17 | | 35 | |
| **Reasons not attending 6 month visit** | | | | | | |
| Lost to follow-up | 21 | | 38 | | 31 | |
| Dead | 10 | | 5 | | 12 | |
| Terminated | 6 | | 5 | | 5 | |
| Attended later visits, missed the 6 month visit | 4 | | 8 | | 11 | |
| **Reasons not attending 12 month visit** | | | | | | |
| Lost to follow-up | 25 | | 37 | | 36 | |
| Dead | 13 | | 12 | | 32 | |
| Terminated | 13 | | 15 | | 10 | |
| Attended later visits, missed the 12 month visit | 3 | | 1 | | 7 | |
| **Reasons not attending 18 month visit** | | | | | | |
| Lost to follow-up | 27 | | 35 | | 35 | |
| Dead | 13 | | 15 | | 35 | |
| Terminated | 17 | | 19 | | 14 | |
| Attended later visits, missed the 18 month visit | 5 | | 3 | | 6 | |
| **Reasons not attending 24 month visit** | | | | | | |
| Lost to follow-up | 0 | | 2 | | 0 | |
| Dead | 17 | | 17 | | 35 | |
| Terminated | 52 | | 65 | | 55 | |

Participants are regarded as lost to follow-up when they have not attended clinic visits for three months. Attempts are made to locate participants who do not attend visits. These attempts are repeated at intervals until study end. Sometimes a participant is regarded as lost to follow-up at a visit and is later found to have died subsequent to being lost to follow-up. These participants are classified as lost to follow-up at the earlier visit and dead at the subsequent visit.

The different patterns of missing data and the number of participants in each of these patterns are given in Table 5.2. Most participants completed the study. The missing data pattern with the most participants was pattern 5, where participants had baseline data only.

**Table 5.2: Patterns of missing data in each of the treatment arms**

| Pattern | Baseline | 6 months | 12 months | 18 months | 24 months | Early integrated treatment arm N = 214 | Late integrated treatment arm N = 215 | Sequential treatment arm N = 213 |
|---|---|---|---|---|---|---|---|---|
| 1 | X | X | X | X | X | 137 (64.0%) | 122 (56.7%) | 108 (50.7%) |
| 2 | X | X | X | X | | 8 (3.7%) | 12 (5.6%) | 5 (2.4%) |
| 3 | X | X | X | | | 8 (3.7%) | 6 (2.8%) | 5 (2.4%) |
| 4 | X | X | | | | 14 (6.5%) | 16 (7.4%) | 30 (14.1%) |
| 5 | X | | | | | 37 (17.3) | 48 (22.3%) | 48 (22.5%) |
| 6 | X | Completers, with some interim missing data | | | X | 8 (3.7%) | 9 (4.2%) | 15 (7.0%) |
| 7 | X | Did not complete, with interim missing data | | | | 2 (0.9%) | 2 (0.9%) | 2 (0.9%) |

X: Indicates data present

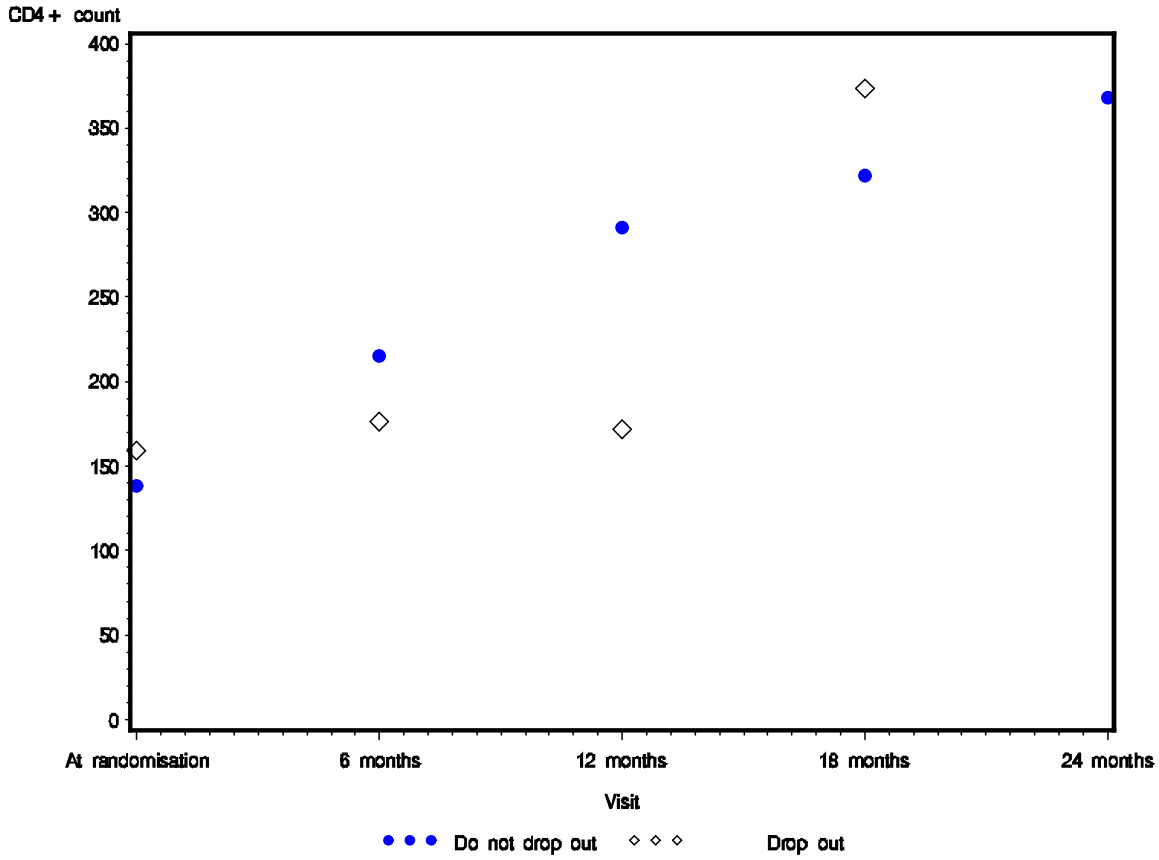



**Figure 5.1: Sample medians of CD4+ counts at each 6 monthly visit.**
The solid dots represent the medians for participants who did not drop out before the subsequent CD4+ measurement. The diamonds represent the medians of participants who dropped out before the subsequent CD4+ measurement

Although the observed data cannot give information about the unobserved missing data, there is an indication that dropout in this study is not MCAR. Figure 5.1 gives the observed median CD4+ count at each of the time points for the participants who dropped out at the subsequent visit and for those who did not drop out at the subsequent visit. At baseline median CD4+ count was similar for those who dropped out and those who did not. However, at 6 and 12 months the median CD4+ count of those who dropped out at the subsequent visit was lower than for those who did not drop out. At 18 months the median CD4+ counts of those who dropped out was higher than for those who did not drop out. This is an interesting phenomenon. A possible reason is that by 18 months participants have improved sufficiently to be less likely to drop out due to ill health. It is also possible that participants who are likely to drop out due to lower CD4+ counts have already done so by 18 months. These theories cannot be validated using the available data. It should also be borne in mind that there were only 29 participants who dropped out after 18 months. The median for those who dropped out was therefore less stable and more variable. The fact that participants with lower median CD4+ count are more likely to drop out at the earlier visits suggests that dropout depends at least on observed CD4+ count, and implies that model-based means could be different from observed means.

At 6 months there was a significant difference between the CD4+ counts of participants who missed the 12 month visit and those who did not miss this visit. The median CD4+ count at 6 months was 198 (interquartile range [IQR]: 65 to 301) cells/$mm^3$ in participants who missed their 12 month visit and 212 (IQR: 117 to 350) cells/$mm^3$ in participants who attended the 12 month visit (median difference = -14, Wilcoxon p = 0.049).

At 12 months there was not a significant difference between the CD4+ counts of participants who missed the 18 month visit and those who did not miss this visit. The median CD4+ count at 12 months was 239 (IQR: 163 to 381) cells/$mm^3$ in participants who missed their 18 month visit and 289 (IQR: 182 to 431) cells/$mm^3$ in participants who attended the 18 month visit (median difference = -50, Wilcoxon p = 0.15).

Carpenter and Kenward (2007) suggested that logistic regression and/or survival analysis be used to determine key independent variables associated with withdrawal. We used logistic regression to identify variables associated with missing CD4+ count at each of the 6-monthly visits. Baseline variables believed by the clinical staff to be related to tuberculosis and HIV outcomes were used in this analysis.

**Table 5.3: Variables associated with missing a visit, results of logistic regression, modelling the probability of withdrawing**

| | Missed visit N (%) | Univariate Odds Ratio (95% CI) | p-value | Multivariate Odds Ratio (95% CI) | p-value |
|---|---|---|---|---|---|
| **Variables associated with missing the 6 month visit** | | | | | |
| Arm (ref sequential) | 59/213 (27.7) | 1.00 | | 1.00 | |
| - Early treatment arm | 41/214 (19.2) | 0.62 (0.39 to 0.97) | 0.04 | 0.61 (0.38 to 0.97) | 0.04 |
| - Late treatment arm | 56/215 (26.1) | 0.92 (0.60 to 1.41) | 0.70 | 0.93 (0.60 to 1.43) | 0.72 |
| Gender (ref male) | 86/319 (27.0) | 1.00 | | 1.00 | |
| - Female | 70/323 (21.7) | 0.75 (0.52 to 1.08) | 0.12 | 0.64 (0.44 to 0.94) | 0.02 |
| Age | - | 0.97 (0.95 to 0.99) | 0.01 | 0.96 (0.94 to 0.98) | 0.001 |
| WHO status (ref = Stage 4) | 150/604 (24.8) | 1.00 | | 1.00 | |
| - Stage 3 | 6/38 (15.8) | 1.76 (0.72 to 4.30) | 0.21 | 1.69 (0.54 to 5.28) | 0.37 |
| CD4+ count at baseline (50 cell/$mm^3$ increments) | - | 1.06 (0.99 to 1.14) | 0.12 | 1.05 (0.98 to 1.14) | 0.16 |
| Extra pulmonary tuberculosis | 148/599 (24.7) | 1.00 | | 1.00 | |
| - Yes | 8/43 (18.6) | 0.70 (0.32 to 1.54) | 0.37 | 1.11 (0.39 to 3.09) | 0.85 |
| Log viral load at baseline | - | 0.83 (0.67 to 1.02) | 0.08 | 0.83 (0.66 to 1.05) | 0.12 |
| History of tuberculosis | 96/428 (22.4) | 1.00 | | 1.00 | |
| - Yes | 60/214 (28.0) | 1.35 (0.93 to 1.96) | 0.12 | 1.47 (1.00 to 2.17) | 0.05 |
| Multidrug resistant tuberculosis | 151/624 (24.2) | 1.00 | | 1.00 | |
| - Yes | 5/18 (27.8) | 1.21 (0.42 to 3.43) | 0.73 | 1.35 (0.44 to 4.09) | 0.60 |
| **Variables associated with missing the 12 month visit for those not lost to follow-up at the 6 month visit** | | | | | |
| Arm (ref sequential) | 37/165 (22.4) | 1.00 | | 1.00 | |
| - Early treatment arm | 17/177 (9.6) | 0.37 (0.20 to 0.68) | 0.002 | 0.47 (0.24 to 0.93) | 0.03 |
| - Late treatment arm | 17/167 (10.2) | 0.39 (0.21 to 0.73) | 0.003 | 0.47 (0.24 to 0.92) | 0.03 |
| Gender (ref male) | 34/241 (14.1) | 1.00 | | 1.00 | |
| - Female | 37/268 (13.8) | 0.98 (0.59 to 1.61) | 0.92 | 0.90 (0.52 to 1.58) | 0.72 |
| Age | - | 0.97 (0.93 to 1.00) | 0.03 | 0.96 (0.92 to 1.00) | 0.03 |
| WHO status (ref = Stage 4) | 8/33 (24.2) | 1.00 | | 1.00 | |
| - Stage 3 | 63/476 (13.2) | 0.48 (0.21 to 1.10) | 0.08 | 0.81 (0.18 to 3.55) | 0.78 |
| CD4+ count at baseline (50 cell/$mm^3$ increments) | - | 1.00 (0.90 to 1.11) | 0.98 | 1.06 (0.92 to 1.23) | 0.42 |
| Extra pulmonary tuberculosis | 62/473 (13.1) | 1.00 | | 1.00 | |
| - Yes | 9/36 (25.0) | 2.21 (0.99 to 4.92) | 0.05 | 0.65 (0.16 to 2.72) | 0.56 |
| Log viral load at baseline | - | 1.22 (0.89 to 1.67) | 0.21 | 1.22 (0.85 to 1.75) | 0.28 |
| History of tuberculosis | 51/294 (14.8) | 1.00 | | 1.00 | |
| - Yes | 20/164 (12.2) | 1.80 (0.46 to 1.39) | 0.43 | 0.86 (0.48 to 1.57) | 0.63 |
| Multidrug resistant tuberculosis | 66/495 (97.3) | 1.00 | | 1.00 | |
| - Yes | 5/14 (35.7) | 3.61 (1.17 to 11.1) | 0.03 | 2.87 (0.78 to 10.56) | 0.11 |
| CD4+ count at 6 months (ref > 200 cell/$mm^3$) | 33/254 (13.0) | 1.00 | | 1.00 | |
| - < 50 cell/$mm^3$ | 13/41 (31.7) | 3.11 (1.47 to 6.60) | 0.006 | 2.28 (0.78 to 6.65) | 0.13 |
| - - 50 – 200 cell/$mm^3$ | 21/191 (11.0) | 0.83 (0.46 to 1.48) | 0.03 | 0.86 (0.41 to 1.79) | 0.68 |
| - -missing | 4/23 (17.4) | 1.41 (0.45 to 4.40) | 0.96 | 1.25 (0.37 to 4.26) | 0.72 |
| **Variables associated with missing the 18 month visit for those not loss to follow-up at 12 months** | | | | | |
| Arm (ref sequential) | 12/135 (8.9) | 1.00 | | 1.00 | |
| - Early treatment arm | 12/163 (7.4) | 0.82 (0.35 to 1.88) | 0.63 | 1.17 (0.47 to 2.90) | 0.74 |
| - Late treatment arm | 8/151 (5.3) | 0.57 (0.23 to 1.45) | 0.24 | 0.73 (0.28 to 1.96) | 0.54 |
| Gender (ref male) | 15/211 (7.1) | 1.00 | | 1.00 | |
| - Female | 17/238 (7.1) | 1.01 (0.49 to 2.07) | 0.99 | 0.86 (0.39 to 1.89) | 0.70 |
| Age | - | 0.95 (0.91 to 1.00) | 0.047 | 0.95 (0.90 to 1.00) | 0.05 |
| WHO status (ref = Stage 4) | 1/25 (4.0) | 1.00 | | 1.00 | |
| - Stage 3 | 31/424 (7.3) | 1.89 (0.25 to 14.46) | 0.54 | 0.83 (0.07 to 10.58) | 0.89 |
| CD4+ count at baseline (50 cell/$mm^3$ increments) | - | 0.97 (0.82 to 1.13) | 0.67 | 1.03 (0.86 to 1.25) | 0.73 |
| Extra pulmonary tuberculosis | 31/422 (7.4) | 1.00 | | 1.00 | |
| - Yes | 1/27 (3.7) | 0.49 (0.06 to 3.70) | 0.49 | 2.42 (0.19 to 30.04) | 0.49 |
| Log viral load at baseline | - | 0.72 (0.48 to 1.08) | 0.11 | 0.59 (0.36 to 0.95) | 0.03 |

**Table 5.3: Variables associated with missing a visit, results of logistic regression, modelling the probability of withdrawing**

| | Missed visit N (%) | Univariate Odds Ratio (95% CI) | p-value | Multivariate Odds Ratio (95% CI) | p-value |
|---|---|---|---|---|---|
| History of tuberculosis | 19/299 (6.4) | 1.00 | | 1.00 | |
| - Yes | 13/150 (8.7) | 1.40 (0.67 to 2.92) | 0.37 | 1.29 (0.60 to 2.81) | 0.51 |
| Multidrug resistant tuberculosis | 32/440 (7.3) | | | | |
| - Yes | 0/9 (0) | Not calculated | | Not calculated | |
| CD4+ count at 12 months (ref > 200 cell/mm$^3$) | 17/307 (5.5) | 1.00 | | 1.00 | |
| - < 50 cell/mm$^3$ | 4/13 (30.8) | 7.58 (2.11 to 27.1) | 0.002 | 9.68 (2.07 to 45.27) | 0.004 |
| - 50 – 200 cell/mm$^3$ | 8/118 (6.8) | 1.24 (0.52 to 2.96) | 0.63 | 1.57 (0.57 to 4.37) | 0.38 |
| - Missing | 3/11 (27.3) | 6.40 (1.56 to 26.3) | 0.01 | 4.66 (1.04 to 20.99) | 0.04 |
| **Variables associated with missing the 24 month visit for those not loss to follow-up at 18 months** | | | | | |
| Arm (ref sequential) | 6/129 (4.7) | 1.00 | | 1.00 | |
| - Early treatment arm | 12/157 (7.6) | 1.70 (0.62 to 4.65) | 0.30 | 2.14 (0.74 to 6.18) | 0.16 |
| - Late treatment arm | 15/146 (10.3) | 2.35 (0.88 to 6.24) | 0.09 | 2.69 (0.96 to 7.53) | 0.06 |
| Gender (ref male) | 21/201 (10.5) | 1.00 | | 1.00 | |
| - Female | 12/231 (5.2) | 0.47 (0.23 to 0.98) | 0.04 | 0.45 (0.21 to 1.00) | 0.05 |
| Age | - | 1.01 (0.97 to 1.05) | 0.81 | 1.00 (0.96 to 1.05) | 0.90 |
| WHO status (ref = Stage 4) | 0/24 (0) | | | | |
| - Stage 3 | 33/408 (8.1) | Not calculated | | Not calculated | |
| CD4+ count at baseline (50 cell/mm$^3$ increments) | - | 1.21 (1.05 to 1.39) | 0.008 | 1.26 (1.08 to 1.47) | 0.004 |
| Extra pulmonary tuberculosis | 30/405 (7.4) | 1.00 | | 1.00 | |
| - Yes | 3/27 (11.1) | 1.56 (0.45 to 5.49) | 0.49 | 0.59 (0.15 to 2.31) | 0.45 |
| Log viral load at baseline | - | 0.81 (0.53 to 1.22) | 0.30 | 0.94 (0.60 to 1.49) | 0.81 |
| History of tuberculosis | 25/288 (8.7) | 1.00 | | 1.00 | |
| - Yes | 8/144 (5.6) | 0.62 (0.27 to 1.41) | 0.25 | 0.57 (0.24 to 1.36) | 0.20 |
| Multidrug resistant tuberculosis | 33/423 (7.8) | | | | |
| - Yes | 0/9 (0) | Not calculated | | Not calculated | |
| CD4+ count <50 cell/mm$^3$ at any follow-up visit | 30/397 (7.6) | 1.00 | | 1.00 | |
| - Yes | 3/35 (8.6) | 1.15 (0.33 to 3.97) | 0.78 | 2.79 (0.67 to 11.67) | 0.16 |
| Missed any visit prior to 24 months | 25/392 (6.4) | 1.00 | | 1.00 | |
| - Yes | 8/40 (20.0) | 3.67 (1.53 to 8.80) | 0.004 | 3.81 (1.47 to 9.84) | 0.006 |

Shaded cells indicate significant variables
Baseline variables believed by the clinical staff to be related to tuberculosis and HIV outcomes were included in this analysis.

Some variables were associated with having missing data (Table 5.3). Missing data at Month 6 were associated with treatment arm (early compared to sequential treatment arm: Odds ratio [OR] = 0.62) and age (OR = 0.97) in the univariate analysis and with treatment arm (OR = 0.61), gender (OR = 0.64), age (OR = 0.96) and history of tuberculosis (OR = 1.47) in the multivariate case. The variables indicative of disease status were not the variables that predicted having missing data. Missing data at Month 12 were associated with treatment arm (early compared to sequential treatment arm: OR = 0.37; late compared to sequential treatment arm: OR = 0.39), age (OR = 0.97), CD4+ count at Month 6 (CD4+ count < 50: OR = 3.11, CD4 count 50-200: OR = 0.83), extra pulmonary tuberculosis (OR = 2.21), and having multidrug resistant tuberculosis (OR = 3.61) in the univariate analysis and with treatment arm (early compared to sequential treatment arm: OR = 0.47; late compared to sequential treatment arm: OR = 0.47) and age (0.96) in the multivariate

analysis. In the univariate analysis some variables related to disease status were predictive of missing data. Missing data at Month 18 were associated with age (OR = 0.95), CD4+ count at Month 12 <50 cell/$mm^3$ (OR = 7.58) or missing CD4+ count (OR = 6.40) in the univariate analysis and with age (OR = 0.95), viral load (OR = 9.68) and CD4+ count < 50 cell/$mm^3$ (OR = 9.68) or missing CD4+ count (OR = 4.66) in the multivariate analysis. Missing data at Month 24 were associated with gender (OR = 0.47 and 0.45, respectively), CD4+ count at baseline (OR = 1.21 and 1.26, respectively) and having had a previous missed visit (OR = 3.67 and 3.81, respectively) in both the univariate and multivariate analyses.

Since the occurrence of missing data was associated with baseline data and previous CD4+ count measurements the data could not be assumed to be MCAR and data analysis that made the assumption of MCAR data would be biased and give incorrect results. It was not possible to rule out that the data were MNAR using the observed data only, since the definition of MNAR depended on unobserved observations. We therefore made the assumption that the CD4+ count data in the SAPiT study were either MAR or MNAR.

The SAPiT study had three treatment arms. Having more than two treatment arms raised the question whether all three arms should be compared or whether two treatments at a time should be analysed in a pairwise comparison. This choice influences the p-value if a linear mixed model is used since model based smoothing of the covariance structure takes place either on two or three arms. Efficiency can be gained by using all arms. Risk of misspecification can be reduced by assuming a treatment arm specific covariance matrix and treatment arm specific mean evolution (Molenberghs & Kenward, 2007; Molenberghs et al., 2004). We decided to include all three arms, since the scientific interest was in determining which of the three time points of antiretroviral therapy initiation resulted in the best HIV endpoints, and to possibly combine two arms if we found no difference between the two arms, while comparing those two to the third arm. At the start of the study, it was not clear that any of these arms would be the natural comparator arm. A different decision could have been made if we were comparing two active treatments and a placebo arm. In that case, it might make sense to compare each of the active treatment arms in a pairwise fashion with the placebo arm.

The CD4+ count data were not normally distributed. A square root transformation of CD4+ count produced a normal distribution, therefore modelling was done and p-values were calculated using square root transformed CD4+ count throughout. Graphs and reported summary statistics are given in the original scale, because clinical researchers find it difficult to interpret square root transformed CD4+ counts and the applied medical statistician can communicate results more clearly to colleagues using actual CD4+ counts instead of square root transformed CD4+ counts.

The Chapter would flow better and be less confusing if all graphs, summaries and tables were done using the square root of CD4+ count. We have however decided that the ability to communicate to clinical researchers was more important and include CD4+ counts in the original scale as much as possible.

The mean profiles did not follow a linear trend over time. A linear model was fitted by including both a term for time and a quadratic term for time. In all models fitted, the quadratic time effect was statistically significant.

In this clinical trial the objective is to estimate the effect of treatment assignment on clinical outcomes over all randomised individuals regardless of what treatment participants actually received. This is Estimand 1 described in the National Research Council report (2010). This clinical trial was conducted to determine what the treatment policy should be for patients co-infected with tuberculosis and HIV and not to test whether the drugs worked, since these were all licensed drugs. We were interested in identifying an appropriate public health policy regarding the time of initiation of antiretroviral treatment, therefore the traditional ITT analysis, or the de facto hypothesis is the most appropriate.

This estimand can be obtained in a parallel-group randomised trial in which outcome data are collected on all participants, regardless of whether the study treatment is received. However, CD4+ count measurements cannot be collected for participants who do not attend visits. We will perform a range of sensitivity analyses and state whether we were able to obtain Estimand 1 in each instance. The estimator defined for this study is the interaction between treatment arm and time in the longitudinal model.

## 5.3 Analysis under MCAR assumptions

Even though there was evidence in the data that the missing data process was not MCAR, we present the naïve analysis that could be done under the MCAR paradigm in this section for comparison purposes. MCAR assumptions were described in Section 2.2.1. We include both the available case analysis (Section 5.3.1) and the complete case analysis (Section 5.3.2).

### 5.3.1 Available case analysis

Under the available case analysis we perform a cross-sectional analysis of the data observed or "available" at each time point. This was done by simply analysing data as observed at each time point. For example, if a participant had data at the 6 month visit, but not at the 12 month visit this

participant was included at 6 months and excluded at 12 months. This is a suboptimal analysis that does not consider the longitudinal nature of the data. A model can be fitted that takes the longitudinal nature of the data into account; such a model would also be valid under MAR and is fitted in Section 5.4.1. The model of interest is the longitudinal model, the cross-sectional model is provided merely to describe the data observed at each visit in detail. The advantage of summarising the data observed at each time point is that it gives a description of the observed data without including any modelling assumptions.

In the available case analysis missing data were ignored and data were summarised at 6, 12, 18 and 24 months using all observed CD4+ counts (Table 5.4 and Figure 5.2).

**Table 5.4: Available case analysis of CD4+ counts (cell/mm$^3$)**

| | Raw data | | | Square root transformed data | | | p-value | |
|---|---|---|---|---|---|---|---|---|
| | **Early integrated** | **Late integrated** | **Sequen-tial** | **Early integrated** | **Late integrated** | **Sequen-tial** | **ANOVA** | **Kruskal -Wallis** |
| **6 months** | | | | | | | <0.001* | <0.001 |
| N | 173 | 159 | 154 | | | | | |
| Mean | 305.0 | 255.4 | 173.5 | 16.6 | 15.1 | 12.0 | | |
| SD | 196.1 | 166.9 | 141.4 | 5.51 | 5.28 | 5.38 | | |
| Standard error | 14.9 | 13.2 | 11.4 | 0.42 | 0.42 | 0.43 | | |
| Median | 250.0 | 225 | 142 | 15.8 | 15.0 | 11.9 | | |
| **12 months** | | | | | | | <0.001$ | <0.001 |
| N | 160 | 150 | 128 | | | | | |
| Mean | 357.7 | 302.1 | 268.6 | 18.2 | 16.7 | 15.5 | | |
| SD | 192.8 | 169.2 | 161.1 | 5.08 | 4.99 | 5.22 | | |
| Standard error | 15.2 | 13.8 | 14.2 | 0.40 | 0.41 | 0.46 | | |
| Median | 322.5 | 286.0 | 243 | 18.0 | 16.9 | 15.6 | | |
| **18 months** | | | | | | | 0.002# | 0.003 |
| N | 152 | 143 | 123 | | | | | |
| Mean | 394.3 | 353.1 | 319.0 | 19.2 | 18.2 | 17.0 | | |
| SD | 202.1 | 185.2 | 206.4 | 5.07 | 4.85 | 5.65 | | |
| Standard error | 16.4 | 15.5 | 18.6 | 0.41 | 0.41 | 0.51 | | |
| Median | 367.5 | 328.0 | 292.0 | 19.2 | 18.1 | 17.1 | | |
| **24 months** | | | | | | | 0.06 | 0.03 |
| N | 145 | 131 | 123 | | | | | |
| Mean | 428.8 | 392.7 | 379.1 | 20.1 | 19.2 | 18.6 | | |
| SD | 212.1 | 210.1 | 238.4 | 5.06 | 4.97 | 5.88 | | |
| Standard error | 17.6 | 18.4 | 21.5 | 0.42 | 0.43 | 0.53 | | |
| Median | 411.0 | 368.0 | 327.0 | 20.3 | 19.2 | 18.1 | | |

ANOVA: Analysis of variance; SD: Standard deviation
* Bonferroni pairwise comparison: All arms significantly different from each other
$ Bonferroni pairwise comparison: Early integrated treatment arm significantly different from the other two arms.
# Bonferroni pairwise comparison: Early integrated treatment arm significantly different from sequential treatment arm

The conclusion drawn was that CD4+ count was significantly different between the arms at all time points. At all visits before 24 months the early integrated treatment arm had significantly higher mean CD4+ counts than the other arms with small p-values. At 24 months the p-value was not as small as at previous time points.

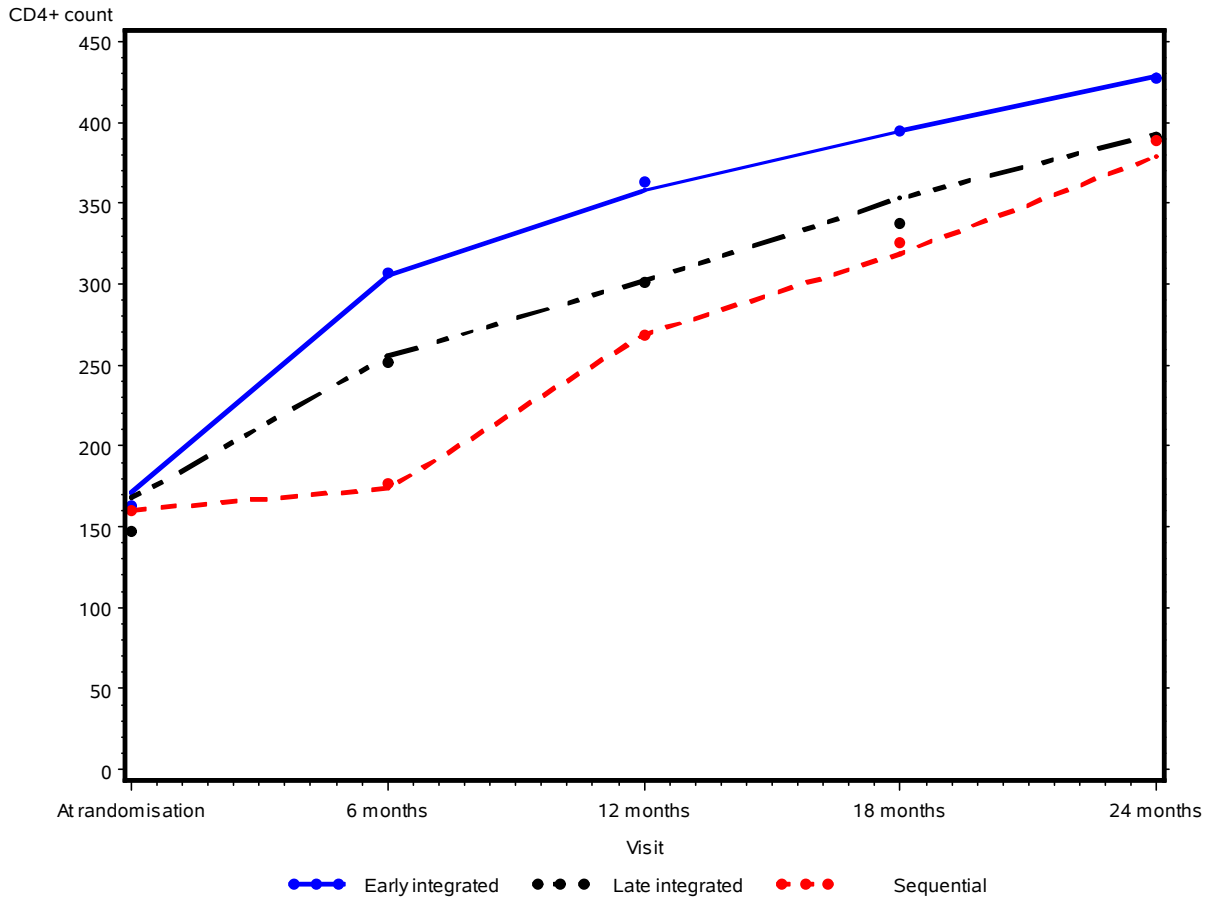



**Figure 5.2: Mean CD4+ counts (cells/mm$^3$) over time, available case analysis**

The lines indicate the available case analysis and the dots indicate the observed means under the complete case analysis.

### 5.3.2 Complete case analysis

In the complete case analysis only participants with no missing data were included in the analysis and data were summarised at 6, 12, 18 and 24 months using all observed CD4+ counts for these participants (Table 5.5 and Figure 5.3). The difference between the complete case analysis and the available case analysis was that list wise deletion was applied in the complete case analysis. In other words, if a participant had missing data at any visit, all data for this participant at all visits were deleted. List wise deletion is often the default in statistical software.

We fitted both a cross-sectional model (Table 5.5), comparing the treatment arms at each visit and a longitudinal model (Table 5.6) under complete case. The cross-sectional analysis is not intended to be similar to the longitudinal model fitted and should not be directly compared. The cross-sectional analysis done under the available case analysis (Section 5.3.1) can be contrasted to the cross-sectional analysis under the complete case analysis.

According to the cross-sectional analysis, at baseline there was no difference between the treatment groups in mean CD4+ count. Mean CD4+ count was significantly different between the arms at 6, 12 and 18 months, but not at 24 months. At all visits before 24 months the early integrated treatment arm had significantly higher mean CD4+ counts than the other arms. The same conclusions were drawn with the complete case and the available case cross-sectional analysis.

**Table 5.5: Cross-sectional complete case analysis of CD4+ counts (cell/mm$^3$)**

| | Treatment arm | | | p-value | |
|---|---|---|---|---|---|
| | **Early integrated N = 137** | **Late integrated N = 122** | **Sequential N = 108** | **ANOVA** | **Kruskal-Wallis** |
| **Baseline** | | | | 0.69 | 0.61 |
| Mean | 163.2 | 147.3 | 160.4 | | |
| Standard deviation | 126.5 | 107.2 | 106.8 | | |
| Standard error | 10.8 | 9.7 | 10.3 | | |
| Median | 146 | 123 | 142 | | |
| **6 months** | | | | <0.001* | <0.001 |
| Mean | 306.7 | 251.8 | 177.0 | | |
| Standard deviation | 199.0 | 171.2 | 144.1 | | |
| Standard error | 17.0 | 15.5 | 13.9 | | |
| Median | 254.0 | 221.0 | 143 | | |
| **12 months** | | | | <0.001$ | <0.001 |
| Mean | 351.8 | 300.5 | 268.2 | | |
| Standard deviation | 192.4 | 160.0 | 163.5 | | |
| Standard error | 16.4 | 14.5 | 15.7 | | |
| Median | 327.0 | 276.5 | 240.5 | | |
| **18 months** | | | | 0.003# | 0.007 |
| Mean | 394.5 | 337.8 | 325.9 | | |
| Standard deviation | 197.4 | 162.9 | 211.7 | | |
| Standard error | 16.9 | 14.7 | 20.4 | | |
| Median | 365.0 | 321.0 | 292.5 | | |
| **24 months** | | | | 0.15 | 0.08 |
| Mean | 427.2 | 390.4 | 388 | | |
| Standard deviation | 210.0 | 198.7 | 245.2 | | |
| Standard error | 17.9 | 18.0 | 23.6 | | |
| Median | 409.0 | 370 | 327 | | |

ANOVA: Analysis of variance
* Bonferroni pairwise comparison: All arms significantly different from each other
$ Bonferroni pairwise comparison: Early integrated treatment arm significantly different from each of the other two arms
# Bonferroni pairwise comparison: Early integrated treatment arm significantly different from sequential treatment arm

The residuals for the repeated measures linear model were not normally distributed (Figure 5.4). The residuals were normally distributed when the square root transformed CD4+ counts were modelled (Figure 5.5). The model was fitted using the square root transformed CD4+ counts. The model fitted with the raw data is also given as reference.

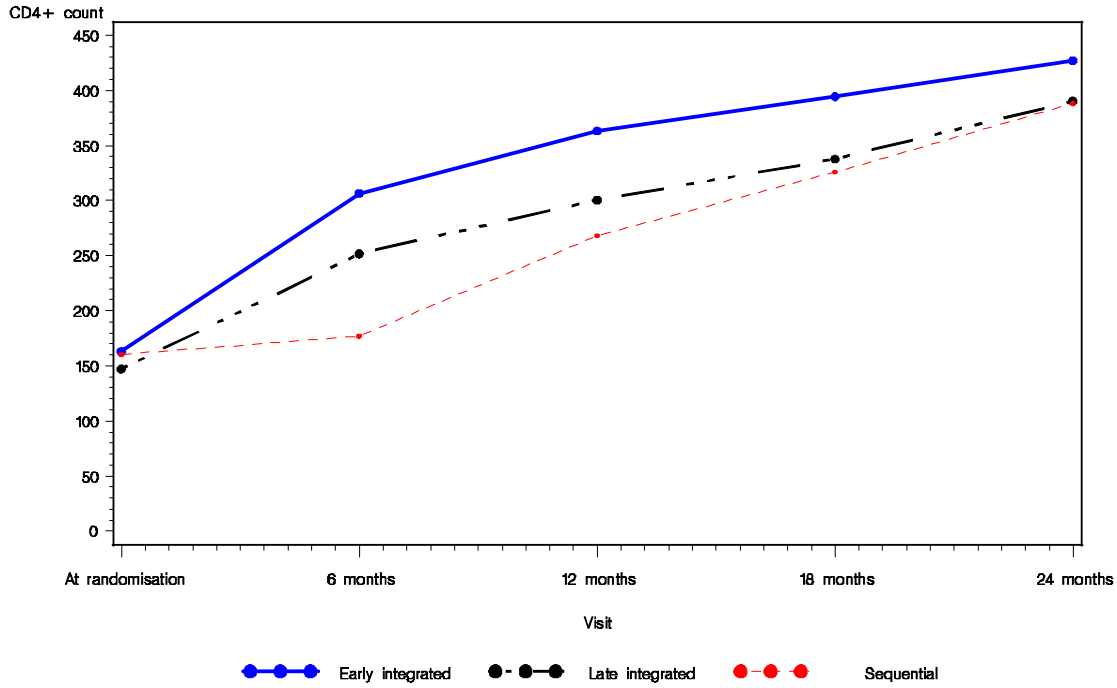



**Figure 5.3: Observed mean CD4+ counts (cells/mm$^3$) over time, complete case analysis**

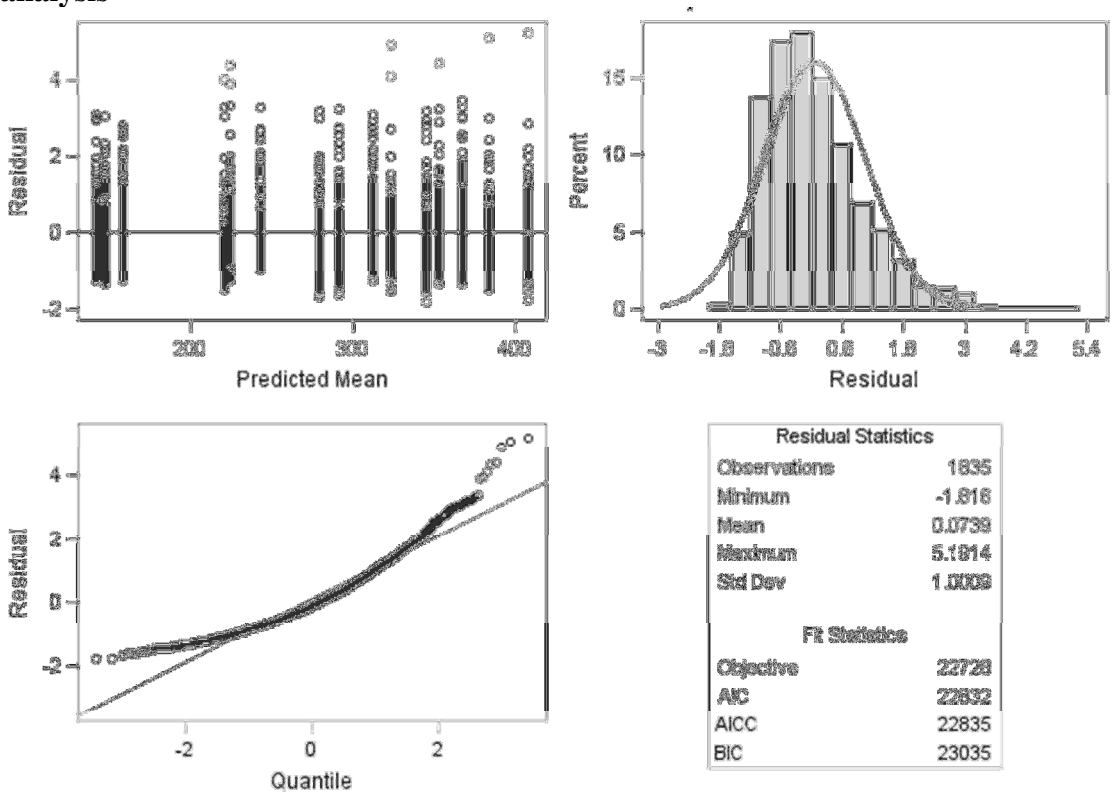



**Figure 5.4: Studentised residuals for the model fitted using SAS procedure MIXED: complete case analysis (Table 5.6)**

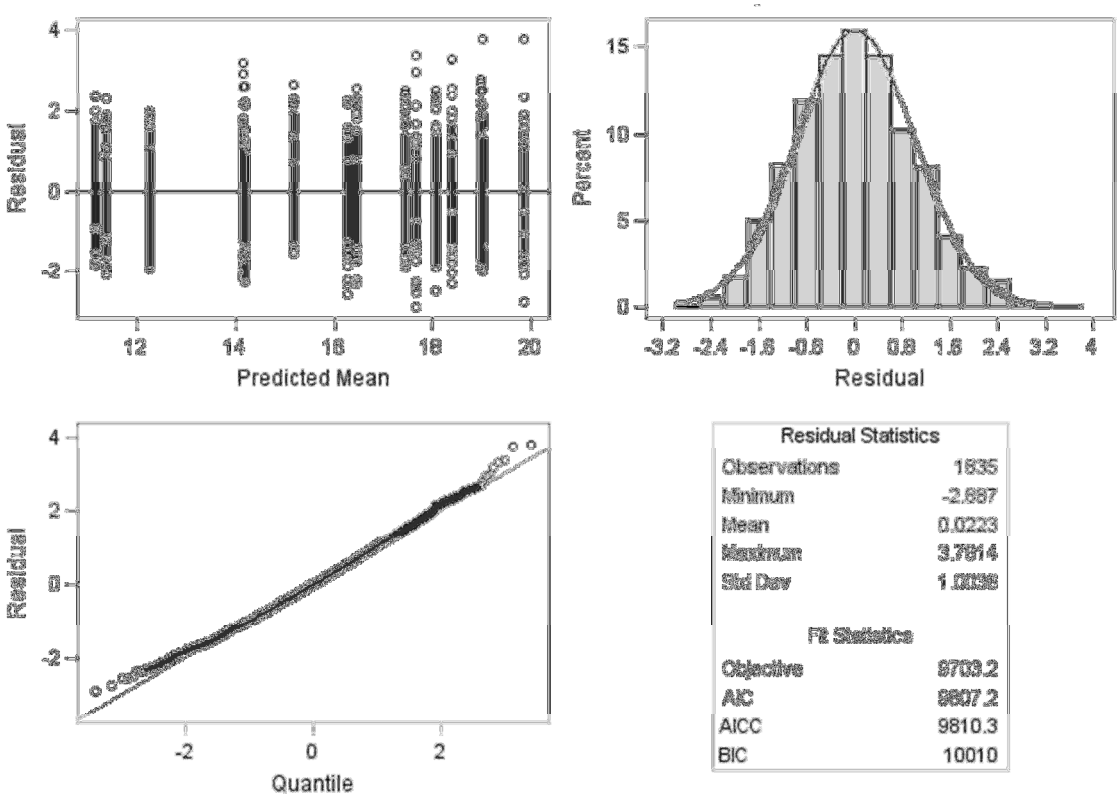



**Figure 5.5: Studentised residuals for the model fitted with square root transformed CD4+ counts using SAS procedure MIXED: complete case analysis (Table 5.6)**

The model for the complete cases was fitted by REML using procedure MIXED in SAS and assumed an unstructured covariance matrix while using the Kenward-Roger (Kenward & Roger, 1997) method for the degrees of freedom. A different covariance matrix was assumed for each treatment group. Because the unstructured covariance structure does not assume any particular pattern about the variance and covariance between measurements it was regarded to be a valid choice of covariance matrix. Fitting time either as a continuous or categorical variable would be consistent with MAR. Time was included as a continuous variable, because that leads to a more parsimonious model and enables a more concise summary of the effect of treatment arm over time. Adding time as a discrete variable will be more comparable to the cross-sectional analysis presented, but the cross-sectional analysis is merely given for additional information and was not the focus of this analysis. The mean profiles were not linear over time and a quadratic time effect was added to the model. This improved the fit of the model. The following code was used:

```
proc mixed data = completecase method = reml ;
  class pid pointx treatment;
  model cd4count = point treatment point*treatment pointsquared/ solution ddfm = kr htype = 2;
  repeated pointx / subject = pid type = un group = treatment;
  lsmeans point treatment;
run;
```

**Table 5.6: Complete case analysis using procedure MIXED in SAS**

| | Square root transformed | | | Raw data | | |
|---|---|---|---|---|---|---|
| **Effect** | **Estimate** | **Standard error** | **p-value** | **Estimate** | **Standard error** | **p-value** |
| Intercept | 7.93 | 0.51 | | 61.08 | 12.07 | |
| Time | 3.81 | 0.23 | <0.001 | 94.41 | 8.20 | <0.001 |
| Time squared | -0.34 | 0.04 | <0.001 | -7.23 | 1.32 | <0.001 |
| Early integrated treatment arm (ref: sequential treatment arm) | 0.73 | 0.68 | 0.28 | -0.55 | 15.71 | 0.97 |
| Late integrated treatment arm (ref: sequential treatment arm) | -0.46 | 0.65 | 0.48 | -15.28 | 14.47 | 0.29 |
| Interaction between treatment arm and time (early integrated compared to sequential treatment arm) | 0.14 | 0.15 | 0.34 | 11.18 | 5.88 | 0.06 |
| Interaction between treatment arm and time (late integrated compared to sequential treatment arm) | 0.22 | 0.15 | 0.15 | 9.25 | 5.73 | 0.11 |

In addition to the p-values given in the table, the F-test comparing the interaction between all three treatment arms and time adjusting for the main effects was not significant for either the raw data (p=0.15) or the square root transformed data (p=0.35). This means that the change in mean CD4+ count over time did not differ between the treatment arms. There is an indication that the interaction between the early integrated and sequential treatment arms and time could be statistically significant (p=0.06), while the interaction between the late integrated and sequential treatment arms was not statistically significant (p=0.11).

### 5.3.3 Conclusions of MCAR analysis

Under MCAR we conclude from the cross-sectional analysis that there is a significant difference between the three treatment arms at each of the time points. From the mixed model we conclude that there is no difference between the treatment arms in mean change in CD4+ count over time.

Both the available case analysis and the complete case analyses are only valid under the MCAR assumption. We have shown in Section 5.2 that the CD4+ count data in the SAPiT study were not MCAR; therefore these analyses were not valid. It is worth remembering that the complete case method is inefficient because it does not utilise all available information, even when it is valid.

None of the analyses done under the MCAR assumption estimate the estimand of interest, namely Estimand 1 defined by the National Research Council (2010). Rather, the cross-sectional analysis estimate the treatment effect at each visit for participants who attended either the respective visit or all visits, whether or not they adhered to treatment. In the longitudinal analysis this estimate is the treatment effect for participants who had complete data, whether or not they adhered to treatment. These estimates do not correspond to the ITT definition. Therefore, in addition to not being valid or optimal, the MCAR analysis also does not measure the estimand of interest.

The models include a second order polynomial of time. The treatment groups were only interacted with time and not with a second order term of time. This means that the interaction interpreted refers to the comparison of the treatment groups with respect to the linear slopes of the curves, excluding the curvature. The curvature is more pronounced later in follow-up than early in follow-up. It is thus predominantly the trend up to 6 months that is being compared for the different treatment arms.

## 5.4 Analysis under MAR assumptions

We have shown that MCAR assumptions are not valid. MAR assumptions, discussed in Section 2.2.2, could however be valid. Using the observed data we cannot show that MAR assumptions are more valid than MNAR assumptions. Under MAR assumptions we use four valid frameworks to analyse the data; these are direct likelihood-based approaches (discussed in Section 3.4), multiple imputation (discussed in Section 3.2), Bayesian analysis (discussed in Section 3.6) and inverse probability weighting (discussed in Section 3.5).

### 5.4.1 Direct likelihood-based approaches

Direct likelihood-based approaches are valid under the MAR assumption and the separability condition. Direct likelihood analyses can be done in SAS with the MIXED, GLIMMIX and NLMIXED procedures. Valid point estimates are obtained from maximizing the likelihood. The observed information matrix gives less biased results than the expected information matrix. The observed information matrix is easy to calculate and available in statistical packages. An unconditional expected information matrix calculated under the correct specification of the MAR missingness mechanism could also be used, but is not standard output of statistical software (Kenward & Molenberghs, 1998). The difference between the two could be small in practice (Molenberghs & Kenward, 2007).

Procedure MIXED in SAS uses the Fisher scoring or Newton-Raphson methods. The MIXED procedure uses the REML method to obtain estimates of parameters by minimising the likelihood of residuals. The expected Hessian matrix rather than the observed matrix is used to calculate standard errors. Procedure GLIMMIX has the same shortcomings as procedure MIXED, namely that the variability of the variance components is not used when calculating the standard errors of the fixed effects. However, procedure NLMIXED uses the full Hessian matrix for the computation of precision estimates and is fully consistent with direct likelihood estimation (SAS Institute Inc., 2004).

In procedure MIXED in SAS the following code was used to fit a model using the direct likelihood approach:

```
proc mixed data = cd4count method = reml ;
  class pointx treatment pid ;
  model cd4count = treatment point point*treatment pointsquared/ solution ddfm = kr;
  repeated pointx / subject = pid type = un group = treatment;
run;
```

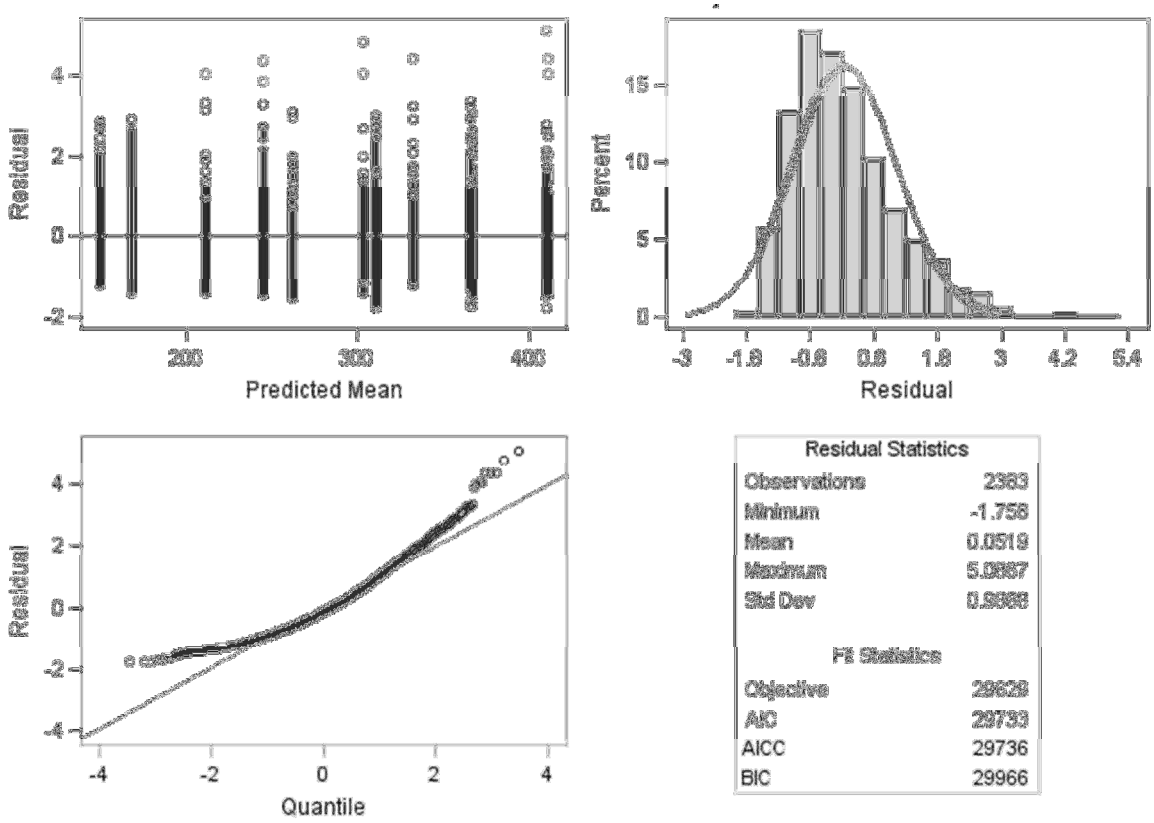



**Figure 5.6: Studentised residuals for the model fitted using procedure MIXED in SAS for the direct likelihood-based analysis under MAR assumptions**

The "ddfm = kr" statement ensures that the Kenward-Roger (Kenward & Roger, 1997) correction is applied when calculating degrees of freedom. An unstructured covariance matrix was assumed, while the group = treatment statement requests a different correlation matrix for each treatment group.

As shown in Figure 5.6 the residuals for the repeated measures linear model did not follow a normal distribution. The residuals were normally distributed when the square root transformed CD4+ counts were modelled (Figure 5.7).

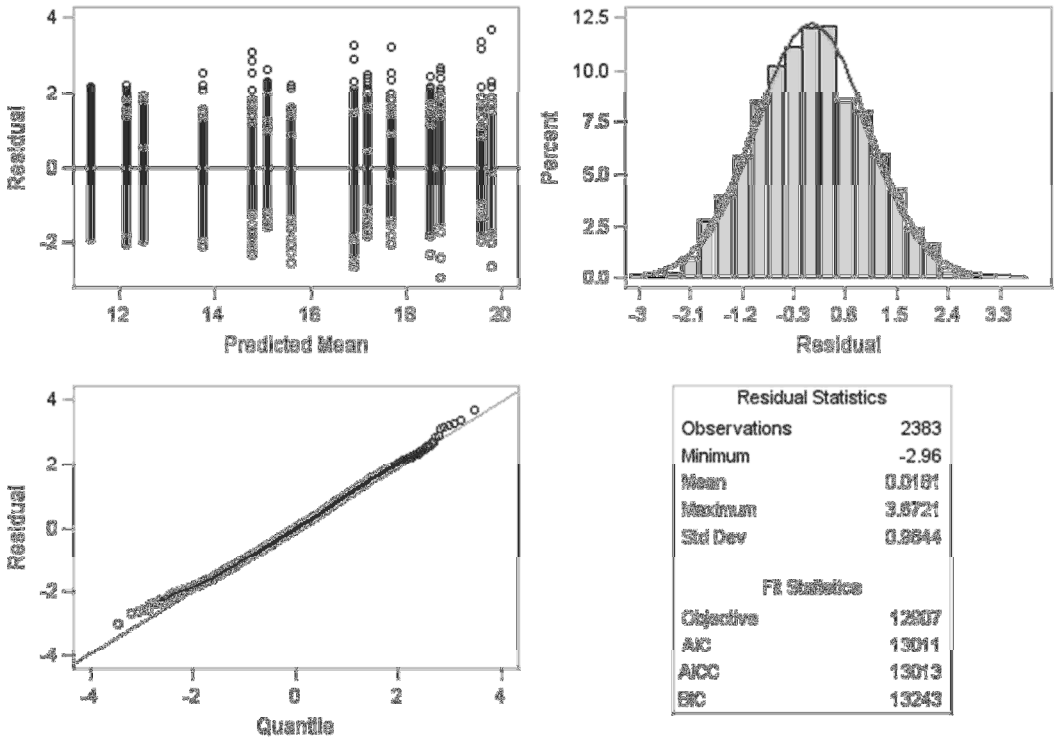



**Figure 5.7: Studentised residuals for the model fitted with square root transformed CD4+ count using procedure MIXED in SAS for the direct likelihood-based analysis under MAR**

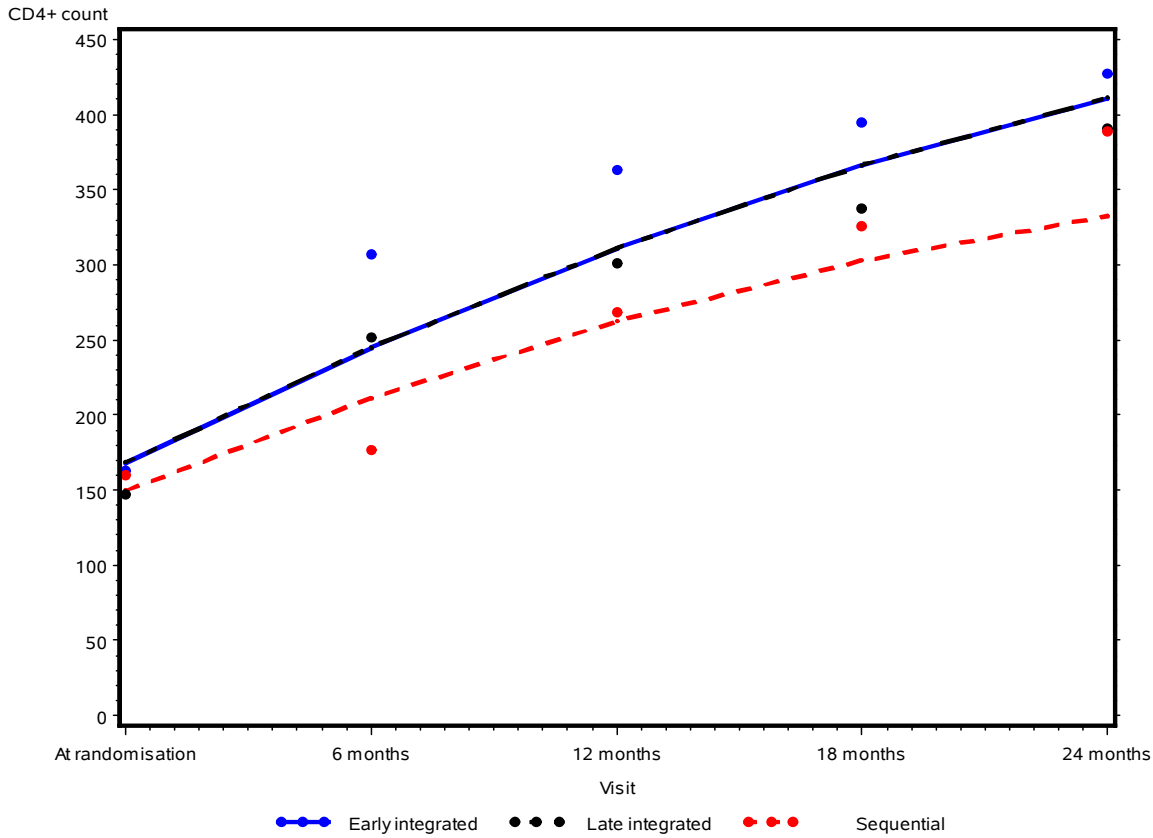



**Figure 5.8: Mean CD4+ counts (cells/mm$^3$) over time, direct likelihood-based approach (mixed model) analysis under MAR assumptions**
The lines indicate fitted model and the dots indicate the observed means

**Table 5.7: CD4+ count over time. Direct likelihood-based parameter estimates and standard errors when using procedure MIXED in SAS under MAR assumptions**

| | Raw data | | | Square root transformed data | | |
|---|---|---|---|---|---|---|
| **Effect** | **Estimate** | **Standard error** | **p-value** | **Estimate** | **Standard error** | **p-value** |
| Intercept | 77.19 | 10.11 | | 8.50 | 0.41 | |
| Early integrated treatment arm (ref: sequential treatment arm) | 3.35 | 12.88 | 0.80 | 0.86 | 0.54 | 0.11 |
| Late integrated treatment arm (ref: sequential treatment arm) | 3.76 | 12.00 | 0.75 | 0.47 | 0.51 | 0.36 |
| Time | 77.72 | 7.36 | <0.001 | 3.14 | 0.21 | <0.001 |
| Time squared | -5.33 | 1.25 | <0.001 | -0.26 | 0.03 | <0.001 |
| Interaction between treatment arm and time (early integrated compared to sequential treatment arm) | 14.94 | 5.31 | 0.005 | 0.25 | 0.14 | 0.08 |
| Interaction between treatment arm and time (late integrated compared to sequential treatment arm) | 14.97 | 5.18 | 0.004 | 0.29 | 0.14 | 0.04 |

The p-value for the interaction between time and the treatment arms, comparing the late integrated treatment arm to the sequential treatment arm (p=0.04) and comparing the early integrated treatment arm to the sequential treatment arm (p=0.08) was small when using the square root transformed data. This means that we can conclude that the change in mean CD4+ count over time was higher in the late integrated treatment arm and early integrated treatment arm than in the sequential treatment arm. In addition to the p-values given in the table, the F-test comparing the interaction between all three treatment arms and time adjusted for the main effects in the model had p=0.006 for the raw data and p=0.009 for the square root transformed data.

### 5.4.2 Multiple imputation

Multiple imputation can be done in SAS by using two procedures, MI and MIANALYZE. Procedure MI does the multiple imputation, corresponding to Step 1 in Section 3.2, after which complete data methods are used to analyse each completed data set, using the appropriate SAS procedure; corresponding to Step 2 in Section 3.2. Afterwards the data are combined using the SAS procedure MIANALYZE, as described in Step 3 (Horton & Kleinman, 2007).

Procedure MI was used to generate 100 different imputed data sets and procedure MIANALYZE was used to combine the results of the 100 analyses on these data sets. In the imputer's model square root CD4+ count at 6, 12, 18 and 24 months and log viral load at baseline were imputed and WHO status, square root CD4+ count at baseline, age, gender, history of tuberculosis, and whether the participant had extra-pulmonary tuberculosis or multidrug resistant tuberculosis were included as covariates in this model. These variables were chosen because the clinical staff believed that these variables were related to tuberculosis and HIV outcomes. It is strongly suggested that multiple imputation should include all variables that could be related to the missingness process,

therefore all possibly relevant baseline variables were included in the imputer's model. Separate imputations were done for each of the three treatment arms.

The multiple imputation procedure in SAS software assumes that the missing data are MAR. Since the missing data pattern in the CD4+ count data was not monotone, the MCMC imputation method was used. It assumes multivariate normality to create multiple imputations by drawing simulations from a Bayesian predictive distribution. The EM algorithm was used and the means and covariances from complete cases were taken as the initial estimates. A single chain was used for all the imputations. The procedure used a non-informative Jeffreys prior to derive the posterior mode from the EM algorithm as the starting values for the MCMC process (SAS Institute Inc., 2004).

The following code was used in SAS to do the multiple imputation step in procedure MI, using the square root of CD4+ count:

```
proc mi data = withcovariate out = full nimpute = 100 seed = 1 round = 1 1 1 1 0.001 . maximum =
      1000 1000 1000 1000 8 minimum = 1 1 1 1 0;
  em initial = cc;
  mcmc impute = full chain = single timeplot (mean(cd4at6 cd4at12 cd4at18 cd4at24)) acfplot
      (mean (cd4at6 cd4at12 cd4at18 cd4at24));
  var cd4at6 cd4at12 cd4at18 cd4at24 logvl who baselinecd4count age gender historytb extrapul
      mdr;
  by treatment;
run;
```

**Table 5.8:Multiple imputation variance (Results of multiple imputation in Table 5.10)**

| | Early integrated treatment | | | | Late integrated treatment | | | | Sequential treatment | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Visit** | **Between** | **Within** | **Total** | **df** | **Between** | **Within** | **Total** | **df** | **Between** | **Within** | **Total** | **df** |
| 6 months | 0.018 | 0.145 | 0.164 | 183 | 0.020 | 0.138 | 0.158 | 180 | 0.020 | 0.148 | 0.168 | 179 |
| 12 months | 0.014 | 0.128 | 0.142 | 187 | 0.019 | 0.115 | 0.134 | 174 | 0.042 | 0.146 | 0.188 | 151 |
| 18 months | 0.027 | 0.120 | 0.148 | 162 | 0.023 | 0.113 | 0.136 | 166 | 0.074 | 0.176 | 0.250 | 130 |
| 24 months | 0.031 | 0.125 | 0.157 | 158 | 0.048 | 0.130 | 0.178 | 139 | 0.093 | 0.178 | 0.273 | 118 |

df = degrees of freedom

Table 5.8 indicates that the major source of variance was the within imputation variance, the between imputation variance added more to the overall variance at the later time points, where more data were missing than at the earlier time points.

The relative increase in variance showed that the uncertainty caused by the missing data increased the variance more over time in the sequential treatment arm. This ratio reached a high of 0.528 at 24 months. The smallest fraction of missing information was 0.097 at 12 months in the early

integrated treatment arm, with the highest fraction, 0.347, at 24 months in the sequential treatment arm.

**Table 5.9: Relative increase in variance and fraction missing information in procedure MI (Results of multiple imputation in Table 5.10)**

| | Early integrated treatment | | Late integrated treatment | | Sequential treatment | |
|---|---|---|---|---|---|---|
| **Visit** | **Relative increase in variance** | **Fraction missing information** | **Relative increase in variance** | **Fraction missing information** | **Relative increase in variance** | **Fraction missing information** |
| 6 months | 0.127 | 0.113 | 0.145 | 0.127 | 0.140 | 0.123 |
| 12 months | 0.107 | 0.097 | 0.170 | 0.146 | 0.288 | 0.224 |
| 18 months | 0.228 | 0.186 | 0.211 | 0.174 | 0.423 | 0.299 |
| 24 months | 0.250 | 0.201 | 0.369 | 0.271 | 0.528 | 0.347 |

The timeplot and acfplot statements in SAS display time-series and autocorrelation plots to check convergence for the single chain. It uses the MCMC method to create an iteration plot for the successive estimates of the mean of the square root CD4+ count at each of the 6-monthly visits. Iterations during the burn-in period are indicated with negative iteration numbers. The plots are given in Figure 5.9 and Figure 5.10 for the early integrated treatment arm at 6 months. No evidence was found of a trend or a significant positive or negative autocorrelation in any of the variables. The plots for the other treatment arms are similar and are not displayed.

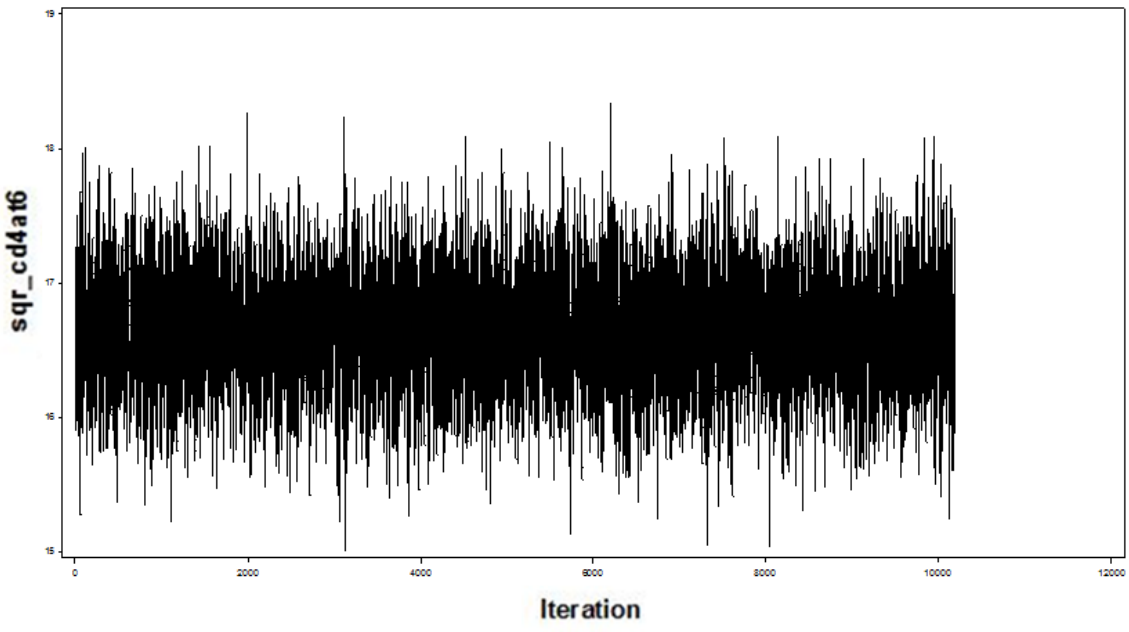



**Figure 5.9: Time series plot for the mean square root CD4+ count at 6 months, early integrated treatment arm**

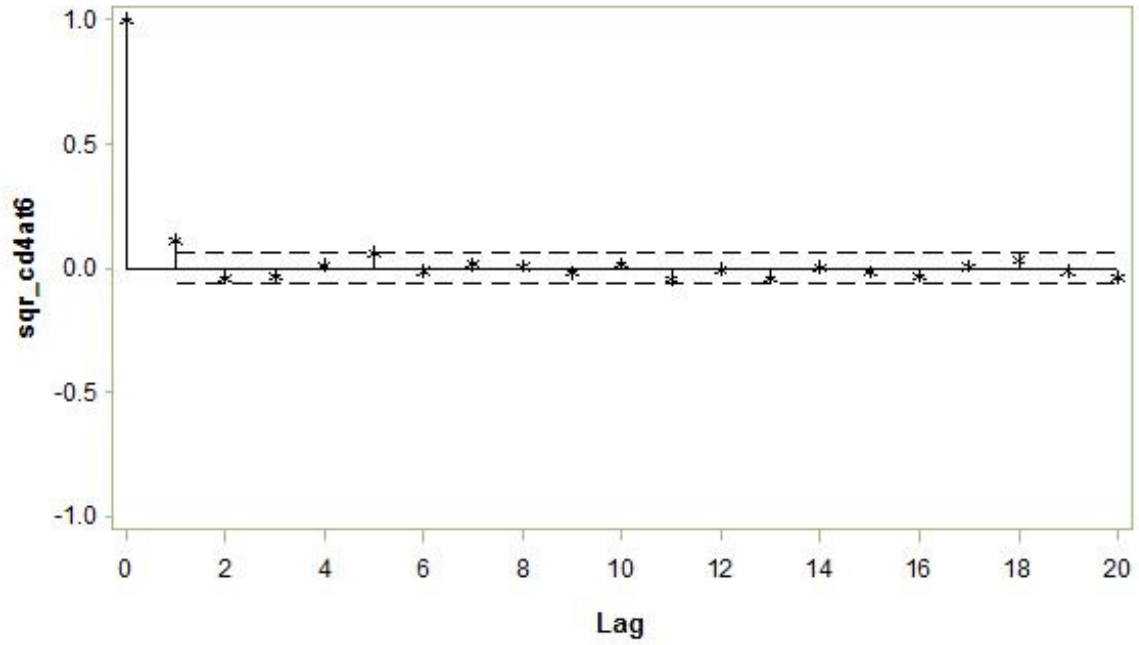



**Figure 5.10: Autocorrelation plot with 95% confidence interval for the mean square root CD4+ count at 6 months, early integrated treatment arm**

After multiple imputation of the missing data was done with procedure MI, the data were analysed using a SAS procedure with a by imputation statement, thus doing the analysis separately for each imputation. The MIANALYZE procedure then reads the parameter estimates and associated standard errors, or the covariance matrix, from a data set produced by this SAS procedure, separately for each imputed data set. The MIANALYZE procedure used the results from each imputation to derive valid univariate inference for these parameters.

Not all parameters of interest or all statistical tests can easily be obtained this way, since not all parameter estimates and the associated covariance matrices can be calculated directly in a SAS procedure. The analyses can be done, but require special techniques and is not a straight forward application of the usual SAS procedures. For example the sample means and covariance matrices for means can be obtained from the CORR procedure, which was written to generate correlations and not means (SAS Institute Inc., 2004).

Some regression procedures, such as procedure REG and LOGISTIC create a data set that contains both the parameter estimates for the regression coefficients and their covariance matrix. Other regression procedures such as GLM, MIXED and GENMOD do not generate such data sets. In procedures MIXED and GENMOD ODS output statements can be used to save the parameter estimates in a data set and the covariance matrix in a separate data set. These data sets can then be read into the MIANALYZE procedure. The PARMS option is used to read the parameter estimates and the COVB option is used to read the covariance matrix. For procedure GLM the ODS output

statement can be used to save parameter estimates and standard errors in a data set and the $(X'X)^{-1}$ matrix in another data set. These data sets are then read in the MIANALYZE procedure with the PARMS option and the XPXI option. Other examples of procedures that do not automatically generate the parameter of interest and covariance matric are correlation coefficients between two variables and the ratios of variable means (SAS Institute Inc., 2004).

One of the advantages of multiple imputation as a technique to handle missing data, is that multiple imputation can be used with the statistical method of your choice. There is no need to change the statistical method you wanted to use in order to adjust for missing data. To illustrate the flexibility and versatility of multiple imputation, we first present a cross-sectional analysis of the imputed data and then we fit a longitudinal model with maximum likelihood analysis of repeated measures, using mixed models (Table 5.12). We acknowledge that the cross-sectional analysis is sub-optimal and that the two models are not directly comparable.

The following code was used in SAS to analyse the imputed data using procedure GLM as the second step of multiple imputation. The same code was repeated for each of the 6-monthly CD4+ count variables, and is not repeated here. Means were given in the table and could be computed using the UNIVARIATE procedure; however procedure MI calculates means for each of the variables and doing the extra steps with procedure UNIVARIATE and MIANALYZE gave the same means as generated by procedure MI. Medians could only be obtained by using procedure SURVEYMEANS because other procedures do not provide the standard errors for medians. The data were square root transformed before the imputations were done, and the imputed values were squared before the data were summarised.

```
Proc glm data=full;
  model cd4at6= treatment/inverse;
  by _Imputation_;
  ods output ParameterEstimates=glmparms InvXPX=glmxpxi;
quit;

proc mianalyze parms=glmparms xpxi=glmxpxi ;
  modeleffects Intercept treatment;
run;
```

Comparing the means and standard errors in Table 5.4 and Table 5.10 is interesting. All the standard errors were lower with the multiple imputation than when using the available case analysis. This was because the multiple imputation analysis is more efficient than the available case analysis and included auxiliary variables in the imputation. All the means at Months 6 and 12 were higher with the multiple imputed data than with the available case analysis. At Months 18 and 24, the means for the late integrated treatment arm were lower with the multiple imputation,

while the means for the early integrated and sequential treatment arms were higher with the multiple imputation than with the available case analysis. The interpretation of the p-values was the same for the available case analysis and the multiple imputation analysis. Having higher mean CD4+ counts at most visits with the multiple imputation analysis than with the available case analysis was surprising.

**Table 5.10: Cross sectional summaries of CD4+ counts (cell/mm$^3$) after multiple imputation of missing values**

| | Early integrated treatment arm N = 214 | | Late integrated treatment arm N = 215 | | Sequential treatment arm N = 213 | | p-value* |
|---|---|---|---|---|---|---|---|
| | Mean (standard error) | Median | Mean (standard error) | Median | Mean (standard error) | Median | |
| 6 months | 308.1 (14.5) | 253.4 | 266.1 (13.1) | 230.9 | 175.5 (11.0) | 139.5 | <0.001 |
| 12 months | 360.9 (14.4) | 322.0 | 309.7 (12.7) | 286.7 | 267.8 (13.9) | 236.9 | <0.001 |
| 18 months | 390.6 (15.2) | 361.3 | 359.1 (14.1) | 327.5 | 314.4 (17.8) | 274.9 | 0.007 |
| 24 months | 424.0 (16.3) | 399.0 | 411.8 (17.7) | 373.1 | 371.8 (20.1) | 320.2 | 0.047 |

* Calculated using procedures GLM and MIANALYZE in SAS

Most of the medians were similar between the available case analysis and the multiple imputation analysis, exceptions were the median in the sequential arm at 18 months and the median at 24 months in the early integrated treatment arm, both were lower with the multiple imputation analysis.

We now move away from the cross-sectional analysis and give the longitudinal analysis. Procedure MI was the same as previously given (Step 1 of the multiple imputation process). The following code was used in SAS to analyse the imputed data using procedure MIXED (Step 2 of the multiple imputation process). This differs from the code used under the likelihood-based analysis in Section 5.4.1 only in the addition of the by _imputation_ statement and the ods output statement. The by statement ensures that a separate analysis is done for each imputed dataset and the ods statement writes the results to datasets that are used in the next step in procedure MIANALYZE. The variance information for this model is given in Table 5.11.

```
proc mixed data = full method = reml;
  class pid treatment pointx;
  model cd4square = point treatment point*treatment  pointsquared / solution covb ddfm = kr ;
  repeated pointx / subject = pid type = un group = treatment;
  by _imputation_;
  ods output solutionf = mixparms covb = mixcovb;
run;
```

A data step is needed to rename the effect variable, in order to have unique effect variables for the treatment and week combinations, since these are captured in two different variables in the mixparms data set.

```
proc mianalyze parms=mixparms ;
    modeleffects Intercept visit trt1 trt2 trt1_week trt2_week;
  run;
```

**Table 5.11: Variance, relative increase in variance and fraction missing information in procedure MI (Results in Table 5.12)**

| Estimate | Variance between | Variance within | Variance total | Df | Relative increase in variance | Fraction missing information | Relative efficiency |
|---|---|---|---|---|---|---|---|
| Intercept | 0.057 | 0.165 | 0.223 | 1465 | 0.351 | 0.261 | 0.997 |
| Early integrated treatment arm (ref: sequential treatment arm) | 0.053 | 0.296 | 0.350 | 4191 | 0.182 | 0.154 | 0.998 |
| Late integrated treatment arm (ref: sequential treatment arm) | 0.050 | 0.283 | 0.334 | 4348 | 0.178 | 0.151 | 0.998 |
| Time | 0.025 | 0.035 | 0.061 | 554 | 0.732 | 0.425 | 0.996 |
| Time squared | 0.0005 | 0.001 | 0.001 | 687 | 0.612 | 0.381 | 0.996 |
| Interaction between treatment arm and time (early integrated compared to sequential treatment arm) | 0.012 | 0.014 | 0.026 | 447 | 0.889 | 0.473 | 0.995 |
| Interaction between treatment arm and time (late integrated compared to sequential treatment arm) | 0.009 | 0.014 | 0.023 | 586 | 0.698 | 0.413 | 0.996 |

df: Degrees of freedom

Multiple imputation and maximum likelihood methods are expected to have similar results when the imputation model is congenial, i.e. the imputation model includes the same variables as the analytic model. A major difference between the two models is that the likelihood-based model did not include any covariates, while the model fitted under multiple imputation included several covariates in the imputer's model. This is strongly encouraged, because the biggest problem with the imputer's model is excluding variables that are associated with the outcome. Rubin (1996) recommended that as many variables as possible are included in the imputer's model. If the goal of this chapter was to compare results obtained using different available methods, the multiple imputation analysis should have been done without including these covariates in the imputer's model. However, the goal is to show how available missing data methods can be used optimally to analyse data, therefore the best possible imputer's model was used. In addition to the model including covariates a secondary model was also fitted where the imputation model did not include any covariates. This was done to investigate whether the inclusion of covariates played an important role in the results observed.

**Table 5.12: CD4+ count longitudinal analysis. Direct likelihood-based parameter estimates and standard errors when using procedure MIXED in SAS, with and without multiple imputation of square root CD4+ counts ($\sqrt{cell/mm^3}$)**

| | Multiple imputation with auxiliary covariates | | | Multiple imputation without auxiliary covariates | | |
|---|---|---|---|---|---|---|
| **Effect** | **Estimate** | **Standard error** | **p-value** | **Estimate** | **Standard error** | **p-value** |
| Intercept | 8.88 | 0.47 | | 9.19 | 0.47 | |
| Early integrated treatment arm (ref: sequential treatment arm) | 1.12 | 0.59 | 0.06 | 0.92 | 0.61 | 0.13 |
| Late integrated treatment arm (ref: sequential treatment arm) | 0.75 | 0.58 | 0.20 | 0.58 | 0.59 | 0.32 |
| Time | 3.01 | 0.25 | <0.001 | 2.76 | 0.22 | <0.001 |
| Time squared | -0.23 | 0.04 | <0.001 | -0.21 | 0.04 | <0.001 |
| Interaction between treatment arm and time (early integrated compared to sequential treatment arm) | 0.08 | 0.16 | 0.62 | 0.22 | 0.15 | 0.14 |
| Interaction between treatment arm and time (late integrated compared to sequential treatment arm) | 0.12 | 0.15 | 0.43 | 0.20 | 0.14 | 0.17 |

Comparing the model fitted using likelihood-based methods with (Table 5.12) and without (Table 5.7) multiple imputation, we draw a different conclusion for the interaction between the two treatment arms and time. This interaction is estimated to be larger when the data are not imputed than when the data are imputed. The standard errors are similar in both models, for all the variables. The covariates included in the imputer's model did make a difference in the effect estimates, which differed substantially between the model with and without covariates. However, the inclusion of covariates did not explain all the difference between the likelihood-based model and the model fitted after multiple imputation of missing data. The results differed slightly between the likelihood-based model and the model fitted after multiple imputation that did not include covariates in the imputer's model. This is because the multiple imputation is done treating time as a categorical variable. Outcome data of all the time points are included in the multiple imputation, but the longitudinal nature of the outcome data is not taken into account. Fitting likelihood-based models, the longitudinal nature of the data is taken into account.

We conclude that there is no significant treatment by time interaction from both the model including covariates in the imputer's model and from the model that did not include covariates in the imputer's model.

Multiple imputation offers advantages if covariates are missing, because likelihood-based analyses in these instances might be impracticable. However, in this instance where covariates were observed and responses only were missing multiple imputation added little advantage over a likelihood-based analysis.

Multiple imputation assumed that the data are MAR and used observed data to imputer unobserved data. If future missing measurements could not be predicted from the past observed data combined with the covariates, it was unlikely that the multiple imputations are correct.

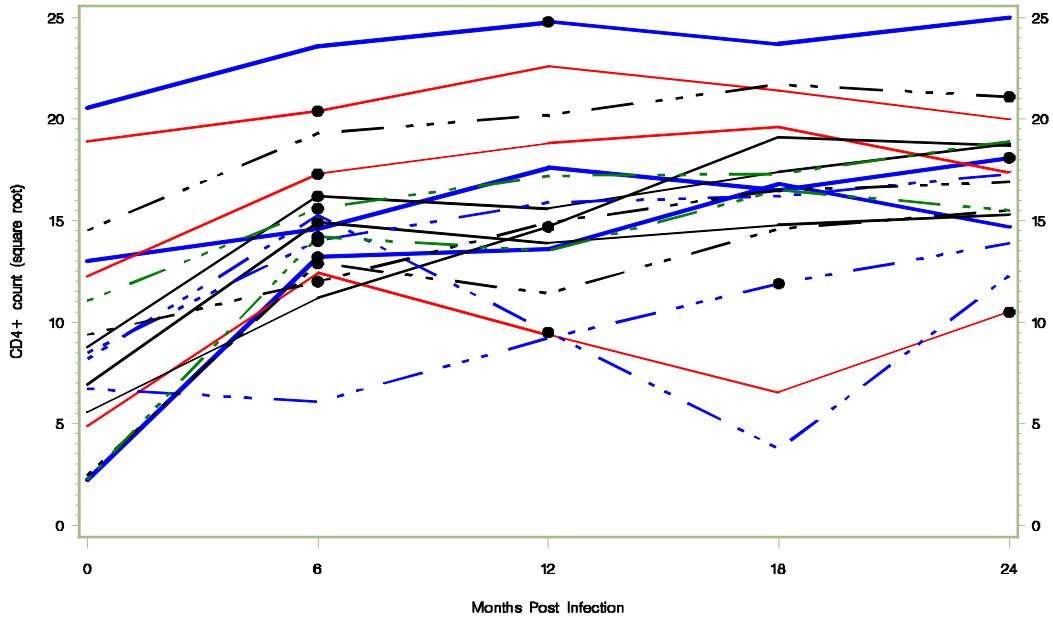



**Figure 5.11: Square root CD4+ counts over time using multiple imputation, by participant. Early integrated treatment arm.**
The graph includes only participants who died and had missing CD4+ counts from the time they died. Time of death is indicated with a black solid dot. CD4+ counts from the black solid dot onwards are imputed CD4+ counts.

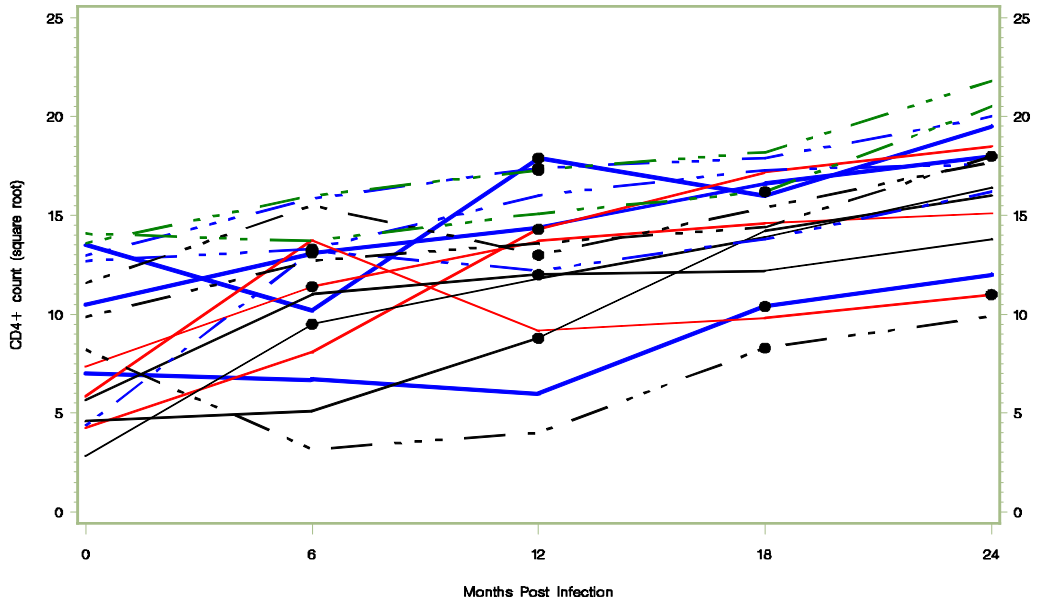



**Figure 5.12: Square root CD4+ counts over time using multiple imputation, by participant. Late integrated treatment arm.**
The graph includes only participants who died and had missing CD4+ counts from the time they died. Time of death is indicated with a black solid dot. CD4+ counts from the black solid dot onwards are imputed.

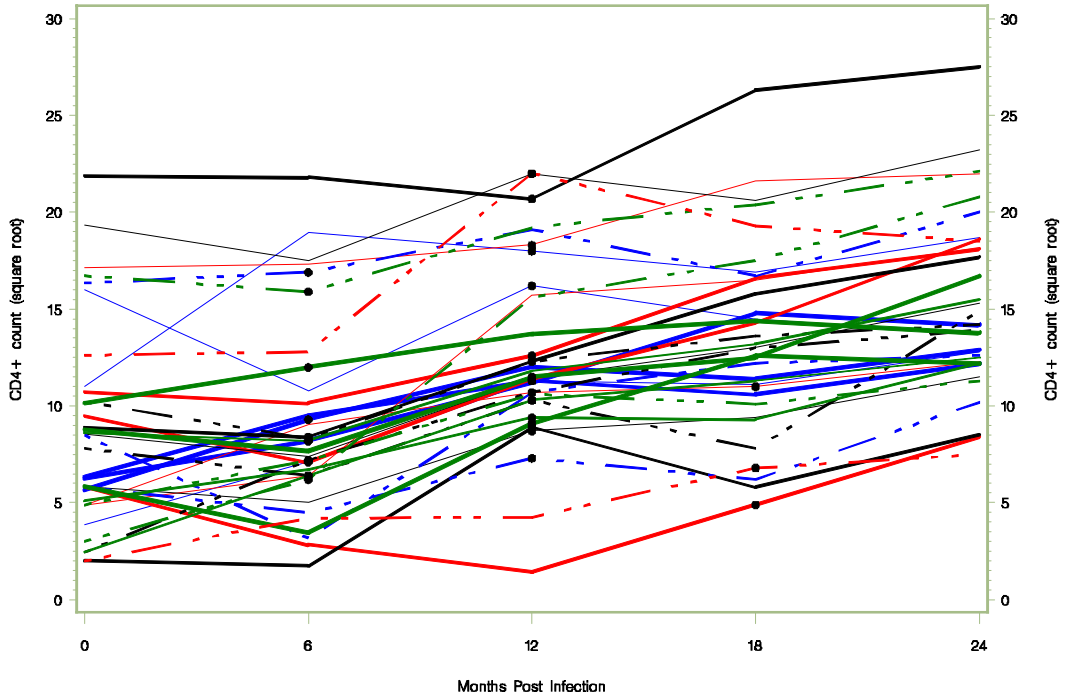



**Figure 5.13: Square root CD4+ counts over time using multiple imputation, by participant. Sequential treatment arm.**
The graph includes only participants who died and had missing CD4+ counts from the time they died. Time of death is indicated with a black solid dot. CD4+ counts from the black solid dot onwards are imputed CD4+ counts.

As can be seen from Figure 5.11, Figure 5.12 and Figure 5.13 the multiple imputation process did not impute lower CD4+ counts for participants who died, in fact for most participants CD4+ counts were higher after death than before.

### 5.4.3 Bayesian MAR analysis

It is fairly straightforward to accommodate missing data while fitting a Bayesian model. The Bayesian model fitted in this section does not take the missing data mechanism into account, but fits all participants, those with and without missing data; it is thus valid under MAR. Its validity is the same as under direct likelihood. Bayesian analysis provides a natural way of handling the missing data, because a probability distribution is estimated for each missing value, allowing for uncertainty to be captured. Missing data are treated as additional unknown quantities, thus no distinction is made between missing data and unknown parameters. OpenBugs will simulate values for the missing observations according to the specified likelihood distribution, given the values of the relevant parameters. The Bayesian model, when the missing data mechanism is ignorable, is the same model that would be used for a complete case analysis. If a model was to be fitted under the complete case, list wise deletion would be done in the data set.

The definition of the statistical notation is given in Section 5.1.4. We fitted a longitudinal model with an unstructured covariance matrix. Consider

$$\boldsymbol{Y}_i|Z_i = k \sim \text{Mutivariate Normal}(\boldsymbol{\mu}_k, \boldsymbol{\Omega}_k)$$

where $Z_i = k$ indicates that the observation belongs to treatment *k.*. For example, if $Z_i = 1$ then this implies $E_i$= 1 and $L_i$ = 0; if $Z_i = 2$ then this implies $E_i$= 0 and $L_i$ = 1 and if $Z_i = 3$ then this implies $E_i$= 0 and $L_i$ = 0. The model fitted was as follows:

$$\mu_{ij} = \beta_0 + \beta_1 E_i + \beta_2 L_i + \beta_3 t_{ij} + (\beta_4 E_i + \beta_5 L_i) t_{ij} + \beta_6 t_{ij}^2$$

This model fits a different unstructured covariance matrix for each of the treatment arms. Uninformative priors were chosen for the unknown parameters of the model of interest. Different sets of prior distributions were assigned. In Set 1, the $\beta$ parameters $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6)$ are assigned Normal(0, 100 000) priors and the inverse of Ω was assigned a Wishart(I, 5) prior, where I is the identity matrix. Prior Set 2 assigned all the $\beta$ parameters a Normal(0, 1000) prior and the inverse of Ω was assigned a Wishart(A, 5) prior, where A is a diagonal matrix with 0.1 on the diagonal. Prior Set 3 assigned all the $\beta$ parameters a Normal(0, 10) prior and the inverse of Ω a Wishart(A, 5) prior, where A is a diagonal matrix with 10 on the diagonal.

All the models were fitted using OpenBugs, which uses MCMC methods and run for 40000 iterations, including 20000 burn-in iterations. Two chains with different starting values were used and convergence was assumed if a visual inspection of the trace plots was satisfactory. All the runs discussed converged.

The following code was used in OpenBugs:

```
model {
        for(i in 1:N) {
      cd4square[i,1:5] ~ dmnorm(mu[i,1:5], omega[S[i], ,])
           S[i] <- 1+1*equals(dummyearly[i],1) + 2*equals(dummylate[i],1)
       for(j in 1:W) {
         mu[i,j] <- beta[1] +beta[2]*(j) + beta[3]*dummyearly[i] + beta[4]*dummylate[i] +
beta[5]*dummyearly[i]*(j)+beta[6]*dummylate[i]*(j) +beta[7]*(j)*(j)}                    }

 # Priors
for(k in 1:3){
omega[k,1:5,1:5] ~ dwish(R[k, ,],5)
sigma[k,1:5,1:5] <- inverse(omega[k, ,])  }

    for(k in 1:7) {
    beta[k] ~ dnorm(0, 0.00001)         }}
```

**Table 5.13: CD4+ count (square root) posterior means, standard deviations and credible intervals according to Bayesian analysis under MAR assumptions**

| | **Prior set 1** $\beta \sim Normal(0, 100\,000)$ $\Omega \sim Wishart(I,5)$ I: Identity matrix | | | **Prior set 2** $\beta \sim Normal(0, 1000)$ $\Omega \sim Wishart(A,5)$ A: diag (0.1, 0.1, 0.1, 0.1, 0.1) | | | **Prior set 3** $\beta \sim Normal(0, 10)$ $\Omega \sim Wishart(A,5)$ A: diag (10, 10, 10, 10, 10) | | |
|---|---|---|---|---|---|---|---|---|---|
| **Effect** | **Mean** | **SD** | **95% credible interval** | **Mean** | **SD** | **95% credible interval** | **Mean** | **SD** | **95% credible interval** |
| $\beta_0$ | 9.02 | 0.48 | | 9.02 | 0.48 | | 8.79 | 0.47 | |
| $\beta_1$ | 1.09 | 0.62 | -0.12; 2.3 | 1.09 | 0.62 | -0.12; 2.30 | 1.23 | 0.60 | 0.04; 2.40 |
| $\beta_2$ | 0.41 | 0.59 | -0.73; 1.55 | 0.41 | 0.58 | -0.73; 1.55 | 0.58 | 0.57 | -0.53; 1.70 |
| $\beta_3$ | 2.71 | 0.24 | 2.23; 3.17 | 2.71 | 0.23 | 2.25; 3.17 | 2.79 | 0.24 | 2.31; 3.26 |
| $\beta_4$ | 0.28 | 0.15 | 0.02; 0.58 | 0.28 | 0.15 | -0.01; 0.58 | 0.27 | 0.15 | -0.03; 0.58 |
| $\beta_5$ | 0.37 | 0.15 | 0.08; 0.67 | 0.38 | 0.15 | 0.09; 0.67 | 0.34 | 0.15 | 0.05; 0.64 |
| $\beta_6$ | -0.21 | 0.04 | -0.29; -0.14 | -0.21 | 0.04 | -0.29; -0.14 | -0.22 | 0.04 | -0.30; -0.14 |

SD: Standard deviation
$\beta_0$: Intercept, $\beta_1$ Early integrated treatment arm (ref: sequential treatment arm); $\beta_2$: Late integrated treatment arm (ref: sequential treatment arm), $\beta_3$: Time; $\beta_4$ Interaction between treatment arm and time (early integrated compared to sequential treatment arm) $\beta_5$ Interaction between treatment arm and time (late integrated compared to sequential treatment arm); $\beta_6$: Time squared
Highlighted sections indicate statistical significance

The choice of vague prior did not change the results appreciably in the MAR analysis (Table 5.13), and the results were almost identical with prior sets 1 and 2. The results are slightly different under prior set 3. This set of priors makes an assumption about the prior distribution of Ω that is very different from the observed covariance matrices, and the prior distributions of the β's are much less uninformative than the priors used under sets 1 and 2. Of course, many more choices of priors are possible. Both the interactions between time and treatment arm were significant, under prior set 1, indicating that both the early integrated and late integrated treatment arms had higher increase in mean CD4+ count than the sequential treatment arm. Under prior sets 2 and 3 only the interaction between time and the late integrated treatment arm compared to the sequential treatment arm was significant, although the posterior mean for the interaction between time and the early integrated treatment arm compared to the sequential treatment arm was similar to the posterior mean when prior set 1 was used.

The results of the Bayesian analysis were similar to the results of the likelihood-based analysis (Table 5.7). In both models we conclude that the interaction between time and the late integrated treatment arm compared to the sequential treatment arm is significant. In the Bayesian model, we conclude that the interaction between time and the early integrated treatment arm compared to the sequential treatment arm was statistically significant. This was not significant in the likelihood-based model, although the p-value was small (0.08).

The addition of a prior distribution to the model means that the validity of the Bayesian answer depends both on the validity of the substantive model as well as the validity of the prior model. By choosing non-informative priors, we hope that the results are not sensitive to the choice of the

prior. Differences between the likelihood-based analysis and the Bayesian analysis can be attributed to the addition of a prior distribution.

### 5.4.4 Inverse probability weighting

Inverse probability weighting requires that the data set follows a monotone missing pattern. We imputed the few intermittent missing values in order to create a monotone missing data set. The data set with a monotone missing pattern was formatted that the last observation for participants with incomplete data was an entry with the missingness indicator equal to 0. Logistic regression was used to model the probability of having a missing observation at each visit. The model included CD4+ counts at all the previous visits. The following SAS code was used to model the probability of missingness at the last time point. Similar code was used for the earlier time points.

```
proc logistic data=dataset;
  class mis treatment;
  model mis = sqr_cd4at0  sqr_cd4at6  sqr_cd4at12  sqr_cd4at18  treatment;
  where point = 5;
output out=predict P=probs;
run;
```

More covariates can also be added to the missingness model to improve the prediction. In the second model in Table 5.14 the covariates WHO status, age, gender, history of tuberculosis, and whether the participant had extra-pulmonary tuberculosis or multidrug resistant tuberculosis were included as covariates in the model used to estimate the probability of being observed. Viral load was not included as a covariate, since some of the viral load values were missing.

Since the baseline visit was observed for all participants, the probability of being observed assigned to the baseline visit was 1. The estimated cumulative probability of being observed was the product of the probability of being observed at all the time points. The inverse probability weight is calculated as the inverse of the cumulative probability. The following code was used in SAS:

```
data wgt (keep=pid visit cumprobs probs);
  merge dataset predict;
  by pid visit;
  retain cumprobs;
  if first.pid then cumprobs=probs;
  else cumprobs=cumprobs*probs;
  if (visit=1) then ipw=1;
  else ipw=1/cumprobs;
run;
```

GEE are then applied with a working independence covariance structure using the weight statement and the GENMOD procedure.

```
proc genmod data=wgt;
weight ipw;
class pid treatment;
model cd4square = treatment visit visit*treatment pointsquare /dist=n link = identity;
where mis=1;
repeated subject=pid /type=ind;
run;
```

Two models were fitted. In one model weights were calculated using a logistic regression model including auxiliary covariates and in the other model weights were calculated using a logistic regression model not including auxiliary covariates. The estimates calculated using these two sets of weights were similar. The conclusion is that the additional covariates do not improve the accuracy of the weights calculated.

The standard errors calculated using procedure GENMOD are in principle not correct because they do not take the estimation of weights into account. Correct standard errors were calculated by drawing bootstrap samples. One hundred bootstrap samples of the same size as the original data set were drawn with replacement. The results using the bootstrap samples, both with and without additional covariates in the calculation of the weights, are also included in Table 5.14. The incorrect standard errors calculated using procedure GENMOD and the standard errors calculated using bootstrap are similar and the conclusions drawn are similar. This underscores that the effect of neglecting uncertainty in the weights may have limited consequences only.

Using weighted GEE, the estimates for the comparison of the effect of the early integrated treatment arm compared to the sequential treatment arm over time as well the effect of the late integrated treatment arm compared to the sequential treatment arm over time are both negative. This means that this model estimates that CD4+ count increases less in the two integrated treatment arms compared to the sequential treatment arm. This is different from any of the other models fitted under MAR. Neither of these effects was statistically significant. In addition to employing inverse probability weighting, this model differs from the other models fitted because it was fitted using GEE as estimation method, while the other methods were fitted using maximum likelihood (Table 5.14).

These models are inefficient, with inconsistent parameter estimates. If some weights are large, some observations could be given undue importance in the estimation of parameters. The estimation could be improved by using doubly robust estimators.

The last model included in Table 5.14 was fitted with doubly robust methods, using bootstrap samples to calculate correct confidence intervals, by following three steps. In step 1 a logistic regression was fitted to estimate the probability of being observed as a function of previous observations of the outcome variable and covariates. In step 2 a mixed model was fitted to the response for the completely observed observations, weighted using the weights of the inverse probability of being observed at each visit as calculated in the previous step. As output from this model, $m_i^*$, the fitted values of the response for all participants, those who were observed and those who had missing values, were saved to a data set. In step 3, the values of the response $Y_i$ were replaced with the fitted values, $m_i^*$ and another mixed model was fitted for all observations (both observed and those with missing values) to the new response, $m_i^*$ including only the covariates of interest (treatment arm) and not the covariates predictive of missingness.

The Vansteelandt macro, available from http://missingdata.lshtm.ac.uk/index.php?option=com_content&view=article&id=225:vansteelandts-doubly-robust-estimation-method&catid=60:inclusive-modeling-approaches&Itemid=137, was used to draw the bootstrap samples and prepare the data sets, but the models were fitted using the following code in SAS.

**Step 2:**

```
proc mixed data= indata ;
class pointx pid ;
model cd4square= point dummyearly dummylate  point*dummyearly  point*dummylate
pointsquare              /solution outp=first_glm_out;
repeated pointx/ subject = pid type = ar(1);
by replicate;
weight prob;
run;
```

In the SAS code prob is the inverse of the probability of being observed that was calculated in step 1 using the Vansteelandt macro. This was done both including and excluding additional covariates predictive of missingness. Replicate refers to the replicate number of the bootstrap sample. In other models fitted throughout this chapter the unstructured covariance matrix was used. These models did not converge if the unstructured covariance structure was used. It did converge with the autoregressive covariance matrix assumed. This covariance matrix is applicable because the visits were scheduled at constant intervals of 6 months.

**Table 5.14: CD4+ count (square root) analysed with inverse probability weighting methods**

| Effect | Weighted GEE without covariates | | | Weighted GEE with covariates | | |
|---|---|---|---|---|---|---|
| | Estimate | Standard error | p-value | Estimate | Standard error | p-value |
| Intercept | 7.18 | 0.47 | | 7.13 | 0.47 | |
| Early integrated treatment arm (ref: sequential treatment arm) | 2.41 | 0.57 | <0.001 | 2.39 | 0.58 | <0.001 |
| Late integrated treatment arm (ref: sequential treatment arm) | 1.88 | 0.57 | 0.001 | 1.89 | 0.57 | 0.001 |
| Time | 3.73 | 0.27 | <0.001 | 3.77 | 0.28 | <0.001 |
| Time square | -0.29 | 0.04 | <0.001 | -0.30 | 0.04 | <0.001 |
| Interaction between treatment arm and time (early integrated compared to sequential treatment arm) | -0.09 | 0.17 | 0.61 | -0.06 | 0.17 | 0.71 |
| Interaction between treatment arm and time (late integrated compared to sequential treatment arm) | -0.22 | 0.18 | 0.20 | -0.19 | 0.18 | 0.28 |
| | **Weighted GEE without covariates using bootstrap** | | | **Weighted GEE with covariates using bootstrap** | | |
| | Estimate | Standard error | p-value | Estimate | Standard error | p-value |
| Intercept | 7.15 | 0.43 | | 7.12 | 0.45 | |
| Early integrated treatment arm (ref: sequential treatment arm) | 2.44 | 0.59 | <0.001 | 2.43 | 0.59 | <0.001 |
| Late integrated treatment arm (ref: sequential treatment arm) | 1.93 | 0.58 | <0.001 | 1.91 | 0.60 | 0.002 |
| Time | 3.72 | 0.26 | <0.001 | 3.74 | 0.28 | <0.001 |
| Time square | -0.29 | 0.04 | <0.001 | -0.30 | 0.04 | <0.001 |
| Interaction between treatment arm and time (early integrated compared to sequential treatment arm) | -0.07 | 0.18 | 0.69 | -0.03 | 0.19 | 0.89 |
| Interaction between treatment arm and time (late integrated compared to sequential treatment arm) | -0.23 | 0.18 | 0.19 | -0.16 | 0.18 | 0.38 |
| | **Doubly robust weighting method without covariates** | | | **Doubly robust weighting method with covariates** | | |
| | Estimate | Standard error | p-value | Estimate | Standard error | p-value |
| Intercept | 9.52 | 0.64 | | 10.52 | 0.56 | |
| Early integrated treatment arm (ref: sequential treatment arm) | 5.15 | 1.03 | <0.001 | 4.41 | 0.81 | <0.001 |
| Late integrated treatment arm (ref: sequential treatment arm) | 3.23 | 0.85 | <0.001 | 2.96 | 0.78 | <0.001 |
| Time | 3.05 | 0.28 | <0.001 | 3.37 | 0.43 | <0.001 |
| Time square | -0.20 | 0.05 | <0.001 | -0.27 | 0.06 | <0.001 |
| Interaction between treatment arm and time (early integrated compared to sequential treatment arm) | -0.95 | 0.31 | 0.003 | -0.94 | 0.30 | 0.002 |
| Interaction between treatment arm and time (late integrated compared to sequential treatment arm) | -0.60 | 0.25 | 0.02 | -0.61 | 0.28 | 0.03 |

**Step 3:**

```
proc mixed data= dataset ;
class pointx pid ;
model pred= point dummyearly dummylate point*dummyearly point*dummylate pointsquare
        /solution outp=abc3;
    repeated pointx/ subject = pid type = ar(1);
    by replicate;
ds output ConvergenceStatus=MixedStatus  solutionf = abmixparms ;
run;
```

The parameter estimates are saved to abmixparms for all the bootstrap samples and these are then averaged to calculate the bootstrap average and standard error. The number of participants who had missing data was higher in the sequential treatment arm than in the two integrated treatment arms. This implies that the observed participants in the sequential arm will have higher weights. The mean of the weights in the early integrated treatment arm was 1.33, in the later integrated treatment arm was 1.45 and in the sequential treatment arm was 1.54. If the participants who had better outcomes were more likely to be observed, these participants with better outcomes will be weighted more heavily in the sequential treatment arm than in the other two treatment arms. Table 5.15 shows that the sequential treatment arm had a larger proportion of participants with large weights than the other treatment arms. In addition, although the mean CD4+ count was lower in the sequential treatment arm, the mean CD4+ count was higher for participants with high weights than for participants with low weights. This was true in all three treatment arms. This means that the inverse probability weighted analysis, where participants with higher weights are more influential, increased the mean CD4+ counts more in the sequential treatment arm than in the other two treatment arms.

**Table 5.15: Mean square root of CD4+ count by treatment arm and probability of being observed**

| | **Early integrated treatment arm** | | **Late integrated treatment arm** | | **Sequential treatment arm** | |
|---|---|---|---|---|---|---|
| | **Mean CD4+ count** | **% of weights** | **Mean CD4+ count** | **% of weights** | **Mean CD4+ count** | **% of weights** |
| Weight > 1.4 | 18.98 | 20.2% | 18.09 | 42.6% | 16.77 | 49.7% |
| Weight ≤ 1.4 | 18.10 | 79.8% | 15.61 | 57.4% | 12.32 | 50.4% |

The doubly robust analysis provides some protection against some of the disadvantages of the inverse probability weighting, such as extreme weights and misspecification of the weighting model. However, it still requires either the missingness model or the imputation model to be correctly specified, as well as the substantive model. If neither of these models is correctly specified, this method does not provide valid results.

### 5.4.5 Conclusions of MAR analysis

Under the MAR framework, longitudinal models were fitted using likelihood-based methods, multiple imputation, Bayesian methods and inverse probability weighting. Although pattern-mixture models are generally regarded as MNAR methods, MAR assumptions can be expressed in the pattern-mixture framework through identifying restrictions related to the available case missing value (ACMV) as discussed in Section 3.3.2 (Molenberghs et al., 1998). This is given in Section 5.5.1.3 where pattern-mixture models are discussed.

**Table 5.16: Summary of MAR sensitivity analyses**

| | **Interaction between treatment arm and time** | | | |
|---|---|---|---|---|
| | **Early integrated treatment arm compared to sequential treatment arm** | | **Late integrated treatment arm compared to sequential treatment arm** | |
| | **Estimate** | **p-value** | **Estimate** | **p-value** |
| Likelihood-based methods | 0.25 | 0.08 | 0.29 | 0.04 |
| Multiple imputation (excluding additional covariates in imputer's model) | 0.22 | 0.14 | 0.20 | 0.17 |
| Bayesian analysis (Prior set 1) | 0.28 | Significant | 0.37 | Significant |
| Inverse probability weighting: weighted GEE without covariates and using bootstrap for standard errors | -0.07 | 0.69 | -0.23 | 0.19 |

Most of the estimated coefficients for the interaction between time and the early integrated treatment arm compared to the sequential treatment arm were between 0.22 and 0.28. The exception was the results of the inverse probability weighting. The interaction between time and the early integrated treatment arm compared to the sequential treatment arm was only significant in the Bayesian model (Table 5.16). Excluding the inverse probability weighting, the interaction between time and the late integrated treatment arm compared to the sequential treatment arm was between 0.20 and 0.37. The estimate was only significant in the likelihood-based model and in the Bayesian model. We therefore conclude that there was some evidence of a higher mean increase over time in CD4+ count in the early and late integrated treatment arms compared to the sequential treatment arm.

If the MAR assumption holds, the missingness mechanism is ignorable and future statistical behaviour of observations from a participant, conditional on the history, is the same for participants who have missing data and participants who are observed, and participants who have missing data continued to adhere to treatment, then these analyses give Estimand 1. Under those conditions these results are the ITT estimand, including all data at the planned end of the study on all participants analysed as randomised. If we want to interpret this estimand as the ITT estimate of the effectiveness of a treatment policy, we need to make the assumption that participants had the same adherence to study treatment after dropout from the study as during follow-up. However, the

effects of the drugs wane after discontinuation and the assumption that participants continued to adhere after drop out is not very plausible; except for a few participants who relocated and continued to take treatment at a different clinic. Some participants had not adhered to treatment during follow-up and for those participants it is reasonable to expect that their adherence after dropout was the same as during follow-up.

Estimand 1 is estimated by all of the MAR analyses reported. Data at the end of the planned study period were available for all randomised participants through a different mechanism in each of these methods. In the likelihood-based analysis, missing data are filled in following the slope estimated for each participant using the observed data. The analysis using multiple imputation created a complete data set by imputing the missing data, and could therefore be valid under the requirement for an ITT analysis, provided that future statistical behaviour can be accurately imputed given past observations, and the imputer's model is correct. In the Bayesian analysis, the missing data are filled in through sampling from the conditional posterior distribution via the Gibbs sampler. The model fitted using inverse probability weighting weights the observed observations and gives observations with a low probability of being observed at the end of the study more weight in order to adjust the final estimate for the unobserved observations. In this way the estimate represents all participants at the end of the study.

Mortality was high in the SAPiT study. Death as an endpoint has been analysed separately using standard survival analysis techniques (Abdool Karim et al., 2010; Abdool Karim et al., 2011) and it has been shown that the sequential treatment arm had higher mortality than the two integrated treatment arms. There is therefore an imbalance in death between the treatment arms.

It is thus important to consider death as a cause of missing data. One could argue that values post death are not missing in the strictest sense. In our analysis of CD4+ count death has just been treated as another cause of missing data. The methods that are valid under MAR with ignorable missing data, such as likelihood-based models, impute CD4+ count after dropout implicitly, because of the structure imposed by correlation between each participant's longitudinal observations. This implies that we are drawing inferences on CD4+ count in an immortal cohort. We are not modelling the association between CD4+ count and treatment in the living only, but also include imputed data after death.

Dufouil et al. (2004) suggested an adjustment to inverse probability weighting that allows one to treat missing data due to dropout and death differently. This has not been applied in our analysis, but could certainly be done.

## 5.5 Analysis under MNAR assumptions

A sensitivity analysis is performed by doing several analyses that are valid under MNAR assumptions. MNAR assumptions are described in Section 2.2.3. The data are analysed using pattern-mixture models (discussed in Section 3.3.2), selection models (discussed in Section 3.3.1) using Bayesian models (discussed in Section 3.6) and shared-parameter models (discussed in Sections 3.3.3 and 3.3.4).

### 5.5.1 Pattern-mixture models

Various applications of pattern-mixture models have been discussed in the literature. We analyse the SAPiT data using four of these applications. The first uses mixed models as discussed by Hedeker and Gibbons (1997). The second uses multiple imputation (Carpenter & Kenward, 2007). The third uses Bayesian methodology as discussed by White et al. (2007). The fourth uses identifying restrictions and multiple imputation (Curran et al., 2004; Thijs et al., 2002). This is by no means an exhaustive list of all available pattern-mixture applications.

Patterns of missing data are identified as in Table 5.2. The first challenge using pattern-mixture models is to determine how the various dropout patterns are grouped. Three different dropout variables were defined; each indicating more detailed dropout groups. In the first model dropout was a binary variable created to indicate whether a participant completed the study or not. Dropout was assigned a value of zero if the study was completed and a value of one if the study was not completed (Model 1 in Table 5.21). In addition, a second dropout variable was created with three categories, an indicator whether the study was completed, whether the study was not completed or whether the participant died (Model 2). The third dropout variable has, in addition to the categories in Model 2, a category for people who had information at baseline only and dropped out after that (Model 3). The pattern where participants had baseline data only was chosen because this was the dropout pattern with the most participants. Table 5.17 gives the number of participants in each of these categories. More dropout patterns can be identified from the data, but some of these patterns include a small number of participants.

In the early integrated treatment arm 69 (32.2%) participants dropped out, in the late integrated treatment arm 84 (39.1%) participants dropped out and in the sequential treatment arm 90 (42.3%) participants dropped out. In Model 2 we have a category for participants who dropped out due to death, but other than that we treat all dropouts the same and do not incorporate the reason for dropout in the analysis. This is not necessarily the optimal strategy, since participants have different reasons for dropping out and these reasons for dropping out, rather than the timing of dropout, may be related to outcomes in different ways.

**Table 5.17: Number of participants in each of the categories of missing data**

| | Early integrated treatment arm N = 214 | Late integrated treatment arm N = 215 | Sequential treatment arm N = 213 |
|---|---|---|---|
| **Model 1: Binary dropout variable** | | | |
| Completed | 145 (67.8%) | 131 (60.9%) | 123 (57.8%) |
| Dropped out | 69 (32.2%) | 84 (39.1%) | 90 (42.3%) |
| **Model 2: Dropout and death included** | | | |
| Completed | 145 (67.8%) | 131 (60.9%) | 123 (57.8%) |
| Dropped out | 52 (24.3%) | 67 (31.2%) | 55 (25.8%) |
| Died | 17 (7.9%) | 17 (7.9%) | 35 (16.4%) |
| **Model 3: Dropout, death and baseline data** | | | |
| Completed | 145 (67.8%) | 131 (60.9%) | 123 (57.8%) |
| Dropped out | 29 (13.6%) | 29 (13.6%) | 26 (12.2%) |
| Baseline data only | 23 (10.8%) | 38 (17.7%) | 29 (13.6%) |
| Died | 17 (7.9%) | 17 (7.9%) | 35 (16.4%) |

**Table 5.18: Estimates and standard errors of coefficients from the different patterns of missing data**

| | Dropout | | Dead | | Completed | |
|---|---|---|---|---|---|---|
| **Effect** | **Estimate** | **Standard error** | **Estimate** | **Standard error** | **Estimate** | **Standard error** |
| Intercept | 8.89 | 1.10 | 7.00 | 1.76 | 8.32 | 0.47 |
| Early integrated treatment arm (ref: sequential treatment arm) | 3.05 | 1.07 | -0.66 | 2.11 | 0.56 | 0.65 |
| Late integrated treatment arm (ref: sequential treatment arm) | 3.51 | 0.84 | -1.61 | 2.02 | -0.41 | 0.62 |
| Time | 2.08 | 1.03 | 2.54 | 1.90 | 3.52 | 0.23 |
| Time squared | -0.05 | 0.23 | -1.02 | 0.52 | -0.31 | 0.03 |
| Interaction between treatment arm and time (early integrated compared to sequential treatment arm) | -0.06 | 0.66 | 1.55 | 1.34 | 0.21 | 0.14 |
| Interaction between treatment arm and time (late integrated compared to sequential treatment arm) | 0.15 | 0.58 | 1.71 | 1.15 | 0.26 | 0.14 |

Table 5.18 gives the estimates for each of the patterns of missing data. MAR models assume the slope of the CD4+ count over time is the same across patterns; however the coefficient for the interaction between treatment arm and time differs substantially between the three patterns of missing data, meaning that the MAR assumption might not be supported by the data.

#### 5.5.1.1 Pattern-mixture models using random-effects mixed models

Procedure MIXED in SAS was used to fit the model. Square root CD4+ count was the dependent variable. Independent variables included the fixed, categorical effects of treatment, dropout, and the dropout by treatment interaction, as well as the continuous effect of time and time squared. The time by treatment, dropout by time and dropout by treatment interaction and the three level interaction of dropout, treatment and time were included. Adding a term for time squared improved the fit of the model since it was not a linear pattern. An unstructured covariance matrix

was used to model the within participant errors, which was modelled separately for each treatment arm. Parameters were estimated using REML with the Newton–Raphson algorithm. Denominator degrees of freedom was estimated using the Kenward-Roger correction (Kenward & Roger, 1997).

The following SAS code was used:

```
proc mixed method = reml ;
  class pointx treatment pid dropout;
  model cd4square = treatment  point  pointsquared  point*treatment  dropout  dropout*point
dropout*treatment  dropout*treatment*point / solution residual ddfm = kr outp = cd4mixedpred;
  repeated pointx / subject = pid type = un group = treatment;
run;
```

**The regression equation for the pattern-mixture model with binary dropout (Model 1)**

$$Y_{ij} = \beta_0 + \beta_1 E_i + \beta_2 L_i + \beta_3 t_{ij} + (\beta_{41} E_i + \beta_{42} L_i) t_{ij} + \beta_5 D_i + \beta_6 t_{ij} D_i + \beta_7 t_{ij}^2 + \beta_{81} D_i E_i + \beta_{82} D_i L_i + (\beta_{91} E_i + \beta_{92} L_i) D_i t_{ij} + \varepsilon_{ij}$$

**Regression equation for pattern-mixture model with three dropout categories (dropout, dead, completed, Model 2)**

$$Y_{ij} = \beta_0 + \beta_1 E_i + \beta_2 L_i + \beta_3 t_{ij} + (\beta_{41} E_i + \beta_{42} L_i) t_{ij} + \beta_{51} C_i + \beta_{52} F_i + (\beta_{61} C_i + \beta_{62} F_i) t_{ij} + \beta_7 t_{ij}^2 + \beta_{811} E_i C_i + \beta_{812} E_i F_i + \beta_{821} L_i C_i + \beta_{822} L_i F_i + (\beta_{911} E_i C_i + \beta_{912} E_i F_i + \beta_{921} L_i C_i + \beta_{922} L_i F_i) t_{ij} + \varepsilon_{ij}$$

Where $C_i = 1$ when completed, 0 otherwise and $F_i = 1$ when died, 0 otherwise

**Regression equation for pattern-mixture model with four dropout categories (dropout, dead, completed, baseline data only, Model 3)**

$$Y_{ij} = \beta_0 + \beta_1 E_i + \beta_2 L_i + \beta_3 t_{ij} + (\beta_{41} E_i + \beta_{42} L_i) t_{ij} + \beta_{51} C_i + \beta_{52} F_i + \beta_{53} B_i + (\beta_{61} C_i + \beta_{62} F_i + \beta_{63} B_i) t_{ij} + \beta_7 t_{ij}^2 + \beta_{811} E_i C_i + \beta_{812} E_i F_i + \beta_{813} E_i B_i + \beta_{821} L_i C_i + \beta_{822} L_i F_i + \beta_{823} L_i B_i + (\beta_{911} E_i C_i + \beta_{912} E_i F_i + \beta_{912} E_i B_i + \beta_{921} L_i C_i + \beta_{922} L_i F_i + \beta_{923} L_i B_i) t_{ij} + \varepsilon_{ij}$$

Where $C_i = 1$ when completed, 0 otherwise; $F_i = 1$ when died, 0 otherwise and $B_i = 1$ when there is baseline data only, 0 otherwise

**Table 5.19: F-test of fixed effects for models including all two way interactions and a three way interaction between treatment arm, dropout and time adjusted for all other variables and interactions in the model**

| | Degrees of freedom | | F-value | p-value |
|---|---|---|---|---|
| **Effect** | **Numerator** | **Denominator** | | |
| **Model 1: Binary dropout variable** | | | | |
| Treatment | 2 | 534 | 0.93 | 0.40 |
| Time | 1 | 537 | 237.20 | <0.001 |
| Time squared | 1 | 432 | 73.95 | <0.001 |
| Dropout | 1 | 739 | 7.69 | 0.006 |
| Interaction time and treatment | 2 | 470 | 3.32 | 0.04 |
| Interaction between time and dropout | 1 | 629 | 14.93 | <0.001 |
| Interaction between treatment and dropout | 2 | 534 | 0.80 | 0.45 |
| Interaction between time, treatment and dropout | 2 | 466 | 1.26 | 0.28 |
| **Model 2: Dropout and death included** | | | | |
| Treatment | 2 | 595 | 0.69 | 0.50 |
| Time | 1 | 721 | 98.12 | <0.001 |
| Time squared | 1 | 437 | 74.97 | <0.001 |
| Dropout | 2 | 798 | 9.22 | <0.001 |
| Interaction time and treatment | 2 | 435 | 2.17 | 0.12 |
| Interaction between time and dropout | 2 | 639 | 22.22 | <0.001 |
| Interaction between treatment and dropout | 4 | 638 | 0.28 | 0.89 |
| Interaction between time, treatment and dropout | 4 | 512 | 0.56 | 0.69 |
| **Model 3: Dropout and death included, with baseline data only** | | | | |
| Treatment | 2 | 568 | 0.76 | 0.47 |
| Time | 1 | 720 | 99.05 | <0.001 |
| Time square | 1 | 437 | 75.30 | <0.001 |
| Dropout | 2 | 766 | 3.92 | 0.02 |
| Interaction time and treatment | 2 | 453 | 1.66 | 0.19 |
| Interaction between time and dropout | 2 | 638 | 21.97 | <0.001 |
| Interaction between treatment and dropout | 4 | 599 | 0.26 | 0.91 |
| Interaction between time, treatment and dropout | 4 | 509 | 0.55 | 0.70 |

The dummy coded variables for dropout were entered into a longitudinal mixed model as a main effect and as interactions with the variables treatment and time (Table 5.19). In all the models the three way interaction and the interaction between treatment and dropout was not significant. The non-significant three way interaction indicated that the treatment arm by time interaction (which indicates a more dramatic improvement over time for participants in the integrated treatment arms compared to the sequential treatment arm) was not more pronounced for any one dropout category (completers, dropouts or participants who died). Although the interaction between time and treatment was not significant in Models 2 and 3, it was kept in the final model fitted, because that was the comparison that was of interest. The final model fitted excluded the non-significant three-way interaction and excluded the non-significant interaction between treatment and dropout (Table 5.20).

**Table 5.20: CD4+ count over time (square root transformed). Final model fitted, with non-significant interactions removed**

| | Model 1 Binary dropout variable | | | Model 2 Dropout and death | | | Model 3 Dropout, death and baseline only | | |
|---|---|---|---|---|---|---|---|---|---|
| **Effect** | **Estimate** | **SE** | **p-value** | **Estimate** | **SE** | **p-value** | **Estimate** | **SE** | **p-value** |
| Intercept | 8.56 | 0.51 | | 10.08 | 0.70 | | 9.34 | 0.97 | |
| $\beta_1$ Early integrated treatment (ref: sequential treatment) | 0.46 | 0.67 | 0.50 | 0.64 | 0.95 | 0.50 | 1.12 | 1.28 | 0.38 |
| $\beta_2$ Late integrated treatment (ref: sequential treatment) | -0.58 | 0.65 | 0.37 | 1.18 | 0.84 | 0.16 | 1.69 | 1.21 | 0.17 |
| $\beta_3$ Time | 3.39 | 0.22 | <0.001 | 3.02 | 0.27 | <0.001 | 3.06 | 0.27 | <0.001 |
| $\beta_7$ Time squared | -0.30 | 0.03 | <0.001 | -0.30 | 0.03 | <0.001 | -0.30 | 0.03 | <0.001 |
| $\beta_{41}$ Interaction treatment arm, time (early integrated compared to sequential treatment arm) | 0.27 | 0.14 | 0.05 | 0.25 | 0.14 | 0.07 | 0.25 | 0.14 | 0.07 |
| $\beta_{42}$ Interaction treatment arm, time (late integrated compared to sequential treatment arm) | 0.35 | 0.14 | 0.01 | 0.34 | 0.14 | 0.01 | 0.34 | 0.14 | 0.01 |
| $\beta_5$ Dropout | 0.19 | 0.73 | 0.79 | - | - | - | - | - | - |
| $\beta_{51}$ Completed compared to dropout | - | - | - | -1.61 | 0.81 | 0.05 | -0.90 | 1.05 | 0.39 |
| $\beta_{52}$ Dead compared to dropout | - | - | - | -1.57 | 1.18 | 0.18 | -0.83 | 1.36 | 0.54 |
| $\beta_{53}$ Baseline data only compared to dropout | - | - | - | - | - | - | 1.34 | 1.27 | 0.29 |
| $\beta_6$ Interaction time and dropout | -0.75 | 0.21 | <0.001 | - | - | - | - | - | - |
| $\beta_{61}$ Interaction time, dropout (Completed compared to dropout) | - | - | - | 0.37 | 0.22 | 0.09 | 0.34 | 0.23 | 0.13 |
| $\beta_{62}$ Interaction time, dropout (Dead compared to dropout) | - | - | - | -2.70 | 0.52 | <0.001 | -2.74 | 0.52 | <0.001 |
| Interaction treatment arm and dropout | | | | | | | | | |
| $\beta_{811}$ Interaction early integrated treatment arm and dropout (Completed compared to drop out; early compared to sequential) | 0.84 | 1.00 | 0.40 | -0.10 | 1.10 | 0.92 | -0.54 | 1.39 | 0.70 |
| $\beta_{812}$ Interaction early integrated treatment arm and dropout (Dead compared to drop out; early compared to sequential) | - | - | - | -0.22 | 1.72 | 0.90 | -0.69 | 1.92 | 0.72 |
| $\beta_{813}$ Interaction early integrated treatment arm and dropout (Baseline data only compared to dropout; early compared to sequential) | - | - | - | - | - | - | -0.80 | 1.87 | 0.67 |
| $\beta_{821}$ Interaction late integrated treatment arm and dropout (Completed compared to drop out; late compared to sequential) | 2.37 | 0.93 | 0.01 | -1.67 | 1.00 | 0.09 | -2.14 | 1.32 | 0.11 |
| $\beta_{822}$ Interaction late integrated treatment arm and dropout (Dead compared to drop out; late compared to | - | - | - | -1.14 | 1.54 | 0.46 | -1.64 | 1.77 | 0.35 |

**Table 5.20: CD4+ count over time (square root transformed). Final model fitted, with non-significant interactions removed**

| | Model 1 Binary dropout variable | | | Model 2 Dropout and death | | | Model 3 Dropout, death and baseline only | | |
|---|---|---|---|---|---|---|---|---|---|
| **Effect** | **Estimate** | **SE** | **p-value** | **Estimate** | **SE** | **p-value** | **Estimate** | **SE** | **p-value** |
| sequential) | | | | | | | | | |
| $\beta_{823}$ Interaction late integrated treatment arm and dropout (Baseline compared to drop out; late compared to sequential) | - | - | - | - | - | - | -0.99 | 1.64 | 0.55 |

SE: Standard error

The degree to which the missing data patterns moderated the influence of other model terms was investigated through interactions with the missing data patterns. In all three models the interaction between dropout and time was statistically significant and the interaction between time and treatment had a p-value of 0.06, which is close to statistical significance. The extra category of baseline data only in Model 3 did not seem to add much value. Therefore Model 2 was chosen as the most parsimonious model to represent the data (Table 5.21).

**Table 5.21: F-test of fixed effects for final models with non-significant interactions removed, adjusted for all other variables and interactions in the model**

| | Degrees of freedom | | F-value | p-value |
|---|---|---|---|---|
| **Effect** | **Numerator** | **Denominator** | | |
| **Model 1: Binary dropout variable** | | | | |
| Treatment | 2 | 419 | 1.51 | 0.22 |
| Time | 1 | 441 | 286.01 | <0.001 |
| Time squared | 1 | 433 | 74.09 | <0.001 |
| Dropout | 1 | 739 | 7.81 | 0.01 |
| Interaction between time and dropout | 1 | 587 | 12.62 | <0.001 |
| Interaction between time and treatment | 2 | 295 | 2.80 | 0.06 |
| **Model 2: Dropout and death included** | | | | |
| Treatment | 2 | 423 | 0.69 | 0.50 |
| Time | 1 | 719 | 102.78 | <0.001 |
| Time squared | 1 | 439 | 75.48 | <0.001 |
| Dropout | 2 | 797 | 10.15 | <0.001 |
| Interaction between time and dropout | 2 | 638 | 21.13 | <0.001 |
| Interaction between time and treatment | 2 | 303 | 2.82 | 0.06 |
| **Model 3: Dropout, death and baseline data only** | | | | |
| Treatment | 2 | 423 | 0.77 | 0.46 |
| Time | 1 | 718 | 103.97 | <0.001 |
| Time squared | 1 | 439 | 75.83 | <0.001 |
| Dropout | 2 | 763 | 4.03 | 0.02 |
| Interaction between time and dropout | 2 | 634 | 20.86 | <0.001 |
| Interaction between time and treatment | 2 | 303 | 2.84 | 0.06 |

The following assumptions were made:

- It was assumed that data are distributed normally within a pattern of missingness

- The conditional distribution of CD4+ count was assumed to depend on dropout time through the assigned category only (completed, dropped out or died) and CD4+ count was assumed to be independent of dropout time within the pattern of missingness
- Intermittent missingness was assumed to be MAR
- It was assumed that covariate effects were the same for missing and observed data within a dropout pattern

Figure 5.14 shows that the trajectory for CD4+ count over time was different for the different patterns of dropout, especially for the participants who died. The predicted mean curve fitted the observed means well. The improvement in CD4+ count over time depended on treatment and dropout status.

On the basis of the model estimates we derived the predicted mean curve for the groups in Figure 5.14.

Completers in the early integrated treatment arm: $\hat{y}_{ij} = 7.35 + 3.64\ \mathrm{t}_{ij} - 0.30\ \mathrm{t}_{ij}^2$

Completers in the late integrated treatment arm: $\hat{y}_{ij} = 7.88 + 3.74\ \mathrm{t}_{ij} - 0.30\ \mathrm{t}_{ij}^2$

Completers in the sequential treatment arm: $\hat{y}_{ij} = 8.48 + 3.40\ \mathrm{t}_{ij} - 0.30\ \mathrm{t}_{ij}^2$

Dropouts (not dead) in the early integrated treatment arm: $\hat{y}_{ij} = 11.03 + 3.27\ \mathrm{t}_{ij} - 0.30\ \mathrm{t}_{ij}^2$

Dropouts (not dead) in the late integrated treatment arm: $\hat{y}_{ij} = 11.26 + 3.36\ \mathrm{t}_{ij} - 0.30\ \mathrm{t}_{ij}^2$

Dropouts (not dead) in the sequential treatment arm: $\hat{y}_{ij} = 10.08 + 3.02\ \mathrm{t}_{ij} - 0.30\ \mathrm{t}_{ij}^2$

Participants who died in the early integrated treatment arm: $\hat{y}_{ij} = 8.94 + 0.57\ \mathrm{t}_{ij} - 0.30\ \mathrm{t}_{ij}^2$

Participants who died in the late integrated treatment arm: $\hat{y}_{ij} = 8.33 + 0.66\ \mathrm{t}_{ij} - 0.30\ \mathrm{t}_{ij}^2$

Participants who died in the sequential treatment arm: $\hat{y}_{ij} = 8.51 + 0.32\ \mathrm{t}_{ij} - 0.30\ \mathrm{t}_{ij}^2$

The predicted mean curve for completers was:

$$\hat{y}_{ij} = 8.48 + 0.79\ \mathrm{E}_i - 0.15\ \mathrm{L}_i\ + 3.40\ \mathrm{t}_{ij}\ - 0.30\ \mathrm{t}_{ij}^2$$

The predicted mean curve for dropouts who did not die was:

$$\hat{y}_{ij} = 10.08 + 0.89\ \mathrm{E}_i + 1.52\ \mathrm{L}_i + 3.02\ \mathrm{t}_{ij} - 0.30\ \mathrm{t}_{ij}^2$$

The predicted mean curve for participants who died was:

$$\hat{y}_{ij} = 8.51 + 0.67\ \mathrm{E}_i + 0.38\ \mathrm{L}_i + 0.32\ \mathrm{t}_{ij} - 0.30\ \mathrm{t}_{ij}^2$$

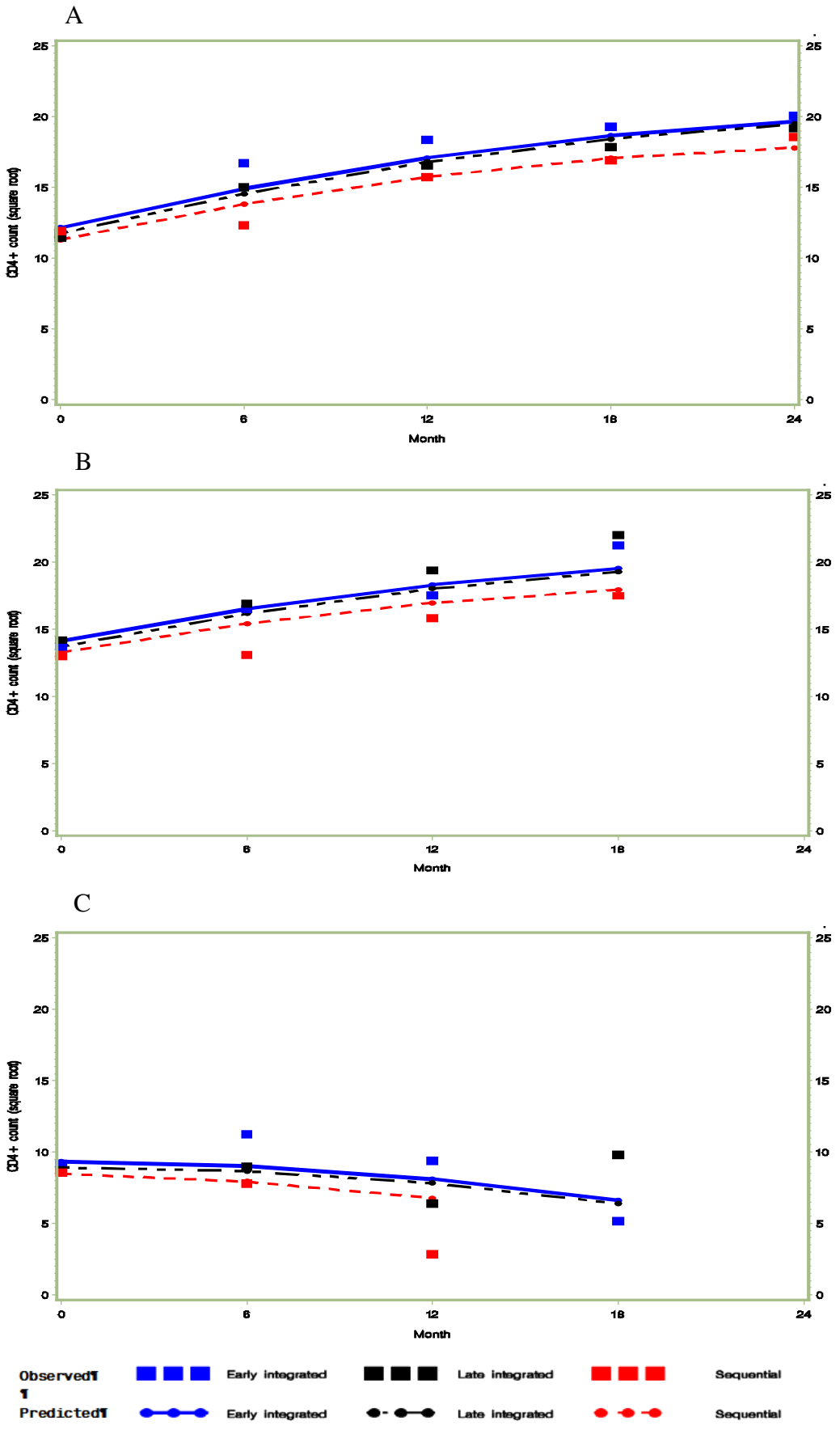



**Figure 5.14: CD4+ counts (square root) over time and treatment group for (a) completers, (b) dropouts (not dead) and (c) participants who died.**

Overall population estimates, averaging over the missing data patterns, are calculated for the fixed effects using the following formula: $\hat{\bar{\beta}} = \sum_{i=1}^{3} \hat{b}_i \hat{\pi}_i$ The sample proportion, $\hat{\pi}_i$, for completers was 0.6215, for dropouts was 0.2710 and for participants who died was 0.1075. This gave the average estimates for $\hat{\bar{\beta}}$ as given in Table 5.23. The corresponding standard errors were calculated using the methodology from Section 3.3.2.1 and are given in Table 5.22.

**Table 5.22: Population averaged estimates and standard errors of pattern mixture model using random-effects mixed model (Model 3: Dropout categories were dropout, died and completed study)**

| Effect | $\hat{\bar{\beta}}$ | Standard error | Z-statistic | Degrees of freedom | p-value |
|---|---|---|---|---|---|
| Intercept | 9.039 | 0.468 | 19.22 | 419 | <0.001 |
| Early integrated treatment arm (ref: sequential treatment arm) | 0.710 | 0.653 | 1.09 | 835 | 0.28 |
| Late integrated treatment arm (ref: sequential treatment arm) | -0.160 | 0.526 | -0.30 | 514 | 0.76 |
| Time | 2.829 | 0.370 | 7.64 | 748 | <0.001 |
| Time squared | -0.283 | 0.084 | -3.37 | 700 | <0.001 |
| Interaction between treatment arm and time (early integrated compared to sequential treatment arm) | 0.237 | 0.213 | 0.97 | 608 | 0.33 |
| Interaction between treatment arm and time (late integrated compared to sequential treatment arm) | 0.506 | 0.244 | 2.37 | 514 | 0.02 |

The assumptions and fit of the pattern-mixture model was checked by assessing how well the predicted means of CD4+ count matched the observed means for each treatment arm and dropout category and looking at residual plots to identify outliers and departures from the model as described by Hogan et al. (2004) in Figure 5.14.

This specific pattern-mixture model can be interpreted as that the increase over time in mean CD4+ count is not significantly different between the early integrated treatment arm and the sequential treatment arm. The increase over time is significantly different between the late integrated and the sequential treatment arms. Participants in the late integrated treatment arm had larger increases in CD4+ count than participants in the sequential treatment arm.

In this analysis, CD4+ counts after dropout were extrapolated. These extrapolations are done separately for each pattern of missing data, assuming the future behaviour of participants who dropped out and died could be predicted by their past behaviour. Although a similar likelihood-based mixed model was fitted, this differed from the MAR likelihood-based analysis by allowing different profiles to be fitted for participants who completed the study, participants who died and participants who dropped out. We thus made different assumptions about the future behaviour of

participants in each of these patterns. These participants had quite different profiles over time (Figure 5.14), and one would expect this model to have different results from the model under MAR.

Demirtas and Schafer (2003) criticised random coefficient pattern-mixture models for their implicit polynomial extrapolation. Their criticism was that in these models the levels of time over which extrapolations have to be made are more than the number of patterns of dropout actually seen. All responses are missing where the time variable is larger than the dropout indicator. One needs to generate predictions for cells where data are not observed. Strong assumptions then need to be made about the shape of the response surface; especially predicting to the far corner is dangerous. Even if the model is correct, the predictions may be highly variable. If the model is misspecified these predictions have large bias. Conclusions from these analyses are heavily model-dependent; while the data do not provide reassurance that the model chosen is appropriate. However, collapsing dropouts into fewer categories, as we did in this example, tends to stabilise the extrapolation for those who leave early. This greater stability should be weighed against loss of information and introduction of bias if the true pattern is gradual, rather than only pattern specific. Molenberghs and Kenward (2007) also highlighted the same concerns about these methods, namely that using the fitted profiles to predict CD4+ trajectory after dropout implies extrapolation. The results are sensitive to the extrapolations made, which may be inappropriate, especially if based on lower order polynomials. The assumptions about the dropout mechanism that these extrapolations imply are not transparent.

Demirtas and Schafer (2003) simulated several data sets and applied random coefficient pattern-mixture models to these data sets. They expressed serious concerns over these methods and found that none of the models tested performed well. They observed that the polynomial coefficient restriction models (similar in spirit to the identifying restrictions) performed better than conventional polynomial surfaces. The complete case polynomial coefficient method that gave the best coverage in their simulated studies gave a treatment effect which lay outside of the 95% interval calculated by Hedeker and Gibbons (1997) using these methods. They have shown that random coefficient pattern-mixture models can be unstable and non-robust. These methods should therefore be used with care. This strategy is computationally simpler than the identifying restriction strategy discussed later, but the identifying restrictions alleviate many of the concerns raised.

### 5.5.1.2 Pattern-mixture models using multiple imputation

White et al. (2007) and Carpenter and Kenward (2007) described how pattern-mixture models can be implemented using multiple imputation. Using the approach 50 multiple imputations were created under MAR, as discussed in Section 3.2 and implemented in Section 5.4.2. With three treatment groups, for each imputation, *k*, we sampled

$$\begin{pmatrix} d_{1k} \\ d_{2k} \\ d_{3k} \end{pmatrix} \sim N\left( \begin{pmatrix} \delta \\ \delta \\ \delta \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{pmatrix} \right)$$

for each participant, in each of the treatment arms ($l = 1, 2, 3$). For each participant we then decreased the first imputed observation by $d_{lk}$, the second by $2d_{lk}$ and so on. The resulting data sets were analysed and combined as described in Section 3.2 and implemented in Section 5.4.2. Correlations between treatment arms are set to *0*, thus $\sigma_{12}$, $\sigma_{13}$ and $\sigma_{23}$ are set to *0* and in the absence of prior information let

$$\sigma^2 = \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1.$$

Sensitivity analyses are done using different values of $\delta$, while $\delta$ is assumed to differ for participants who dropped out and those who died. $\delta$ indicates the association between low unobserved CD4+ count and dropout.

The variance increased with increasing missing data because this methodology incorporated the uncertainty introduced by missing values. The fairly large increase in variance meant that p-values at later time points were not significant. This should not be taken to mean that the effect of treatment has been removed using this methodology, but rather that uncertainty has been increased so much that conclusions are harder to draw.

In Table 5.23 $\sigma^2$ was set equal to one. If a larger variance was chosen, the values increased more and if a smaller variance was chosen, the values increased less. Different values of $\sigma$ and $\delta$ lead to different answers, but the following trends were seen: In almost all scenarios included in Table 5.23 the mean CD4+ counts were the highest in the early integrated treatment arm, followed by the late integrated treatment arm, which was higher than the sequential treatment arm. The only exceptions was Model 5 at 24 months where the late integrated arm had the highest mean. The first four models assumed that participants who died had lower mean CD4+ counts than other participants and that participants who did not complete the study had lower mean CD4+ counts than participants who completed the study. The SAS code is given in Appendix 3.

The same imputations were done in Table 5.23, Table 5.24 and Figure 5.15. Table 5.23 gives cross-sectional results; means and standard errors at each of the time points are given. The

intention is not to imply that the cross-sectional model is the same as the longitudinal model fitted in Table 5.24. Rather, the goal is to show the versatility of the method. It can be used to impute data that can then be modelled in many different modelling frameworks. The cross-sectional analysis is also useful when communicating with clinicians and other non-statisticians, who find it harder to interpret regression coefficients than means and standard errors. The implications of the different assumptions can be communicated more clearly when clinicians can see how varying $\sigma$ and $\delta$ lead to either higher or lower means or standard errors. In that regard, graphical representations, as in Figure 5.15 are also useful.

**Table 5.23: Multiple imputation of CD4+ counts (cell/mm$^3$) using a pattern-mixture approach, mean (standard error)**

| | Early integrated treatment arm N = 214 | Late integrated treatment arm N = 215 | Sequential treatment arm N = 213 | p-value* |
|---|---|---|---|---|
| **Assuming decrease in CD4+ count for participants who dropped out and died** | | | | |
| **Model 1: δ for dropout = 0.5 and δ for participants who died = 1** | | | | |
| 6 months | 303.8 (15.0) | 259.8 (13.8) | 172.1 (11.4) | <0.001 |
| 12 months | 347.5 (27.8) | 290.9 (31.1) | 252.5 (28.7) | 0.04 |
| 18 months | 368.9 (60.5) | 326.2 (65.4) | 288.4 (63.1) | 0.42 |
| 24 months | 416.3 (118.1) | 382.6 (136.0) | 365.4 (126.5) | 0.79 |
| **Model 2: δ for dropout = 1 and δ for participants who died = 2** | | | | |
| 6 months | 300.1 (14.8) | 255.8 (13.6) | 168.8 (11.2) | <0.001 |
| 12 months | 332.2 (25.4) | 274.4 (27.9) | 233.7 (25.4) | 0.02 |
| 18 months | 343.0 (47.0) | 296.8 (49.5) | 258.3 (47.1) | 0.25 |
| 24 months | 397.2 (87.0) | 355.7 (105.0) | 349.2 (99.4) | 0.72 |
| **Model 3: δ for dropout = 0.2 and δ for participants who died = 0.5** | | | | |
| 6 months | 306.0 (15.1) | 262.1 (13.9) | 174.1 (11.5) | <0.001 |
| 12 months | 357.3 (29.4) | 301.7 (33.0) | 264.8 (30.7) | 0.06 |
| 18 months | 390.6 (69.2) | 350.8 (75.5) | 316.3 (74.0) | 0.52 |
| 24 months | 451.6 (144.4) | 424.3 (163.9) | 413.1 (158.6) | 0.87 |
| **Model 4: δ for dropout = 0.3 and δ for participants who died = 1** | | | | |
| 6 months | 304.6 (15.0) | 261.0 (13.9) | 172.9 (11.5) | <0.001 |
| 12 months | 351.5 (28.6) | 296.3 (32.30 | 256.9 (29.7) | 0.05 |
| 18 months | 378.6 (65.1) | 338.6 (71.4) | 299.3 (68.5) | 0.46 |
| 24 months | 434.5 (133.1) | 404.4 (151.6) | 386.0 (143.6) | 0.82 |
| **Assuming increase in CD4+ count for participants who dropped out and decrease for those who died** | | | | |
| **Model 5: δ for dropout = -1 and δ for participants who died = 2** | | | | |
| 6 months | 307.9 (15.4) | 268.1 (14.6) | 176.4 (11.9) | <0.001 |
| 12 months | 375.8 (33.8) | 333.8 (40.6) | 282.1 (36.0) | 0.09 |
| 18 months | 458.7 (95.5) | 446.8 (112.7) | 390.0 (104.0) | 0.66 |
| 24 months | 641.8 (239.5) | 662.6 (278.2) | 630.6 (270.6) | 0.98 |
| **Model 6: δ for dropout = -0.5 and δ for participants who died = 0.5** | | | | |
| 6 months | 308.7 (15.3) | 266.5 (14.3) | 176.8 (11.8) | <0.001 |
| 12 months | 373.0 (32.3) | 323.2 (37.5) | 282.2 (34.4) | 0.10 |
| 18 months | 433.5 (85.9) | 406.7 (97.9) | 365.3 (93.2) | 0.64 |
| 24 months | 545.5 (200.7) | 543.4 (231.1) | 521.6 (222.9) | 0.94 |

* Calculated using general linear models and procedure MIANALYZE in SAS; σ = 1

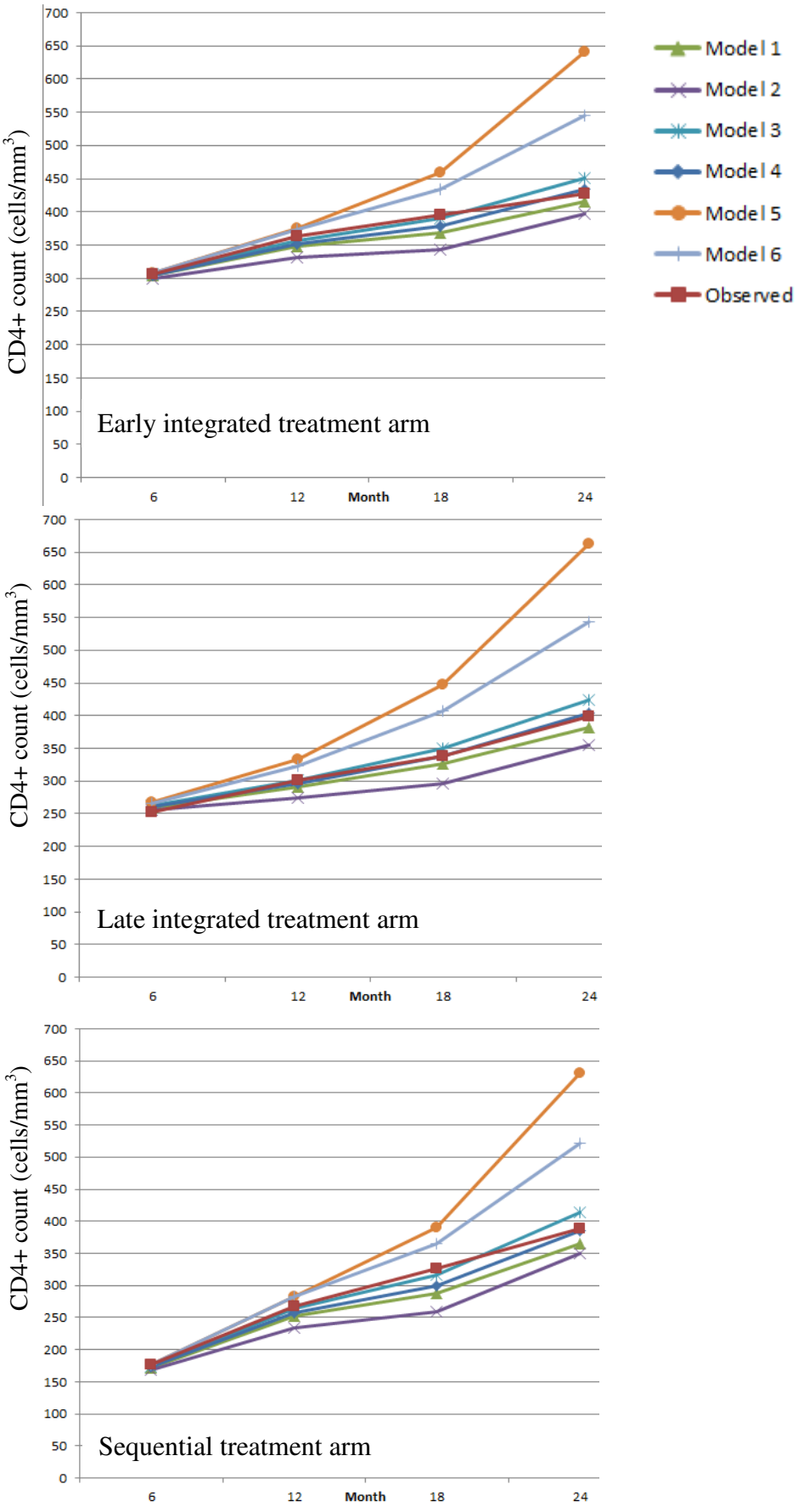



**Figure 5.15: Mean CD4+ count (cell/mm$^3$): Multiple imputation of CD4+ counts using a pattern-mixture approach, Models as in Table 5.23**

It is unrealistic to assume that CD4+ counts would be increasing over time for participants who died, therefore this assumption was not made for any of the sensitivity analyses. However, it is possible, but unlikely, that CD4+ counts could increase over time for participants who dropped out of the study (Models 5 and 6). If this assumption was made, the CD4+ counts were much higher at later time points. In addition, the difference between the early and late integrated treatment arms became smaller, whereas the sequential treatment arm had lower mean CD4+ counts than the other two arms.

In Models 1, 3, 4 and 6 the two interaction terms between treatment arm and time were not statistically significant. However, in Model 2, where a larger decrease in CD4+ count was assumed for participants who died and dropped out, the interaction between time and the early integrated arm compared to the sequential arm was significant. This is interpreted as that mean CD4+ count increased more in the early integrated treatment arm than in the sequential treatment arm. It has been shown previously that survival was different between the integrated treatment arms and the sequential treatment arm. It is possible that the larger increase in CD4+ counts observed with these models reflect the higher number of deaths in the sequential treatment arm. Model 6 that assumed an increase in CD4+ count for participants who dropped out did not have a significant treatment by time interaction, whereas Model 5 did. In Model 5 a large increase in CD4+ count was assumed for participants who dropped out, while a large decrease in CD4+ count was assumed for participants who died. The interaction between both treatment arms and time was statistically significant in Model 5, indicating that the mean increase in CD4+ count was different in the early and late integrated treatment arms compared to the sequential treatment arm.

The sensitivity analysis showed that although the actual CD4+ counts depended on the size of the mean difference assumed, the only models where the interaction between time and treatment arm was significant, were Model 2, where a large decrease in CD4+ count was assumed for participants who died and dropped out, and Model 6, which made the rather unlikely assumption that CD4+ count increased after dropout.

This method fitted the same substantive model as under the MAR multiple imputation in Section 5.4.2. The difference is that the data were changed after imputation according to the pattern assumed in the different drop out groups. If the assumed patterns were correct, these results should provide more accurate estimates of the parameters of interests. However, if these assumptions were incorrect, these models could provide less accurate estimates.

These models lead to curves that follow a different pattern from the curves fitted previously. The reason is that for participants with monotone missing data, the assumption is that each subsequent missing value is further and further away from the imputed value. This means that when a participant dropped out the assumptions compound for the later time points, giving results that are increasingly away from the MAR model fitted.

**Table 5.24: Multiple imputation of CD4+ counts using a pattern-mixture approach, results of mixed model**

| | **Model 1** δ for dropout = 0.5, δ for participants who died = 1 | | | **Model 2** δ for dropout = 1, δ for participants who died = 2 | | | **Model 3** δ for dropout = 0.2; δ for participants who died = 0.5 | | |
|---|---|---|---|---|---|---|---|---|---|
| **Effect** | **Estimate** | **SE** | **p-value** | **Estimate** | **SE** | **p-value** | **Estimate** | **SE** | **p-value** |
| Intercept | 11.33 | 0.42 | <0.001 | 11.34 | 0.41 | | 11.36 | 0.43 | <0.001 |
| Early integrated treatment arm (ref: sequential treatment arm) | 1.00 | 0.61 | 0.11 | 0.92 | 0.60 | 0.13 | 1.02 | 063 | 0.10 |
| Late integrated treatment arm (ref: sequential treatment arm) | 0.70 | 0.58 | 0.22 | 0.72 | 0.57 | 0.21 | 0.67 | 0.59 | 0.26 |
| Time | 2.41 | 0.53 | <0.001 | 2.31 | 0.48 | <0.001 | 2.43 | 0.50 | <0.001 |
| Time squared | -0.45 | 0.16 | 0.01 | -0.66 | 0.18 | 0.0004 | -0.24 | 0.16 | 0.03 |
| Interaction between treatment arm and time (early integrated compared to sequential treatment arm) | 1.11 | 0.91 | 0.23 | 1.56 | 0.75 | 0.04 | 0.88 | 0.94 | 0.35 |
| Interaction between treatment arm and time (late integrated compared to sequential treatment arm) | 0.61 | 0.70 | 0.39 | 0.84 | 0.64 | 0.19 | 0.52 | 0.68 | 0.45 |
| | **Model 4** δ for dropout = 0.3; δ for participants who died = 1 | | | **Model 5** δ for dropout = -1; δ for participants who died = 2 | | | **Model 6** δ for dropout = -0.5; δ for participants who died = 0.5 | | |
| | **Estimate** | **SE** | **p-value** | **Estimate** | **SE** | **p-value** | **Estimate** | **SE** | **p-value** |
| Intercept | 11.31 | 0.41 | | 11.28 | 0.41 | | 11.29 | 0.41 | |
| Early integrated treatment arm (ref: sequential treatment arm) | 1.04 | 0.62 | 0.09 | 1.09 | 0.67 | 0.10 | 1.15 | 0.64 | 0.07 |
| Late integrated treatment arm (ref: sequential treatment arm) | 0.72 | 0.58 | 0.22 | 0.66 | 0.61 | 0.28 | 0.70 | 0.60 | 0.24 |
| Time | 2.38 | 0.51 | <0.001 | 1.96 | 0.36 | <0.001 | 2.29 | 0.45 | <0.001 |
| Time squared | -0.41 | 0.16 | 0.01 | -0.24 | 0.16 | 0.14 | -0.20 | 0.15 | 0.18 |
| Interaction between treatment arm and time (early integrated compared to sequential treatment arm) | 1.07 | 0.91 | 0.25 | 1.46 | 0.66 | 0.03 | 0.84 | 0.94 | 0.37 |
| Interaction between treatment arm and time (late integrated compared to sequential treatment arm) | 0.62 | 0.69 | 0.37 | 1.00 | 0.49 | 0.04 | 0.60 | 0.63 | 0.35 |

SE = standard error

#### 5.5.1.3 Pattern-mixture models using identifying restrictions

Multiple imputation was done in order to create monotone missing data, because this model required monotone missing data. A model was then fitted for each pattern. An estimate was calculated over all patterns where the required components were observed, and the conditional distributions of the unobserved outcomes given the observed outcomes were calculated. Separate likelihood-based models were fitted for each pattern of missing data using procedure MIXED in SAS.

We imputed 70 data sets using each of the identifying restrictions. Imputed data were truncated at 35 and 3 respectively to ensure that only biologically plausible values were generated. Only a small fraction of the imputed values were affected by the truncation. After the imputations under each of the restrictions, the data were analysed using the same substantive model as previously under multiple imputation and in Section 5.4.1. This mixed model was fitted for each imputation with continuous time, treatment and treatment by time interaction as fixed effects, using an unstructured covariance matrix. The separate imputations were then combined using procedure MIANALYZE. The SAS code for fitting pattern-mixture models using identifying restriction is included in Appendix 3.

The goal of fitting the pattern-mixture models was to represent a MNAR mechanism. The MAR condition can be imposed using pattern-mixture models by fitting models using the ACMV identifying restriction. This model is given here in order to contrast the MAR case with the MNAR case.

The covariates in the imputer's model differed from the covariates included in the model fitted in the multiple imputation section. In the previous model using multiple imputation, several baseline covariates were included in the imputers' model. In the pattern-mixture model using identifying restrictions only the observed outcomes are included in the imputers' model. Because both the analysis and the imputation are done by pattern of missing data, it is more elaborate than the model fitted in Section 5.4.1.

**Table 5.25: CD4+ count over time, pattern-mixture model with identifying restrictions**

| | ACMV | | | CCMV | | | NCMV | | |
|---|---|---|---|---|---|---|---|---|---|
| **Effect** | **Estimate** | **SE** | **p-value** | **Estimate** | **SE** | **p-value** | **Estimate** | **SE** | **p-value** |
| Intercept | 8.54 | 0.45 | | 8.39 | 0.44 | | 9.22 | 0.71 | |
| Early integrated treatment arm (ref: sequential treatment arm) | 1.02 | 0.53 | 0.06 | 1.25 | 0.53 | 0.02 | 1.00 | 0.62 | 0.10 |
| Late integrated treatment arm (ref: sequential treatment arm) | 0.66 | 0.53 | 0.21 | 0.80 | 0.52 | 0.13 | 0.40 | 0.60 | 0.50 |
| Time | 2.99 | 0.29 | <0.001 | 3.04 | 0.26 | <0.001 | 2.34 | 0.62 | <0.001 |
| Time squared | -0.27 | 0.05 | <0.001 | -0.27 | 0.04 | <0.001 | -0.17 | 0.10 | 0.09 |
| Interaction between treatment arm and time (early integrated compared to sequential treatment arm) | 0.23 | 0.13 | 0.09 | 0.16 | 0.12 | 0.19 | -0.11 | 0.24 | 0.65 |
| Interaction between treatment arm and time (late integrated compared to sequential treatment arm) | 0.15 | 0.13 | 0.22 | 0.11 | 0.11 | 0.35 | 0.15 | 0.25 | 0.55 |

SE = standard error; ACMV: available case missing values; CCMV: complete case missing values; NCMV: neighbouring case missing values

Both the likelihood-based method and the pattern-mixture model using ACMV identifying restrictions are valid under MAR assumptions, and the sizes of the regression coefficients are relatively similar with the two models. In the pattern-mixture model with ACMV identifying restriction the interaction between time and the late integrated treatment arm compared to the sequential treatment arm was not statistically significant, while it was statistically significant under the likelihood-based method. The interaction between time and the early integrated treatment arm compared to the sequential treatment arm was similar in the two models.

The results for the CCMV identifying restriction were fairly similar to the results for the ACMV identifying restrictions (the MAR case). The same conclusions are drawn. However, the results with the NCMV identifying restrictions were different. The NCMV model was also the only model fitted where the estimate for the interaction between time and the early integrated treatment arm compared to the sequential treatment arm was negative. This was not statistically significant.

In these models the non-identified distributions are calculated using linear combinations of the identified distributions. Comparing the CCMV and NCMV identifying restrictions models to the ACMV model indicates how these pattern-mixture models differ from the MAR model. In the ACMV case identifying restriction uses all available patterns where $f_t(y_s|y_1, \dots, y_{s-1})$ is identified by the data. With CCMV constraints the distributions for those with missing data are equated to those with complete data. Therefore participants who stayed in the study longer, who probably had better outcomes because they did not die during the course of the study, had a larger influence on results. CCMV are therefore more influenced by completers and thus more similar to the ACMV (MAR) case. NCMV borrowed information from participants with the closest dropout time. Usually there are many completers, but some of the closest neighbours the NCMV identifying

restrictions borrow information from could be sparse. These models fit the same substantive model as in previous analyses, but use information borrowed according to the specific identifying restriction under the MNAR assumption.

For all three treatment arms the ACMV and CCMV models were similar. However for the early and late integrated treatment arms, the NCMV model (the thick black line with squares as symbols for the early integrated treatment arm and the thick dotted black line with circles as symbols for the late integrated arm in Figure 5.16) was different from the other two models. It imputed a lower mean CD4+ count at each time point than the other two models. For the sequential treatment arm, the NCMV model also had lower mean CD4+ counts at all time points, but these were much closer to the other models (the black dotted line with triangles as symbols). Because there were more dropouts in the sequential arm, the nearest neighbours were more similar to the complete case or available case than in the other treatment arms.

Under the pattern-mixture model using ACMV identifying restrictions, the missing data are imputed. This model makes the additional assumption that data that follow the same missingness pattern has the same future statistical behaviour conditional on observed past data.

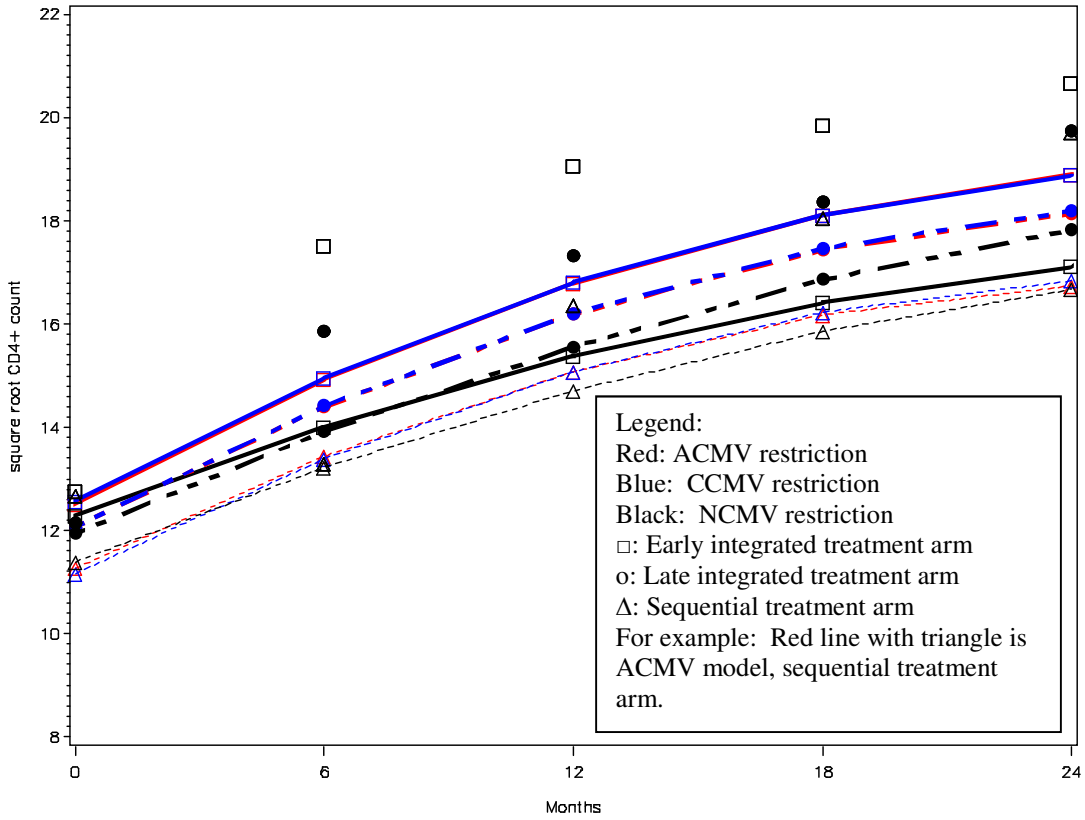



**Figure 5.16: Models fitted using identifying restrictions**

The lines indicate fitted model and the unconnected symbols indicate the observed means

### 5.5.2 Selection model approaches

Under an MNAR assumption, and using selection model factorisation, we fitted several Bayesian models under several different assumptions about the missing data mechanism. The Bayesian paradigm was chosen because it lends itself to building complex models by linking smaller sub-models into a coherent model of the joint distribution for the full data. In a non-Bayesian paradigm, integrating the joint model created through a selection model is not always possible analytically, and the integration would need to be performed numerically. Often this is non-trivial. An alternative is to do it through some form of an EM algorithm. In the Bayesian framework, if we use uninformative priors, the results will approximately agree with maximum likelihood estimates and the calculations are easier. The Gibbs sampler can be used to draw samples from the posterior distribution of the full joint distribution. This can be done in a routine way using software such as OpenBUGS.

The same model for the observed data was fitted in the Bayesian analysis under MNAR as was fitted under MAR using an unstructured covariance matrix, namely

$$\boldsymbol{Y}_i|Z_i = k \sim \text{Mutivariate Normal}(\boldsymbol{\mu}_k, \boldsymbol{\Omega}_k)$$

where $Z_i = k$ indicates that the observation belongs to treatment k

$$\mu_{ij} = \beta_0 + \beta_1 \mathrm{E}_i + \beta_2 \mathrm{L}_i + \beta_3 t_{ij} + (\beta_4 E_i + \beta_5 L_i) t_{ij} + \beta_6 t_{ij}^2$$

The priors for the parameters included in the substantive model was the same as the priors defined in Section 5.4.3 and are given in Table 5.26.

In addition, a model of missingness was added. We fitted four different models of missingness. The first model of missingness had the form

$$\mathrm{m_{ij}} \sim \text{Bernoulli}(\mathrm{p_{ij}}),$$

$$\text{logit}(\mathrm{p_{ij}}) = \theta_0 + \Delta \mathrm{y_{ij}}$$

where $m_{ij}$ was a binary missing value indicator for $Y_{ij}$. A flat prior was specified on the scale of $p_{ij}$ by specifying a logistic(0, 1) prior for $\theta_0$ in all sets. For Prior Set 1 $\Delta$ was assigned a non-informative Normal(0, 10 000) distribution. For Prior Set 2 the prior for $\Delta$ was Normal(1, 1000) and for Prior Set 3 the prior for $\Delta$ was Normal(1, 10).

The OpenBUGS code for this missingness model was:

```
mis[i,w]~dbern(p.bound[i,w])
logit(p[i,w])<-theta0+delta*cd4square[i,w]
p.bound[i,w]<-max(0, min(1,p[i,w]))
```

Since there is not just one MNAR model that is consistent with the observed data, a second MNAR model was also fitted. In the second MNAR model the model of missingness had the following form:

$$m_{ij} \sim \text{Bernoulli}(p_{ij}),$$
$$logit\ p_{ij} = \theta_1 + \theta_2 y_{i,j-1} + \theta_3 (y_{ij} - y_{i,j-1}).$$

This model allowed the missingness to depend on the previous CD4+ count and the change in CD4+ count from the previous visit to the visit where the missing data occurred. The prior sets were:

Prior set 1: $\theta_1$, $\theta_2$ and $\theta_3$ ~ logistic(0,1)

Prior set 2: $\theta_1$, $\theta_2$, $\theta_3$ ~ Normal(0, 10000)

Prior set 3: $\theta_1$, $\theta_2$, $\theta_3$ ~ Normal(0, 100)

The OpenBUGS code for this missingness model was:

```
mis[i,w]~dbern(p.bound[i,w])
logit(p[i,w])<-theta[1] + theta[2]*cd4square[i,w-1] + theta[3]*(cd4square[i,w] –  cd4square[i,w-1])
p.bound[i,w]<-max(0, min(1,p[i,w]))
```

The third MNAR model fitted allowed missingness to depend on all previous CD4+ counts, not only the CD4+ count immediately prior to drop out. The model for missingness was:

$$m_{ij} \sim \text{Bernoulli}(p_{ij}),$$

If visit = 2 then: $logit\ p_{ij} = \theta_1 + \theta_2 y_{ij}$

If visit = 3 then: $logit\ p_{ij} = \theta_1 + \theta_2 y_{i,j-1} + \theta_3 y_{ij}$

If visit = 4 then: $logit\ p_{ij} = \theta_1 + \theta_2 y_{i,j-2} + \theta_3 y_{i,j-1} + \theta_4 y_{ij}$

If visit = 5 then: $logit\ p_{ij} = \theta_1 + \theta_2 y_{i,j-3} + \theta_3 y_{i,j-2} + \theta_4 y_{i,j-1} + \theta_5 y_{ij}$

The prior sets for the missingness model were:

Prior set 1: $\theta_1$, $\theta_2$, $\theta_3$, $\theta_4$ and $\theta_5$~ logistic(0,1)

Prior set 2: $\theta_1$, $\theta_2$, $\theta_3$, $\theta_4$ and $\theta_5$~ ~ Normal(0,10000)

Prior set 3: $\theta_1$, $\theta_2$, $\theta_3$, $\theta_4$ and $\theta_5$~ ~ Normal(0, 100)

The OpenBUGS code for this missingness model, using the first set of prior distributions, was:

```
for (w in 2:2)   {  # W weeks with drop-out
for (i in 1:509) {  # exclude individuals who have already dropped out
mis[i,w]~dbern(p.bound[i,w])
logit(p[i,w])<- theta[1]+theta[2]*cd4square[i,w]
p.bound[i,w]<-max(0, min(1,p[i,w]))                    }
                    }
```

```
for (w in 3:3)   {  # W weeks with drop-out
for (i in 1:449) {  # exclude individuals who have already dropped out
mis[i,w]~dbern(p.bound[i,w])
logit(p[i,w])<-theta[1] + theta[2]*cd4square[i,w-1] + theta[3]*cd4square[i,w]
p.bound[i,w]<-max(0, min(1,p[i,w]))                    }
                    }
for (w in 4:4)   {  # W weeks with drop-out
for (i in 1:428) {  # exclude individuals who have already dropped out
mis[i,w]~dbern(p.bound[i,w])
logit(p[i,w])<-theta[1] + theta[2]*cd4square[i,w-2] + theta[3]*cd4square[i,w-1]  +
theta[4]*cd4square[i,w]
p.bound[i,w]<-max(0, min(1,p[i,w]))                    }
                    }
for (w in 5:5)   {  # W weeks with drop-out
for (i in 1:399) {  # exclude individuals who have already dropped out
mis[i,w]~dbern(p.bound[i,w])
logit(p[i,w])<-theta[1] + theta[2]*cd4square[i,w-3] + theta[3]*cd4square[i,w-2] +
theta[4]*cd4square[i,w-1] + theta[5]*cd4square[i,w]
p.bound[i,w]<-max(0, min(1,p[i,w]))                    }                    }

# Priors
for(k in 1:3){
omega[k,1:5,1:5] ~ dwish(R[k, ,],5)
sigma[k,1:5,1:5] <- inverse(omega[k, ,])  }
for(m in 1:5) {
   theta[m] ~ dlogis(0,1)         }  }
```

The fourth MNAR model fitted allowed missingness to depend on the interaction between treatment group and the unobserved CD4+ count. The model for missingness was:

$$m_{ij} \sim \text{Bernoulli}(p_{ij}),$$

$$logit\ p_{ij} = \theta_1 + \theta_2 E_i y_{ij} + \theta_3 L_i y_{ij} + \theta_4 y_{ij} + \theta_5 E_i + \theta_6 L_i$$

The prior sets for the missingness model were:

Prior set 1: $\theta_1$ to $\theta_6$, ~ logistic(0,1)

Prior set 2: $\theta_1$ to $\theta_6$, ~ ~ Normal(0,10000)

Prior set 3: $\theta_1$ to $\theta_6$, ~ ~ Normal(0, 100)

The OpenBUGS code for the missingness model was:

```
hamdmis[i,w]~dbern(p.bound[i,w])
logit(p[i,w])<-theta[1] + theta[2]*cd4square[i,w]*dummyearly[i] +
theta[3]*cd4square[i,w]*dummylate[i] + theta[4]*cd4square[i,w] + theta[5]*dummyearly[i]
+theta[6]*dummylate[i]
p.bound[i,w]<-max(0, min(1,p[i,w]))
```

Many other models for non-random missingness could be fitted, depending on the assumptions made regarding the missing data mechanism. These are not the only plausible models.

**Table 5.26: CD4+ count (square root) posterior means, standard deviations and credible intervals according to MNAR Bayesian analysis (selection models)**

**MNAR model 1** $m_{ij} \sim$ Bernoulli($p_{ij}$); $\text{logit}(p_{ij}) = \theta_0 + \Delta y_{ij}$

| | **MNAR model 1: Prior Set 1** β~Normal(0, 100 000), Ω~Wishart(I,5) I: Identity matrix, $\theta_0$~logistic(0,1), Δ~Normal(0,10 000) | | | **MNAR model 1: Prior Set 2** β~Normal(0, 1000), Ω~Wishart(A,5) A: diag (0.1, 0.1, 0.1, 0.1, 0.1 $\theta_0$~logistic(0,1), Δ~Normal(0,100) | | | **MNAR model 1: Prior Set 3** β~Normal(0, 10), Ω~Wishart(A,5) A: diag (10, 10, 10, 10, 10) $\theta_0$~logistic(0,1), Δ~Normal(0,10) | | |
|---|---|---|---|---|---|---|---|---|---|
| **Effect** | **Mean** | **SD** | **95% credible interval** | **Mean** | **SD** | **95% credible interval** | **Mean** | **SD** | **95% credible interval** |
| $\beta_0$ | 8.78 | 0.51 | 7.77;9.79 | 8.78 | 0.51 | 7.78;9.78 | 8.61 | 0.49 | 7.63; 9.57 |
| $\beta_1$ | 1.12 | 0.65 | -0.17; 2.38 | 1.11 | 0.65 | -0.16; 2.37 | 1.25 | 0.62 | 0.03; 2.47 |
| $\beta_2$ | 0.38 | 0.62 | -0.83; 1.59 | 0.36 | 0.61 | -0.84; 1.55 | 0.53 | 0.58 | 0.62; 1.67 |
| $\beta_3$ | 2.97 | 0.24 | 2.48; 3.44 | 2.97 | 0.24 | 2.49; 3.45 | 2.99 | 0.25 | 2.50; 3.46 |
| $\beta_4$ | 0.25 | 0.16 | -0.06; 0.56 | 0.25 | 0.16 | -0.06; 0.58 | 0.24 | 0.16 | -0.06; 0.56 |
| $\beta_5$ | 0.32 | 0.16 | 0.01; 0.62 | 0.32 | 0.16 | 0.02; 0.63 | 0.31 | 0.16 | 0.003; 0.61 |
| $\beta_6$ | -0.25 | 0.04 | -0.32; -0.17 | -0.25 | 0.04 | -0.32; -0.17 | -0.25 | 0.04 | -0.32; -0.17 |
| Δ | -0.06 | 0.04 | -0.13; 0.01 | -0.06 | 0.04 | -0.14; 0.01 | -0.06 | 0.04 | -0.32; -0.17 |
| $\theta_0$ | -2.76 | 0.55 | -3.84; -1,71 | -2.69 | 0.60 | -3.92; -1,58 | -2.67 | 0.61 | -3.90; -1.51 |

**MNAR model 2** $m_{ij} \sim$ Bernoulli($p_{ij}$); $logit\ p_{i,w} = \theta_1 + \theta_2 y_{i,w-1} + \theta_3(y_{i,w} - y_{i,w-1})$.

| | **MNAR model 2: Prior Set1** β~Normal(0, 100 000), Ω~Wishart(I,5) I: Identity matrix, $\theta_1, \theta_2, \theta_3$ ~logistic(0,1) | | | **MNAR model 2: Prior Set 2** β~Normal(0, 1000), Ω~Wishart(A,5) A: diag (0.1, 0.1, 0.1, 0.1, 0.1 $\theta_1, \theta_2, \theta_3$ ~Normal(0, 10000) | | | **MNAR model 2: Prior Set 3** β~Normal(0, 10), Ω~Wishart(A,5) A: diag (10, 10, 10, 10, 10) $\theta_1, \theta_2, \theta_3$ ~Normal(0, 100) | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Mean** | **SD** | **95% credible interval** | **Mean** | **SD** | **95% credible interval** | **Mean** | **SD** | **95% credible interval** |
| $\beta_0$ | 8.80 | 0.51 | 7.77; 9.78 | 8.83 | 0.50 | 7.84; 9.81 | 8.61 | 0.49 | 7.64; 9.56 |
| $\beta_1$ | 1.11 | 0.64 | -0.14; 2.38 | 1.09 | 0.65 | -0.18; 2.36 | 1.25 | 0.62 | 0.02; 2.47 |
| $\beta_2$ | 0.37 | 0.61 | -0.82;1.59 | 0.33 | 0.60 | -0.84; 1.52 | 0.52 | 0.59 | -0.63; 1.67 |
| $\beta_3$ | 2.95 | 0.24 | 2.45; 3.43 | 2.94 | 0.24 | 2.47; 3.42 | 2.99 | 0.25 | 2.50; 3.48 |
| $\beta_4$ | 0.26 | 0.16 | -0.05; 0.58 | 0.26 | 0.16 | -0.05; 0.57 | 0.24 | 0.16 | -0.06; 0.56 |
| $\beta_5$ | 0.33 | 0.16 | 0.02; 0.64 | 0.32 | 0.15 | 0.02; 0.63 | 0.31 | 0.16 | <-0.001; 0.61 |
| $\beta_6$ | -0.24 | 0.04 | -0.32; -0.17 | -0.24 | 0.04 | -0.32; -0.17 | -0.25 | 0.04 | -0.33; -0.17 |
| $\theta_1$ | -2.75 | 0.59 | -3.97; -1.66 | -2.64 | 0.99 | -4.15; -0.47 | -2.94 | 0.60 | -4.17; -1.81 |
| $\theta_2$ | -0.06 | 0.04 | -0.13; 0.02 | -0.07 | 0.06 | -0.21; 0.02 | -0.05 | 0.04 | -0.12; 0.02 |
| $\theta_3$ | -0.08 | 0.07 | -0.22; 0.05 | -0.09 | 0.08 | -0.24; 0.05 | -0.07 | 0.07 | -0.21; 0.06 |

**MNAR model 3** $m_{ij} \sim$ Bernoulli($p_{ij}$); $logit\ p_{i,w} = \theta_1 + \theta_2 y_{i,w-3} + \theta_3 y_{i,w-2} + \theta_4 y_{i,w-1} + \theta_5 y_{i,w}$

| | **MNAR model 3: Prior Set1** β~Normal(0, 100 000), Ω~Wishart(I,5) I: Identity matrix, $\theta_1$ *to* $\theta_5$,~logistic(0,1) | | | **MNAR model 3: Prior Set 2** β~Normal(0, 1000), Ω~Wishart(A,5) A: diag (0.1, 0.1, 0.1, 0.1, 0.1 $\theta_1$ to $\theta_5$ ~ Normal(0, 10000) | | | **MNAR model 3: Prior Set 3** β~Normal(0, 10), Ω~Wishart(A,5) A: diag (10, 10, 10, 10, 10) $\theta_1$ to $\theta_5$ ~Normal(0, 100) | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Mean** | **SD** | **95% credible interval** | **Mean** | **SD** | **95% credible interval** | **Mean** | **SD** | **95% credible interval** |
| $\beta_0$ | 8.77 | 0.50 | 7.78; 9.75 | 8.80 | 0.51 | 7.77; 9.77 | 8.63 | 0.49 | 7.67; 9.58 |
| $\beta_1$ | 1.12 | 0.64 | -0.14; 2.37 | 1.12 | 0.65 | -0.15; 2.38 | 1.23 | 0.61 | 0.01; 2.42 |
| $\beta_2$ | 0.38 | 0.61 | -0.82; 1.56 | 0.36 | 0.61 | -0.83; 1.56 | 0.51 | 0.58 | -0.63; 1.65 |
| $\beta_3$ | 2.97 | 0.25 | 2.49; 3.48 | 2.96 | 0.24 | 2.48; 3.42 | 2.99 | 0.24 | 2.52; 3.48 |
| $\beta_4$ | 0.38 | 0.16 | -0.82; 1.56 | 0.25 | 0.16 | -0.05; 0.57 | 0.25 | 0.16 | -0.05; 0.56 |
| $\beta_5$ | 0.31 | 0.16 | 0.01; 0.63 | 0.32 | 0.16 | 0.02; 0.63 | 0.31 | 0.16 | 0.01; 0.63 |
| $\beta_6$ | -0.25 | 0.04 | -0.33; -0.17 | -0.24 | 0.04 | -0.32; -0.17 | -0.25 | 0.04 | -0.33; -0.17 |
| $\theta_1$ | -3.28 | 0.54 | -4.39; -2.27 | -3.46 | 0.55 | -4.60; -2.43 | -3.44 | 0.54 | -4.53; -2.39 |
| $\theta_2$ | 0.01 | 0.03 | -0.06; 0.08 | 0.02 | 0.03 | -0.04; 0.09 | 0.02 | 0.03 | -0.05; -0.09 |
| $\theta_3$ | -0.03 | 0.02 | -0.08; 0.01 | -0.03 | 0.02 | -0.07; 0.01 | -0.03 | 0.02 | -0.08; 0.01 |
| $\theta_4$ | -0.01 | 0.03 | -0.06; 0.04 | -0.01 | 0.02 | -0.06; 0.04 | -0.01 | 0.03 | -0.06; 0.04 |
| $\theta_5$ | -1.64 | 1.27 | -4.88; -0.19 | -17.6 | 13.26 | -61.74; -0.80 | -8.42 | 6.28 | -23.47; -0.45 |

**Table 5.26: CD4+ count (square root) posterior means, standard deviations and credible intervals according to MNAR Bayesian analysis (selection models)**

**MNAR model 4** $m_{ij} \sim \text{Bernoulli}(p_{ij})$, $logit\ p_{ij} = \theta_1 + \theta_2 E_i y_{ij} + \theta_3 L_i y_{ij} + \theta_4 y_{ij} + \theta_5 E_i + \theta_6 L_i$

| | **MNAR model 4: Prior Set1** β~Normal(0, 100 000), Ω~Wishart(I,5) I: Identity matrix, $\theta_1$ *to* $\theta_5$,~logistic(0,1) | | | **MNAR model 4: Prior Set 2** β~Normal(0, 1000), Ω~Wishart(A,5) A: diag (0.1, 0.1, 0.1, 0.1, 0.1 $\theta_1$ to $\theta_5$ ~ Normal(0, 10000) | | | **MNAR model 4: Prior Set 3** β~Normal(0, 100), Ω~Wishart(A,5) A: diag (10, 10, 10, 10, 10) $\theta_1$ to $\theta_5$ ~Normal(0, 100) | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Mean** | **SD** | **95% credible interval** | **Mean** | **SD** | **95% credible interval** | **Mean** | **SD** | **95% credible interval** |
| $\beta_0$ | 8.79 | 0.50 | 7.81; 9.77 | 8.81 | 0.50 | 7.81; 9.79 | 8.63 | 0.50 | 7.63; 9.59 |
| $\beta_1$ | 1.14 | 0.65 | -0.14; 2.40 | 1.11 | 0.64 | -0.15; 2.38 | 1.23 | 0.63 | 0.03; 2.48 |
| $\beta_2$ | 0.38 | 0.61 | -0.82; 1.56 | 0.34 | 0.61 | -0.86; 1.54 | 0.49 | 0.60 | -0.69; 1.67 |
| $\beta_3$ | 2.94 | 0.25 | 2,45; 3.42 | 2.95 | 0.24 | 2.49; 3.45 | 3.00 | 0.25 | 2.52; 3.48 |
| $\beta_4$ | 0.25 | 0.16 | -0.07; 0.57 | 0.26 | 0.16 | -0.06; 0.57 | 0.24 | 0.16 | -0.07; 0.56 |
| $\beta_5$ | 0.32 | 0.16 | 0.01; 0.63 | 0.32 | 0.16 | 0.01; 0.64 | 0.31 | 0.16 | 0.004; 0.62 |
| $\beta_6$ | -0.24 | 0.04 | -0.32; -0.16 | -0.24 | 0.04 | -0.32; -0.17 | -0.25 | 0.04 | -0.33; -0.17 |
| $\theta_1$ | -2.65 | 0.71 | -4.10; -1.32 | -0.70 | 0.65 | -2.08; 0.22 | -2.83 | 0.81 | -4.49; -1.34 |
| $\theta_2$ | -0.04 | 0.08 | -0.19; 0.11 | 0.04 | 0.09 | -0.09; 0.22 | -0.03 | 0.09 | -0.20; 0.14 |
| $\theta_3$ | -0.04 | 0.07 | -0.18; 0.11 | -0.06 | 0.09 | -0.24; 0.10 | -0.05 | 0.09 | -0.23; 0.11 |
| $\theta_4$ | -0.04 | 0.05 | -0.13; 0.06 | -0.16 | 0.05 | -0.24; -0.06 | -0.02 | 0.05 | -0.13; 0.08 |
| $\theta_5$ | -0.36 | 1.22 | -2.77; 2.04 | -1.44 | 1.65 | -4.77; 0.19 | -0.42 | 1.50 | -3.45; 2.30 |
| $\theta_6$ | -0.27 | 1.11 | -2.54; 1.82 | 0.07 | 1.30 | -2.47; 2.29 | -0.10 | 1.35 | -2.72; 2.65 |

MNAR: Missing not at random, SD: Standard deviation

$\beta_0$ Intercept; $\beta_1$ Early integrated treatment arm (ref: sequential treatment arm); $\beta_2$ Late integrated treatment arm (ref: sequential treatment arm), $\beta_3$ Time; $\beta_4$ Interaction between treatment arm and time (early integrated compared to sequential treatment arm); $\beta_5$ Interaction between treatment arm and time (late integrated compared to sequential treatment arm); $\beta_6$: Time squared

Highlighted sections indicate statistical significance

All the models were fitted using OpenBugs, which uses MCMC methods and run for 40 000 iteration, including 20 000 burn-in iterations. Two chains with different starting values were used and convergence was assumed if a visual inspection of the trace plots was satisfactory. All the runs discussed converged.

The choice of prior distribution did not appreciably change the results. In addition, similar conclusions were drawn from all four of the MNAR models. In all four of the MNAR Bayesian models fitted the interaction between time and the early integrated treatment arm compared to the sequential treatment arm was not significant, under any of the prior sets used. In all four of the MNAR models fitted, the interaction between time and the late integrated treatment arm compared to the sequential treatment arm was significant. One of the 95% credible intervals included 0 (Model 2, prior set 3), but only barely so. We therefore conclude that the change in mean CD4+ count increase is larger in the late integrated treatment arm than in the sequential treatment arm, while the change over time in mean CD4+ count is similar in the early integrated treatment arm and the sequential treatment arm.

### 5.5.3 Shared parameter models

In this setting, the occurrence of an event, dropout, also corresponds to the discontinuation of the longitudinal process. Joint models were fitted for the longitudinal CD4+ counts and the time to dropout where both the measurement and dropout processes share random effects, conditional upon which they are assumed to be independent. It is assumed that some underlying individual characteristic, such as disease process, influences both whether a participant drops out of the study or whether CD4+ counts improve. The linear mixed model for CD4+ count included as fixed effects an intercept, the treatment and time variables and the interaction between the treatment variables and time. The measurement error terms are assumed to be independent and to follow a normal distribution with mean 0 and variance $\sigma^2$. Dropout is modelled through a linear effect of the treatment variables, intercept and a random disturbance term in a time to event model. The baseline hazard function follows a Weibull distribution. The connection between CD4+ count and time to dropout is accomplished via the random effects U and V. The random effects have mean 0 and the parameters $a_3$ and $a_4$ rescale their variances to account for the different scales of CD4+ count and the time variable. A Cholesky parameterisation of the random effects covariance matrix was used to ensure positive definiteness. Conditional on random effects we have independent Gaussian contributions. The initial values of the parameters were estimated by fitting the longitudinal model and survival model separately.

The following model was fitted where the random effect for the intercept, $U_i$, and the random effect for the slope, $V_i$, were shared between the survival and the longitudinal sub-models.

Survival sub-model: $h_i(t) = h_0(t)\exp[a_0 + a_1 E_i + a_2 L_i + a_3 U_i + a_4 V_i]$

where $h_0(t)$ is the baseline hazard function and follows a Weibull distribution.

Longitudinal sub-model:

$$Y_{ij} = \beta_1 + U_i + \beta_2 t_{ij} + V_i t_{ij} + \beta_3 E_i + \beta_4 L_i + \beta_5 E_i t_{ij} + \beta_6 L_i t_{ij} + \beta_7 t_{ij}^2 + \varepsilon_i$$

The following SAS code was used for the model with random intercept and slope linkage:

```
proc nlmixed data = dataset;
if (last) then do;
  linpsurv = a0 + adumearly*dummyearly + adumlate*dummylate + aInt*U0 + r2*u1;
  alpha = exp(-linpsurv*gamma);
  G_t = exp(-alpha*last_t**gamma);
  g = gamma*alpha*last_t**(gamma-1)*G_t;
  llsurv = (death=1)*log(g) + (death=0)*log(G_t);
 end;
else llsurv = 0;

v11 = a11*a11;
v12 = a11*a12;
v22 = a12*a12 + a22*a22;
```

```
linplong = (bl0 + u0) + (point*bt + u1*point)  + dummyearly*bdumearly +dummylate*bdumlate +
dummyearly*point*btdumearly +dummylate*point*btdumlate+ pointsquared*btsquare;
resid = (cdsquare-linplong);

   if (abs(resid) > 1E100) or (s2 < 1e-12) then do;
     Llong = -1e20;
   end; else do;
     Llong = -0.5*(1.837876 + resid**2 / s2  + log(s2));
   end;
model last ~ general (llong + llsurv);
random u0 u1 ~ normal ([0,0],[v11,v12,v22]) subject = pid;
estimate 'Var[u0]' v11;
estimate 'cov[u0,u1]' v12;
estimate 'var[u1]' v22;
run;
```

The following model was fitted where the random effect for intercept, $U_i$, only was shared between the survival and the longitudinal sub-models. In the longitudinal sub-model the $U_i$ serve as participant specific random effects and in the survival sub-model the $U_i$ serve as participant specific covariates.

Survival sub-model: $h_i(t) = h_0(\mathrm{t})\exp[a_0 + a_1\mathrm{E_i} + a_2\mathrm{L_i} + a_3\mathrm{U}_i]$

where $h_0(\mathrm{t})$ is the baseline hazard function and follows a Weibull distribution

Longitudinal sub-model: $Y_{ij} = \beta_1 + \mathrm{U}_i + \beta_2 t_{ij} + \beta_3\mathrm{E_i} + \beta_4\mathrm{L_i} + \beta_5\mathrm{E_i}t_{ij} + \beta_6\mathrm{L_i}t_{ij} + \beta_7 t_{ij}^2 + \varepsilon_i$

The following SAS code was used for the model with intercept linkage:

```
proc nlmixed data = dataset;
PARMS a0 = 10 adumearly = 0.38 adumlate = -0.13 bdumearly = 1.74 bdumlate = 1.34 btdumlate
= 0.05 btsquare = -0.28 bt = 3.445 btdumearly = 0.156 aint = 0;
if (last) then do;
   linpsurv = a0 + adumearly*dummyearly + adumlate*dummylate + aInt*U0 ;
   alpha = exp(-linpsurv*gamma);
   G_t = exp(-alpha*last_t**gamma);
   g = gamma*alpha*last_t**(gamma-1)*G_t;
   llsurv = (death=1)*log(g) + (death=0)*log(G_t);
  end;
else llsurv = 0;

var = s11*s11;

linplong = (bl0 + u0) + t*bt  + dummyearly*bdumearly +dummylate*bdumlate +
dummyearly*t*btdumearly +dummylate*t*btdumlate+ pointsquared*btsquare;
    resid = (cd4square-linplong);

   if (abs(resid) > 1E100) or (s2 < 1e-12) then do;
     Llong = -1e20;
   end; else do;
     Llong = -0.5*(1.837876 + resid**2 / s2  + log(s2));
   end;
         model last ~ general (llong + llsurv);
```

```
    random u0 ~ normal(0,var) subject= pid;
estimate 'Var[u0]' var;
    run;
```

**Table 5.27: Joint model for CD4+ count over time and time to dropout**

| | | Intercept linkage | | | Intercept and linear slope linkage | | |
|---|---|---|---|---|---|---|---|
| **Effect** | **Parameter** | **Estimate** | **Standard error** | **p-value** | **Estimate** | **Standard error** | **p-value** |
| **Longitudinal sub-model** | | | | | | | |
| Intercept | $\beta_1$ | 7.68 | 0.44 | <0.001 | 7.70 | 0.43 | <0.001 |
| Time | $\beta_2$ | 3.45 | 0.23 | <0.001 | 3.45 | 0.22 | <0.001 |
| Early integrated treatment arm (ref: sequential treatment arm) | $\beta_3$ | 1.76 | 0.53 | 0.001 | 1.70 | 0.53 | 0.002 |
| Late integrated treatment arm (ref: sequential treatment arm) | $\beta_4$ | 1.32 | 0.53 | 0.013 | 1.32 | 0.54 | 0.014 |
| Interaction between treatment arm and time (early integrated compared to sequential treatment arm) | $\beta_5$ | 0.15 | 0.11 | 0.152 | 0.18 | 0.13 | 0.168 |
| Interaction between treatment arm and time (late integrated compared to sequential treatment arm) | $\beta_6$ | 0.05 | 0.11 | 0.624 | 0.08 | 0.14 | 0.575 |
| Time squared | $\beta_7$ | -0.28 | 0.04 | <0.001 | -0.29 | 0.03 | <0.001 |
| **Time to dropout sub-model** | | | | | | | |
| Intercept survival model only | $a_0$ | 10.03 | 22.89 | 0.661 | 10.18 | 31.97 | 0.750 |
| Early treatment arm compared to sequential | $a_1$ | 0.39 | 21.53 | 0.986 | 0.43 | 22.08 | 0.984 |
| Late integrated treatment arm (ref: sequential treatment arm) | $a_2$ | 0.14 | 20.29 | 0.995 | 0.22 | 21.84 | 0.992 |
| Intercept shared | $a_3$ | 0.01 | 2.03 | 0.995 | -0.21 | 2.67 | 0.939 |
| Slope shared | $a_4$ | | | | 1.06 | 16.13 | 0.948 |

The results were similar for the joint model where a random slope and intercept were shared between the sub-models and where a random intercept only was shared. Both models lead to the conclusion that neither the early nor the late integrated treatment arm was associated with a reduction in mean CD4+ count compared to the sequential treatment arm. The shared intercept ($a_3$) estimate was close to 0 in the random intercept joint model and not significant in both models, indicating that baseline CD4+ count did not influence dropout time. The shared slope was positive and not significant. The positive slope indicated that participants with larger linear increase in CD4+ count also had an increase in dropout time, but this was not statistically significant (Table 5.27).

These joint models differ from the MAR models by allowing the longitudinal CD4+ count profiles and the missingness process to share random effects. The models need to be interpreted conditional on these random effects. Shared-parameter and selection models differ in how they relate the probability of a missing observation and the longitudinal CD4+ counts. Shared-random

effects models link the two by relating a participant's propensity to respond with his or her propensity to miss a visit, while selection models directly model the probability of a missed visit as a function of the response. Shared-parameter models are more appropriate when the missingness is related to an individual's disease process.

The substantive model in these joint models differs from the substantive model fitted previously. SAS is not able to fit a joint model with the same substantive model as previously fitted. In order to compare the joint model fitted with a model under the MAR case, the MAR model was fitted as a random intercept model. The within-participant effect was taken into account by adding a random participant specific effect to the model. This model was fitted with a compound symmetry covariance matrix. A model with a subject specific random effect and random slope was also fitted. These models are given in Table 5.28.

**Table 5.28: MAR longitudinal model with random subject specific effects: comparable to the longitudinal sub-model in Table 6.27**

| | | Intercept linkage | | | Intercept and linear slope linkage | | |
|---|---|---|---|---|---|---|---|
| **Effect** | **Parameter** | **Estimate** | **Standard error** | **p-value** | **Estimate** | **Standard error** | **p-value** |
| **Longitudinal sub-model** | | | | | | | |
| Intercept | $\beta_1$ | 7.68 | 0.44 | <0.001 | 7.78 | 0.37 | <0.001 |
| Time | $\beta_2$ | 3.45 | 0.23 | <0.001 | 3.37 | 0.27 | <0.001 |
| Early integrated treatment arm (ref: sequential treatment arm) | $\beta_3$ | 1.74 | 0.53 | 0.001 | 1.71 | 0.38 | <0.001 |
| Late integrated treatment arm (ref: sequential treatment arm) | $\beta_4$ | 1.34 | 0.53 | 0.011 | 1.20 | 0.38 | 0.002 |
| Interaction between treatment arm and time (early integrated compared to sequential treatment arm) | $\beta_5$ | 0.16 | 0.11 | 0.145 | 0.20 | 0.18 | 0.269 |
| Interaction between treatment arm and time (late integrated compared to sequential treatment arm) | $\beta_6$ | 0.05 | 0.11 | 0.639 | 0.14 | 0.18 | 0.452 |
| Time squared | $\beta_7$ | -0.28 | 0.04 | <0.001 | -0.28 | 0.04 | <0.001 |

Comparing the β-estimates between Table 5.27 and Table 5.28 there is almost no difference in the random intercept model. The random intercept and slope model had small differences between the MAR and MNAR models, but the same conclusions are drawn from both models. It seems that the deviations from MAR assumed by the joint model do not change the conclusions.

Death as a cause of missing data deserves special attention. In order to investigate the relationship between CD4+ count and death, the same joint models were fitted, with the only change that the survival model modelled the time to death instead of the time to dropout. Participants who dropped

out were censored. This model creates a joint model of the probability of being alive and healthy. The results are given in Table 5.29 and were similar to the results of the joint model for time to dropout and longitudinal CD4+ count. The similarity in results can partly be explained by the fact that a large proportion of the dropouts were caused by death. For the participants who died time to death and time to dropout was the same.

**Table 5.29: Joint model of longitudinal CD4+ count and time to death**

| | | Intercept linkage | | | Intercept and linear slope linkage | | |
|---|---|---|---|---|---|---|---|
| **Effect** | **Parameter** | **Estimate** | **Standard error** | **p-value** | **Estimate** | **Standard error** | **p-value** |
| **Longitudinal sub-model** | | | | | | | |
| Intercept | $\beta_1$ | 7.68 | 0.44 | <0.001 | 7.70 | 0.43 | <0.001 |
| Time | $\beta_2$ | 3.45 | 0.23 | <0.001 | 3.45 | 0.22 | <0.001 |
| Early integrated treatment arm (ref: sequential treatment arm) | $\beta_3$ | 1.77 | 0.53 | 0.001 | 1.71 | 0.53 | 0.002 |
| Late integrated treatment arm (ref: sequential treatment arm) | $\beta_4$ | 1.32 | 0.53 | 0.013 | 1.30 | 0.54 | 0.016 |
| Interaction between treatment arm and time (early integrated compared to sequential treatment arm) | $\beta_5$ | 0.15 | 0.11 | 0.153 | 0.18 | 0.13 | 0.173 |
| Interaction between treatment arm and time (late integrated compared to sequential treatment arm) | $\beta_6$ | 0.05 | 0.11 | 0.623 | 0.08 | 0.14 | 0.564 |
| Time squared | $\beta_7$ | -0.28 | 0.04 | <0.001 | -0.29 | 0.03 | <0.001 |
| **Survival sub-model** | | | | | | | |
| Intercept survival model only | $a_0$ | 10.59 | 12.85 | 0.410 | 11.29 | 14.80 | 0.446 |
| Early treatment arm compared to sequential | $a_1$ | -0.08 | 10.65 | 0.994 | 0.18 | 10.42 | 0.986 |
| Late integrated treatment arm (ref: sequential treatment arm) | $a_2$ | 0.08 | 12.10 | 0.995 | 0.32 | 11.24 | 0.977 |
| Intercept shared | $a_3$ | 0.01 | 1.05 | 0.992 | 0.05 | 1.15 | 0.963 |
| Slope shared | $a_4$ | | | | 1.00 | 7.45 | 0.894 |

### 5.5.4 Conclusions of MNAR models

Several MNAR models were fitted using both pattern-mixture models and selection models. These methods form part of a sensitivity analysis and is by no means an exhaustive list of plausible models that could be fitted. In fact, this is not even an exhaustive list of the different methods that could be used to fit these models. Nor do the models fitted give any indication as to which of the models fitted are the most appropriate models, since these models rely on unverifiable assumptions.

**Table 5.30: Summary of findings in MNAR sensitivity analyses**

| | Interaction between treatment arm and time | | | |
|---|---|---|---|---|
| | Early integrated compared to sequential treatment arm | | Late integrated compared to sequential treatment arm | |
| | Estimate | p-value | Estimate | p-value |
| **Pattern-mixture model** | | | | |
| Pattern-mixture model using random-effects mixed models for each of the patterns of missing data (Table 5.22FTa) | 0.24 | 0.33 | 0.51 | 0.02 |
| Pattern-mixture models using multiple imputation | | | | |
| Model 1: δ for dropout = 0.5 and δ for participants who died = 1 | 1.11 | 0.23 | 0.61 | 0.39 |
| Model 2; δ for dropout = 1, δ for participants who died = 2 | 1.56 | 0.04 | 0.84 | 0.19 |
| Model 3; δ for dropout = 0.2; δ for participants who died = 0.5 | 0.88 | 0.35 | 0.52 | 0.45 |
| Model 4; δ for dropout = 0.3; δ for participants who died = 1 | 1.07 | 0.25 | 0.62 | 0.37 |
| Model 5; δ for dropout = -1; δ for participants who died = 2 | 1.46 | 0.03 | 1.00 | 0.04 |
| Model 6; δ for dropout = -0.5; δ for participants who died = 0.5 | 0.84 | 0.37 | 0.60 | 0.35 |
| Pattern-mixture models using identifying restrictions | | | | |
| CCMV | 0.16 | 0.19 | 0.11 | 0.35 |
| NCMV | -0.11 | 0.65 | 0.15 | 0.55 |
| **Selection model approaches (Prior set 1)** | | | | |
| Bayesian MNAR model 1 | 0.25 | NS | 0.32 | Sign |
| Bayesian MNAR model 2 | 0.26 | NS | 0.33 | Sign |
| Bayesian MNAR model 3 | 0.38 | NS | 0.31 | Sign |
| Bayesian MNAR model 4 | 0.25 | NS | 0.32 | Sign |
| **Shared-parameter model** | | | | |
| Joint model with intercept shared | 0.15 | 0.15 | 0.05 | 0.62 |
| Joint model with intercept and slope shared | 0.18 | 0.17 | 0.08 | 0.58 |

Sign = significant, NS = Not significant

The answer to the question of whether mean CD4+ count increases over time differed between the treatment arms was complicated. If we look at the interaction between time and the early integrated treatment arm compared to the sequential treatment arm, we conclude that this interaction was not statistically significant. It was statistically significant only in Models 2 and 5 with multiple imputation which makes quite strong assumptions, namely that participants who died had much lower mean CD4+ counts than participants who remained in the study. We therefore conclude that unless very strong assumptions are made about CD4+ counts that are not observed, there is no difference between the early integrated and sequential treatment arms over time in the change in CD4+ count (Table 5.30).

Regarding the interaction between time and the late integrated treatment arm compared to the sequential treatment arm, half the models found a significant interaction and the other half did not.

The models that found significant interactions were the pattern-mixture model using random-effects mixed models for each of the patterns of missing data, Model 5 with multiple imputation and all the selection models. In the former, participants who died had lower mean CD4+ counts than other participants. There were more participants who died in the sequential treatment arm and we speculate that this pattern reduced the mean CD4+ count in the sequential treatment arm. As mentioned previously, Model 5 with multiple imputation makes quite strong assumptions, namely that the imputed CD4+ count for participants who died was lower and the imputed CD4+ count for participants who dropped out was higher than if they had remained in the study. The coefficients estimated under the selection model were similar to the coefficients estimated under the MAR case with the MAR Bayesian model. Therefore these models did not show major sensitivity for the deviations from MAR assumed under these models.

The estimand considered under MNAR assumptions is the same as the one under the MAR assumptions. We would like to have an estimator for Estimand 1 as described by the National Research Council (2010). This is the de facto or ITT estimand, including all data on all participants analysed as randomised. The different MNAR models fitted merely made different assumptions about the data not observed, but did not change the estimand of interest.

Under the pattern-mixture model using random-effects models for each of the patterns of missing data, if the future behaviour of observations from a participant in a specific pattern of missing data, conditional on the history, was correctly modelled then this analysis leads to an estimate for Estimand 1. We assumed that participants who dropped out continued to adhere to treatment post dropout and we also assumed that participants who died continued with the same drug taking behaviour as prior to death.

The analysis using multiple imputation includes data on participants after withdrawal or loss to follow-up, because the missing data are imputed according to previously observed data and the assumptions made about the missing data. This analysis could therefore give a valid analysis under the requirement for an ITT analysis when data are incomplete, provided the imputer's model used is correct and the assumptions made about missing data after dropout is correct.

The analyses using identifying restrictions include data after withdrawal or loss to follow-up, because the identifying restrictions borrow information from the appropriate pattern of data to complete the profiles. This analysis could therefore give a valid analysis under the requirement for an ITT analysis provided assumptions made about missing data after dropout is correct. The selection model analysis gives Estimand 1, but is sensitive to correct specification of the model.

Pattern-mixture models, fitted under the MNAR assumption, allow one to treat missing data due to drop out and death differently. The models fitted in Sections 5.5.1.1 and 5.5.1.2 explicitly fitted different models for participants who died and those who dropped out. Although these models did not make the assumption that trajectories were the same for participants who died and those who dropped out, these models still imputed CD4+ counts for participants post death; thus creating an immortal cohort. These models could be adjusted to only impute data for participants who dropped out and to censor participants who died. In Table 5.29 a joint model for longitudinal CD4+ count and time to death was fitted. This model modelled the effect of treatment arm on changes in CD4+ count jointly with the probability of surviving. None of the other models fitted under the MNAR assumption treated missingness due to death differently from missingness due to drop out.

We cannot interpret the CD4+ count trajectories over time as the CD4+ counts of participants who were still alive. Where declining CD4+ count preceded death and where lower CD4+ counts were imputed for participants who died, one would assume that treatment arms with higher death rates would have lower CD4+ counts modelled at later time points. If CD4+ count is modelled because it is a surrogate marker for disease status, with very low CD4+ count as predictive of pending death, this might not be as incorrect as it sounds.

## 5.6 Conclusion

The conclusion reached when the survival data were analysed was that integrated treatment saved lives and the recommendation was made that tuberculosis and HIV treatment should be integrated (Abdool Karim et al., 2010; Abdool Karim et al., 2011). A secondary question is whether integrated treatment changed the HIV treatment outcomes, measured by CD4+ counts. In this study about a third of the CD4+ counts are missing, either due to loss to follow-up or death. The CD4+ count over time can therefore not be evaluated without taking missing data into account.

There is evidence that the missing data process is not MCAR. The missing data process could be MAR, and MNAR cannot be ruled out. Under the MAR assumption, the likelihood-based method and the Bayesian method showed a significant difference between treatment arms in mean CD4+ count over time. The early and late integrated treatment arms both had a higher mean increase in CD4+ count over time than the sequential treatment arm. The other methods did not show a significant effect of treatment arm over time, although the pattern-mixture model under the ACMV identifying restriction had a small p-value for the interaction between time and the early integrated treatment arm. Under MNAR the majority of methods did not find a significant interaction between time and the early integrated treatment arm compared to the sequential treatment arm. The interaction between time and the late integrated treatment arm compared to the sequential treatment arm was not significant under most assumptions, the selection model approach is an

exception. With the selection models, which gave results similar to the MAR case, this interaction was significant. The conclusion therefore needs to be tempered slightly by stating that treatment arm could have an effect on mean CD4+ count over time, under certain assumptions made about the unobserved data.

This secondary analysis was done to determine whether integrated HIV and tuberculosis treatment would compromise CD4+ counts. Although the results were not consistent under all the assumptions made and with all the different methods used, none of these analyses found any evidence that CD4+ count increased more in the sequential treatment arm than in the other treatment arms. We can therefore safely recommend integrated treatment, knowing that as a treatment policy this will save lives and not compromise CD4+ counts. CD4+ count measures immune system health and is not specific to HIV. Tuberculosis treatment on its own is known to increase in CD4+ count (Martin et al., 1995), which probably explains some of the similarity in all the treatment arms, since all participants started tuberculosis treatment shortly prior to being enrolled in the study.

# Chapter 6

# St John's Wort trial

In the previous chapter we applied the theory around analysis of incomplete longitudinal data to a clinical trial with three treatment arms in HIV-infected patients. We continue our investigation of a proper analysis of a clinical trial with missing data with a second example. We follow the same organisation as in the previous chapter. In Section 1 we describe the design and published results of the trial. Section 2 follows with a description of the missing data in the trial. We then analyse the data using MCAR methods in Section 3, MAR methods in Section 4 and MNAR methods in Section 5. We conclude the chapter by contrasting these analyses and drawing a conclusion regarding the findings.

## 6.1 The St John's Wort (hypericum perforatum) trial

This trial was chosen to illustrate missing data analysis because of large amounts of missing data. In addition to participants discontinuing from the trial for various reasons and lost to follow up as in the previous SAPiT example, the trial also introduced missing data from Week 8 onwards by design. The design of the trial is discussed in Section 6.1.2. The next section describes the design and available results of the hypericum trial.

### 6.1.1 Background to the St John's Wort trial

Hypericum perforatum, also known as St John's Wort, is a herbal remedy believed to treat, among other things, depression. Especially in European countries, such as Germany, hypericum perforatum is often prescribed for treatment of mild depression, also in children and adolescents (Fegert et al., 2006). It was shown to be more effective than placebo (Kalb, Trautmann-Sponsel, & Kieser, 2001) in treating depression and is believed to have fewer side effects than standard antidepressive therapies, thus making it an attractive option for many patients (Linde et al., 1996). Some studies and meta-analyses have found hypericum to be as effective as standard antidepressive therapies (Linde, Berner, & Kriston, 2008; Rahimi, Nikfar, & Abdollahi, 2009), while other studies found no difference between hypericum and placebo (Shelton et al., 2001)

Because different studies had contradictory results about the relative effectiveness of hypericum when compared to placebo and standard antidepressive drugs, a trial was designed to compare both a standard antidepressive therapy (sertraline) and hypericum to placebo (Hypericum Depression Trial Study Group, 2002). Sertraline is an antidepressant of the selective serotonin reuptake inhibitor (SSRI) class. It was brought to market by Pfizer in 1991 under the trademark name Zoloft.

### 6.1.2 Methods

The trial was a randomised, double blind, parallel-arm, 8-week, outpatient trial of hypericum, sertraline, or placebo treatment for major depressive disorder, followed by 18 weeks of double blind continuation treatment in participants meeting response criteria at Week 8 (Hypericum Depression Trial Study Group, 2002). Participants with major depression were recruited from December 1998 to June 2000 at 12 centres.

The trial included outpatients with major depression of moderate severity. Specifically, the inclusion criteria included: aged 18 years or more; diagnosis of major depression; a score of 20 or more on the Hamilton Depression scale (HAM-D) (Hamilton, 1960), and a score of 60 or less on the Global Assessment of Functioning (GAF) scale at screening and baseline, capable of giving informed consent and identification of a personal contact. Exclusion criteria included: suicide or homicide risk, current or planned pregnancy, breastfeeding or not using birth control; and some relevant medical conditions, such as liver disease, use of either hypericum or sertraline in the previous 6 months or current use of other psychotropic drugs or current psychotherapy. Complete inclusion and exclusion criteria are available in the trial protocol and in the trial paper (Hypericum Depression Trial Study Group, 2002).

The eligibility of participants was assessed, after which they gave written informed consent and participated in a one-week placebo run-in. Participants meeting eligibility criteria after this one-week run-in were randomised to one of the three treatment arms in a 1:1:1 ratio. Participants were assessed weekly from Week 1 to Week 8 (with the exception of Week 5). Participants who were regarded as responders at Week 8 (the end of the acute phase) could enter the continuation phase, with visits at Weeks 10, 14, 18, 22 and 26. The HAM-D, Global Assessment of Functioning (GAF) scale, Clinical Global impressions scales for severity (CGI-S) and improvement (CGI-I) and the Beck Depression Inventory (BDI) were assessed at all visits. Other safety related information was also collected; such as vital signs, adverse events and blood chemistry and haematology (Hypericum Depression Trial Study Group, 2002). For the purpose of this study we analyse the response over time on the HAM-D. The HAM-D scale is a measure of depression, with a higher score indicative of more severe depression.

Hypericum and matching placebo were provided by Lichtwer Pharma and sertraline and matching placebo were provided by Pfizer. Dosing was three times daily. Participants were started on a standard dose, which could be increased according to a predetermined algorithm if required. Medication was provided in double-dummy fashion (Hypericum Depression Trial Study Group, 2002).

Response at Week 8 was defined as either full or partial response. Full response was defined as a CGI-I score of 1 or 2 and a HAM-D total score of 8 or less. Partial response was defined as a CGI-I score of 1 or 2, a decrease in the HAM-D score from baseline of at least 50% and a HAM-D score between 9 and 12 (Hypericum Depression Trial Study Group, 2002).

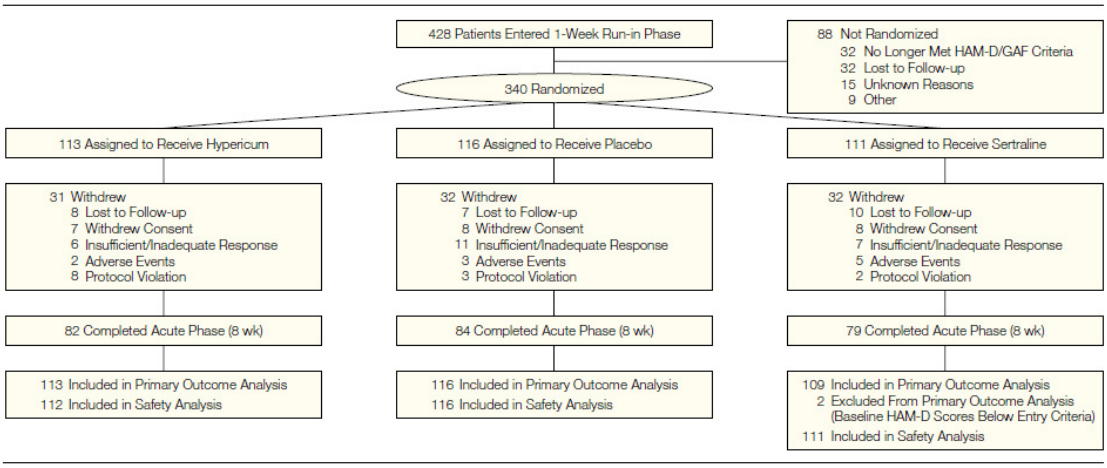
### 6.1.3 Primary and secondary endpoints

The main hypothesis was whether hypericum was superior to placebo after 8 weeks. The endpoints were defined as the change in the HAM-D score from baseline to Week 8 and the proportion of participants who had a full response at Week 8. The principal comparison was between the hypericum and placebo arms. The sertraline arm was included as an active control arm to validate the study, but no comparison between hypericum and sertraline was intended and the trial was not powered for such a comparison or for multiple comparisons with placebo (Hypericum Depression Trial Study Group, 2002).

In the original analysis treatment differences in the change in HAM-D score from baseline to Week 8 were evaluated through a random-coefficient linear regression model. Fixed effects were treatment, site, sex, week and treatment by week interaction. A random intercept and slope over time were included for each participant. Tests of treatment differences for the change in HAM-D

scores from baseline to Week 8 were equivalent to tests of treatment differences in the slopes over time. The original paper also included a completers' analysis and a LOCF analysis of the acute phase. No endpoint data were presented from the continuation phase (Hypericum Depression Trial Study Group, 2002).

The trial enrolled 428 participants in the run-in phase and 340 were randomised to the three treatment arms (Figure 6.1). The demographic data of the sample is described elsewhere (Hypericum Depression Trial Study Group, 2002). They found that the HAM-D scores declined with time and differed by site and sex. The scores were lower for men. They did not find that trend with time differed significantly by treatment; either for the comparison between hypericum and placebo or for the comparison between sertraline and placebo. The full response rate at Week 8 did not differ between hypericum and placebo or between sertraline and placebo. The conclusion was that neither hypericum nor sertraline were superior to placebo on the primary efficacy measures. The authors highlighted the high level of improvement in placebo arms often seen in depression trials.




HAM-D indicates Hamilton Depression scale; GAF, Global Assessment of Functioning.

**Figure 6.1: Disposition of the trial participants during the acute phase taken from Hypericum depression trial study group (2002).**

A separate publication described the results during the continuation phase using a completers' analysis and LOCF. This analysis confirmed the findings during the acute phase, namely that participants in all three treatment arms showed a large improvement over time, but there was no difference between the treatment arms (Sarris et al., 2012).

Our re-analysis of the original data included an analysis of the HAM-D score only. The longitudinal pattern of this variable was evaluated, taking account of missing data, rather than just

looking at the change from baseline to Week 8 as in the original analysis. In addition, the data from Week 8 to Week 26 were analysed. This analysis was not done in the original paper.

The decision to continue a participant to the continuation phase was based on the response of the participant up to Week 8. Since only responders continued with the continuation phase, the data up to Week 8 can be used to predict whether a participant continued beyond Week 8. Missingness after Week 8 was, by design, based on observed data (HAM-D score), therefore the missing data were likely MAR. However, data could also be missing for reasons not related to treatment response, therefore the data were not guaranteed to be MAR. MCAR could, however be ruled out.

This study created missing data by design, because participants were discontinued from the study at Week 8 if they did not respond. It is therefore questionable whether the missing data from Week 8 onwards should really be regarded as missing data.

Since there was a large amount of missing data after Week 8, some estimates were unstable and the analysis could be questionable, since it relied on the assumptions about the missing data pattern more than on the actual observed data. To mitigate this, the data were analysed from Week 0 to Week 8, accounting for missing data, and the data were also separately analysed from Week 0 to Week 26, while accounting for missing data. This also answered the research question of treatment effect during the acute phase (First 8 weeks) only.

As was the case with the SAPiT study in Chapter 5, the hypericum trial had three treatment arms. The relevant issues when more than two arms are to be compared were discussed in Section 5.3. This study was designed to primarily compare the hypericum and placebo arms; therefore the primary focus of the analysis is on that pairwise comparison. The pairwise comparison between the sertraline and placebo arms was also done. In this study the interest lay in the pairwise comparisons between each of the active arms and placebo, in contrast to the study in Chapter 5 where it was not clear what the control arm was. All three treatment arms were included in the covariance matrix to increase the precision of models.

### 6.1.4 Statistical notation

The statistical notation was introduced in Section 2.1 and is adapted to this data set as follows. We assume $N$ independent participants, indicated by $i = 1, \ldots, N$. The HAM-D score for the $i^{th}$ participant at the $j^{th}$ occasion is given by $Y_{ij}$. Treatment is indicated by the two variables $H_i$ and $S_i$. If the $i^{th}$ participant is in the hypericum arm, then $H_i = 1$, else $H_i = 0$. If the $i^{th}$ participant is in the sertraline arm, then $S_i = 1$, else $S_i = 0$. $t_{ij}$ is the week for the $i^{th}$ participant at the $j^{th}$ occasion.

When dropout is defined as a binary variable, dropout is indicated by $D_i$. If participant $i$ dropped out, then $D_i = 1$, if participant $i$ did not drop out then $D_i = 0$.

## 6.2 Patterns of missing data

The number of missing data points over time and the reasons for having missing data are provided in Table 6.1. Trial discontinuation prior to Week 8 was high; 31 (27.4%) participants in the hypericum arm, 32 (28.8%) participants in the sertraline arm and 32 (27.6%) participants in the placebo arm discontinued prior to Week 8. A large number of these discontinued because of insufficient response (24), although some were lost to follow-up (25) and some withdrew consent (23), often because they wanted to pursue other therapies since they perceived the therapy not to be of benefit. There is no relationship between treatment arm and the number of participants not completing the acute phase of the study (Chi-square test, 2 degrees of freedom, p = 0.99).

A large number of participants did not meet the protocol definition of a responder in order to be allowed to continue with the continuation phase. Only a small number of participants entered the continuation phase; 49 (44.1%) in the sertraline arm, 38 (33.6%) in the hypericum arm and 42 (36.2%) in the placebo arm. Of those who entered the continuation phase, only a small number completed the entire study to Week 26; 28 (57.1%) in the sertraline arm, 24 (63.2%) in the hypericum arm and 27 (64.3%) in the placebo arm. Table 6.2 gives the pattern of missing data. It is clear that there is a large amount of missing data in this data set.

A central question is whether the missing data are MCAR, MAR or MNAR. Since only participants who showed a response on treatment entered the continuation phase, missing data are probably MAR by design from Week 8 onwards. There was also missing data caused by reasons other than discontinuation due to non-response, therefore MNAR cannot be ruled out. The large amount of missing data prior to Week 8 could be MAR or MNAR. As mentioned in Section 2.2.3, although the data cannot exclude MNAR, it can hold some evidence of informative missingness.

Figure 6.2 compares the mean HAM-D score of participants who dropped out at the next visit to those of participants who did not drop out at the next visit. It seems that participants who dropped out are different from those who did not drop out. Prior to Week 4 those who dropped out had lower HAM-D scores than those who did not drop out. From Week 4 onwards participants who dropped out had higher HAM-D scores (indicative of more severe depression) than participants who did not drop out. This is also true during the continuation phase. This suggests that dropout depends at least on observed HAM-D score, and implies that model-based means could be different from observed means. It also suggests that the observed HAM-D scores were associated with

dropout. Figure 6.3 gives the same summary by treatment arm. The pattern over time was similar in all three treatment arms for participants who did not drop out.

**Table 6.1: Number of participants attending each visit in the hypericum trial**

| | Sertraline N = 111 | | Hypericum N = 113 | | Placebo N = 116 | |
|---|---|---|---|---|---|---|
| **Number with Hamilton depression scale score at** | **N** | **Number missing (%)** | **N** | **Number missing (%)** | **N** | **Number missing (%)** |
| Baseline | 111 | - | 113 | - | 116 | - |
| Week 1 | 101 | 10 (9.0) | 101 | 12 (10.6) | 111 | 5 (4.3) |
| Week 2 | 90 | 21 (18.9) | 102 | 11 (9.7) | 107 | 9 (7.8) |
| Week 3 | 90 | 21 (18.9) | 100 | 13 (11.5) | 94 | 22 (19.0) |
| Week 4 | 89 | 22 (19.8) | 97 | 16 (14.2) | 99 | 17 (14.7) |
| Week 6 | 82 | 29 (25.1) | 91 | 22 (19.5) | 93 | 23 (19.8) |
| Week 7 | 79 | 32 (28.8) | 82 | 31 (27.4) | 84 | 32 (27.6) |
| Week 8 | 79 | 32 (28.8) | 82 | 31 (27.4) | 84 | 32 (27.6) |
| **Enter continuation phase** | **49** | **62 (55.9)** | **38** | **75 (66.4)** | **42** | **74 (63.8)** |
| Week 10 | 48 | 63 (56.8) | 35 | 78 (69.0) | 39 | 77 (66.4) |
| Week 14 | 43 | 68 (61.3) | 33 | 80 (70.8) | 37 | 79 (68.1) |
| Week 18 | 39 | 72 (64.9) | 27 | 86 (76.1) | 32 | 84 (72.4) |
| Week 22 | 32 | 79 (71.2) | 25 | 88 (77.8) | 27 | 89 (76.7) |
| Week 26 | 31 | 80 (72.1) | 24 | 89 (78.8) | 27 | 89 (76.7) |
| **Reasons not completing acute phase (Week 8)** | | | | | | |
| Loss to follow-up | 10 | | 8 | | 7 | |
| Insufficient response | 7 | | 6 | | 11 | |
| Withdrew consent | 8 | | 7 | | 8 | |
| Adverse event | 5 | | 2 | | 3 | |
| Protocol violation | 2 | | 8 | | 3 | |
| **Reasons not enrolled into continuation phase** | | | | | | |
| Did not complete acute phase | 32 | | 31 | | 32 | |
| Non-responder | 27 | | 40 | | 38 | |
| Responder, but did not want to continue to continuation phase | 3 | | 4 | | 4 | |
| **Reasons not completing continuation phase (Week 26)** | | | | | | |
| Completed | 28 | | 24 | | 27 | |
| Insufficient response | 3 | | 2 | | 4 | |
| Loss to follow up | 1 | | 2 | | 1 | |
| Adverse event | 3 | | | | 2 | |
| Withdrew consent | 12 | | 8 | | 5 | |
| Protocol violation: ineligible | 2 | | 2 | | 3 | |

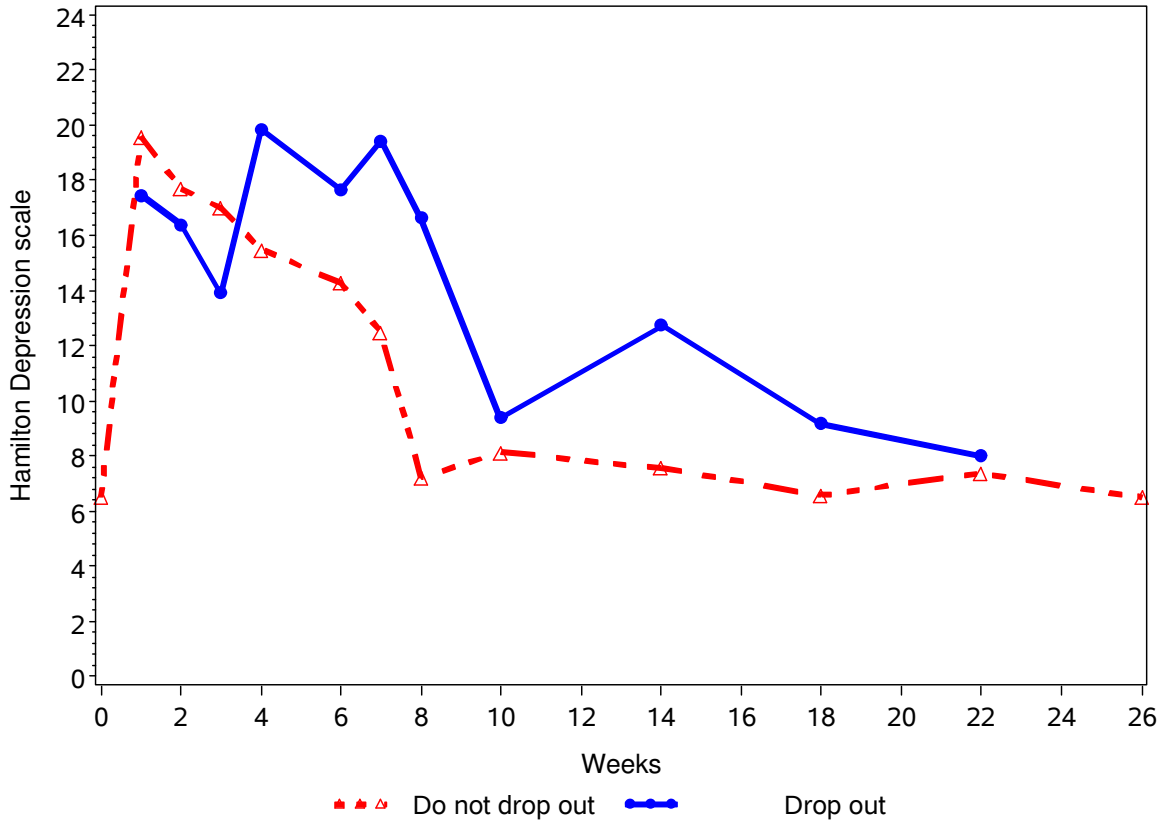



**Figure 6.2: Sample mean of Hamilton depression scale at each week for all treatment arms combined.**

The triangles in red are the means for participants who did not drop out before the subsequent measurement. The blue solid dots are the means for participants who dropped out before the subsequent measurement.

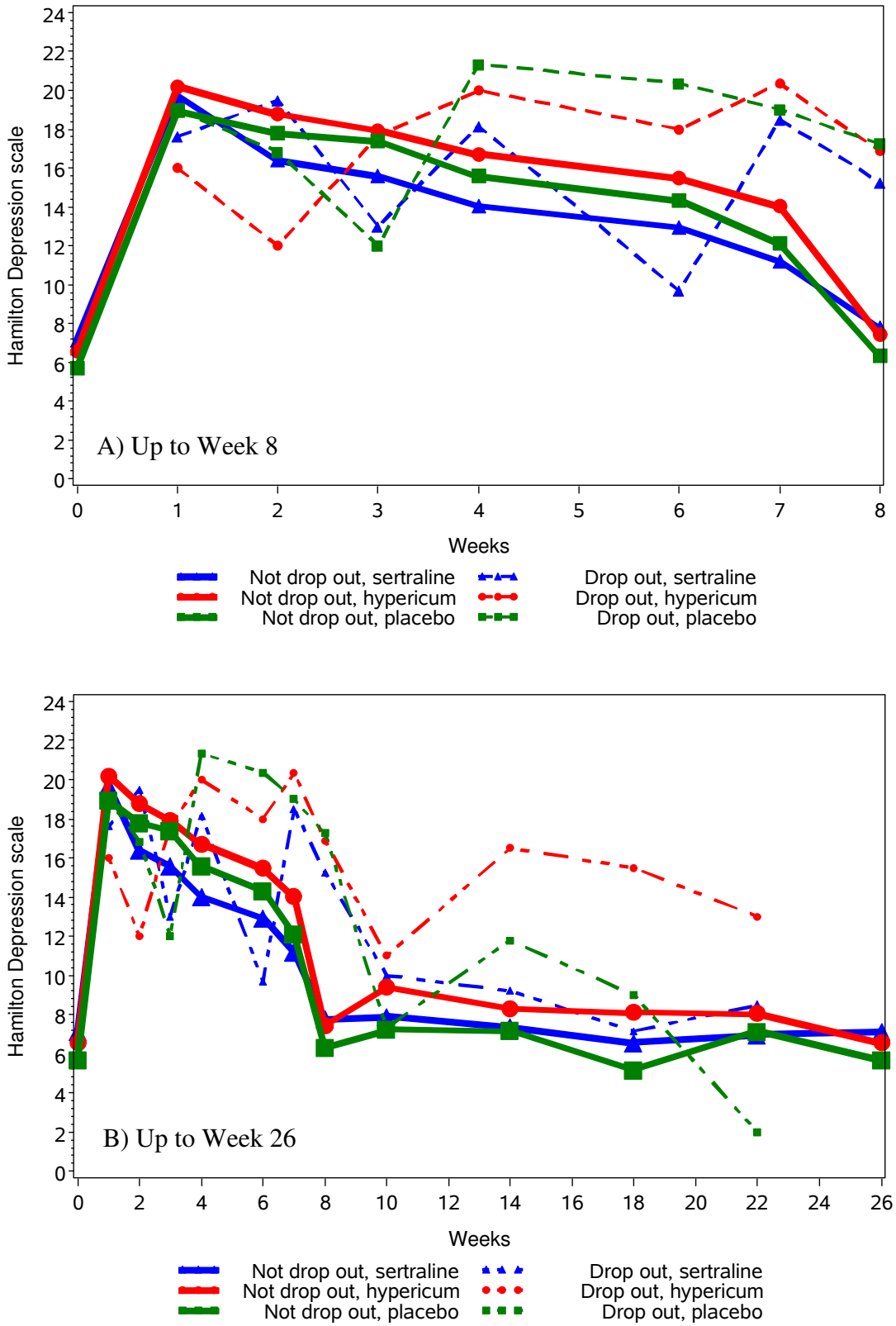



**Figure 6.3: Sample mean of Hamilton depression scale at each week by treatment group.**

The solid lines are the means for participants who did not drop out before the subsequent measurement. The dotted lines are the means for participants who dropped out before the subsequent measurement.

**Table 6.2: Pattern of missing data in each of the treatment arms in the hypericum trial**

| | Week | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acute phase | | | | | | | Continuation phase | | | | | | Sertraline | Hypericum | Placebo |
| Pattern | 1 | 2 | 3 | 4 | 6 | 7 | 8 | 10 | 14 | 18 | 22 | 26 | Description | N = 111 | N = 113 | N =116 |
| 1 | X | X | X | X | X | X | X | X | X | X | X | X | Complete* | 31 | 24 | 27 |
| 2 | X | X | X | X | X | X | X | X | X | X | X | | | 2 | 1 | 1 |
| 3 | X | X | X | X | X | X | X | X | X | X | | | * | 6 | 2 | 4 |
| 4 | X | X | X | X | X | X | X | X | X | | | | * | 5 | 6 | 5 |
| 5 | X | X | X | X | X | X | X | X | | | | | * | 5 | 2 | 3 |
| 6 | X | X | X | X | X | X | X | | | | | | Non-response* | 27 | 39 | 37 |
| 7 | X | X | X | X | X | X | X | | | | | | Response but refused continuation* | 3 | 4 | 4 |
| 8 | X | X | X | X | X | X | X | | | | | | Continued to continuation phase, did not have data | 0 | 3 | 2 |
| 9 | X | X | X | X | X | X | | | | | | | * | 2 | 2 | 2 |
| 10 | X | X | X | X | X | | | | | | | | * | 2 | 7 | 5 |
| 11 | X | X | X | X | | | | | | | | | * | 6 | 5 | 6 |
| 12 | X | X | X | | | | | | | | | | * | 2 | 3 | 5 |
| 13 | X | X | | | | | | | | | | | * | 2 | 2 | 6 |
| 14 | X | | | | | | | | | | | | | 10 | 2 | 1 |
| 15 | | | | | | | | | | | | | Baseline only | 7 | 9 | 4 |
| 16 | With more than one interim missing data point | | | | | | X | | | | | | | 1 | 2 | 4 |

* A minority of participants in this pattern missed a visit, but attended all visits thereafter.

As suggested by Carpenter and Kenward (2007) we used logistic regression to determine key independent variables associated with withdrawal in either the acute or continuation phases (Table 6.3). In the acute phase age (OR = 0.97) and duration of current depression (OR = 1.04) were associated with dropping out prior to Week 8. Older people were less likely to drop out than younger people and participants with longer duration of depression were more likely to drop out. None of the baseline demographic variables were indicative of drop out during the continuation phase. The only variable associated with dropout during the continuation phase was the HAM-D score at Week 8 (OR = 1.24) (Table 6.3).

Since withdrawal in the acute phase was associated with baseline variables the data could not be assumed to be MCAR and any data analysis that made the assumption of MCAR data could be biased and give incorrect results. It was not possible to rule out that the data were MNAR using the observed data only, since the definition of MNAR depended on unobserved observations. We therefore made the assumption that the HAM-D data were either MAR or MNAR and proceeded to do a sensitivity analyses under these assumptions.

The high withdrawal rate (about a quarter of participants did not complete the acute phase of the study) combined with the fact that many withdrawals were due to insufficient response suggested

that HAM-D score could not be analysed without taking missing data into account. The goal of this analysis was to analyse the HAM-D score in a valid way, while taking missing data into account.

**Table 6.3: Variables associated with dropping out, modelling the probability of withdrawing (logistic regression)**

| | | Univariate | | Multivariate | |
|---|---|---|---|---|---|
| | n/N (%) | Odds Ratio (95% CI) | p-value | Odds Ratio (95% CI) | p-value |
| **Variables associated with not completing the acute phase** | | | | | |
| Treatment arm (ref placebo) | 32/116 (27.6) | | | | |
| - Sertraline | 32/111 (28.8) | 1.06 (0.60; 1.90) | 0.84 | 1.23 (0.67; 2.25) | 0.51 |
| - Hypericum | 31/113 (27.4) | 0.99 (0.56; 1.77) | 0.98 | 1.04 (0.57; 1.90) | 0.91 |
| Baseline HAM-D | - | 1.00 (0.91; 1.09) | 0.94 | 1.02 (0.93; 1.12) | 0.74 |
| Duration of current depression (years) | - | 1.04 (1.00; 1.08) | 0.04 | 1.05 (1.01; 1.10) | 0.02 |
| Age | - | 0.97 (0.95; 0.99) | 0.002 | 0.97 (0.95; 0.99) | 0.001 |
| Gender (ref female) | 66/224 (29.5) | | | | |
| - Male | 29/116 (25.0) | 0.80 (0.48; 1.33) | 0.38 | 1.27 (0.75; 2.16) | 0.37 |
| Race (ref = white) | 71/257 (27.6) | | | | |
| - Hispanic | 12/27 (44.4) | 2.10 (0.94; 4.70) | 0.07 | 1.79 (0.76; 4.20) | 0.18 |
| - Black | 9/35 (25.7) | 0.91 (0.41; 2.03) | 0.81 | 0.87 (0.38; 2.00) | 0.75 |
| - Other | 3/21 (14.3) | 0.44 (0.13; 1.53) | 0.19 | 0.36 (0.10; 1.29) | 0.12 |
| **Variables associated with not completing the continuation phase given the acute phase was completed** | | | | | |
| Treatment arm (ref placebo) | 57/84 (67.9) | | | | |
| - Sertraline | 48/79 (60.8) | 0.73 (0.39; 1.40) | 0.34 | 0.74 (0.34; 1.61) | 0.47 |
| - Hypericum | 58/82 (70.7) | 1.15 (0.59; 2.22) | 0.69 | 0.97 (0.44; 2.16) | 0.66 |
| Baseline HAM-D | - | 1.05 (0.95; 1.17) | 0.36 | 0.90 (0.78; 1.04) | 0.34 |
| Duration of current depression (years) | - | | | 1.02 (0.94; 1.11) | 0.33 |
| Age | - | 0.99 (0.97; 1.01) | 0.46 | 0.99 (0.96; 1.01) | 0.44 |
| Gender (ref female) | 104/158 (65.8) | | | | |
| - Male | 59/87 (67.8) | 1.09 (0.63; 1.91) | 0.75 | 1.29 (0.65; 2.55) | 0.83 |
| Race (ref = white) | 127/186 (68.3) | | | | |
| - Hispanic | 11/15 (73.3) | 1.28 (0.39; 4.18) | 0.69 | 3.15 (0.79; 12.5) | 0.73 |
| - Black | 15/26 (57.7) | 0.63 (0.27; 1.46) | 0.29 | 0.75 (0.28; 2.04) | 0.27 |
| - Other | 10/18 (55.6) | 0.58 (0.22; 1.55) | 0.28 | 0.52 (0.19; 1.43) | 0.21 |
| Week 8 HAM-D | - | 1.24 (1.16; 1.32) | <0.001 | 1.27 (1.19; 1.36) | <0.001 |

HAM-D: Hamilton depression scale

The primary analysis during the drug development process is usually the traditional ITT analysis, and is the estimand we were interested in. This corresponds to what the National Research Council (2010) called Estimand 1. This estimand gives the difference in outcome at the planned endpoint of the trial for all participants. This is not the most relevant estimand when one wants to determine the efficacy of a new drug. The de jure estimand is more relevant if one is interested in the efficacy of the drug, versus the effectiveness of the treatment policy. In ideal conditions one would like to determine Estimand 3, however in a trial like this with high drop out and possible non-adherence to the protocol this may not be possible. Estimand 6, defined by Mallinckrodt et al. (2012), which

gives the difference in outcome in all randomised participants attributable to the randomised medication, is also a viable secondary objective of the trial.

SAS code was given in the corresponding sections of Chapter 5 and was not repeated in this chapter.

## 6.3 Analysis of HAM-D score in the hypericum study under MCAR assumptions

Although MCAR assumptions were not valid, we analysed the data using an available case analysis and complete case analysis (discussed in Section 3.1) under MCAR assumptions. This is done solely to contrast the naïve and incorrect analysis to the more appropriate analyses done under the MAR and MNAR assumptions.

### 6.3.1 Available case analysis

In the available case analysis missing data were ignored and observed data were summarised at each of the visits using all observed HAM-D scale total scores (Table 6.4 and Figure 6.4). We performed a cross-sectional analysis of the data observed or "available" at each time point. A model that takes the longitudinal nature of the data into account was also fitted and is given in Section 6.4.1. The cross-sectional analysis is provided to give additional information, but we do not imply that the cross-sectional analysis is equivalent to the longitudinal analysis done later. The summaries at all time points of the available data also provide a description of the observed data, without including any modelling assumptions.

Mean HAM-D scores decreased over time. The mean HAM-D scores did not differ between the treatment arms for most of the weeks. The exceptions were Weeks 2 and 7 where a statistically significant difference between the treatment arms existed. Bonferroni pairwise comparisons showed that the sertraline arm had lower mean scores than the other two arms at these visits.

**Table 6.4: Available case analysis of Hamilton depression scale**

| | Sertraline | Hypericum | Placebo | ANOVA p-value |
|---|---|---|---|---|
| **Baseline** | | | | 0.25 |
| N | 111 | 113 | 116 | |
| Mean | 22.5 | 23.1 | 22.7 | |
| SD | 2.5 | 2.7 | 2.7 | |
| Standard error | 0.24 | 0.26 | 0.25 | |
| Median | 22 | 23 | 22 | |
| **Week 1** | | | | 0.33 |
| N | 101 | 101 | 111 | |
| Mean | 19.5 | 20.1 | 18.9 | |
| SD | 5.1 | 5.2 | 4.9 | |
| Standard error | 0.51 | 0.51 | 0.46 | |
| Median | 20 | 20 | 19 | |
| **Week 2** | | | | 0.05[a] |
| N | 90 | 102 | 107 | |
| Mean | 16.5 | 18.6 | 17.7 | |
| SD | 5.5 | 5.7 | 5.7 | |
| Standard error | 0.58 | 0.57 | 0.55 | |
| Median | 17 | 18.5 | 18.0 | |
| **Week 3** | | | | 0.08 |
| N | 90 | 100 | 94 | |
| Mean | 15.5 | 17.9 | 17.1 | |
| SD | 6.3 | 6.4 | 6.5 | |
| Standard error | 0.67 | 0.64 | 0.67 | |
| Median | 16.5 | 19 | 18 | |
| **Week 4** | | | | 0.16 |
| N | 89 | 97 | 99 | |
| Mean | 14.3 | 16.9 | 15.9 | |
| SD | 6.1 | 7.1 | 6.5 | |
| Standard error | 0.65 | 0.72 | 0.65 | |
| Median | 14 | 18 | 16 | |
| **Week 6** | | | | 0.07 |
| N | 82 | 91 | 93 | |
| Mean | 12.8 | 15.6 | 14.9 | |
| SD | 6.6 | 7.1 | 7.0 | |
| Standard error | 0.72 | 0.75 | 0.73 | |
| Median | 13 | 17 | 15 | |
| **Week 7** | | | | 0.01[a] |
| N | 79 | 82 | 84 | |
| Mean | 11.4 | 14.3 | 12.3 | |
| SD | 6.5 | 6.5 | 7.1 | |
| Standard error | 0.73 | 0.72 | 0.77 | |
| Median | 11 | 14 | 11 | |
| **Week 8** | | | | 0.22 |
| N | 79 | 82 | 84 | |
| Mean | 10.6 | 12.9 | 12.0 | |
| SD | 5.9 | 7.1 | 7.5 | |
| Standard error | 0.66 | 0.79 | 0.82 | |
| Median | 10 | 12 | 10 | |
| **Week 10** | | | | 0.21 |
| N | 48 | 35 | 39 | |
| Mean | 8.1 | 9.5 | 7.3 | |
| SD | 4.6 | 5.0 | 4.8 | |
| Standard error | 0.66 | 0.85 | 0.77 | |
| Median | 7 | 9 | 7 | |

**Table 6.4: Available case analysis of Hamilton depression scale**

| | Sertraline | Hypericum | Placebo | ANOVA p-value |
|---|---|---|---|---|
| **Week 14** | | | | 0.09 |
| N | 43 | 33 | 37 | |
| Mean | 7.6 | 9.8 | 7.8 | |
| SD | 5.0 | 5.0 | 5.0 | |
| Standard error | 0.77 | 0.86 | 0.82 | |
| Median | 6 | 8 | 7 | |
| **Week 18** | | | | 0.21 |
| N | 39 | 27 | 32 | |
| Mean | 6.7 | 8.7 | 5.7 | |
| SD | 4.7 | 6.5 | 4.0 | |
| Standard error | 0.76 | 1.24 | 0.71 | |
| Median | 6 | 7 | 6.5 | |
| **Week 22** | | | | 0.34 |
| N | 32 | 25 | 27 | |
| Mean | 7.1 | 8.2 | 7.0 | |
| SD | 6.1 | 4.6 | 6.0 | |
| Standard error | 1.07 | 0.91 | 1.15 | |
| Median | 5 | 7 | 5 | |
| **Week 26** | | | | 0.36 |
| N | 31 | 24 | 27 | |
| Mean | 7.1 | 6.6 | 5.7 | |
| SD | 5.4 | 4.5 | 5.4 | |
| Standard error | 0.97 | 0.93 | 1.04 | |
| Median | 6 | 5.5 | 4 | |

[a]: Bonferroni pairwise comparison: Sertraline arm significantly different from hypericum arm
ANOVA: Analysis of variance

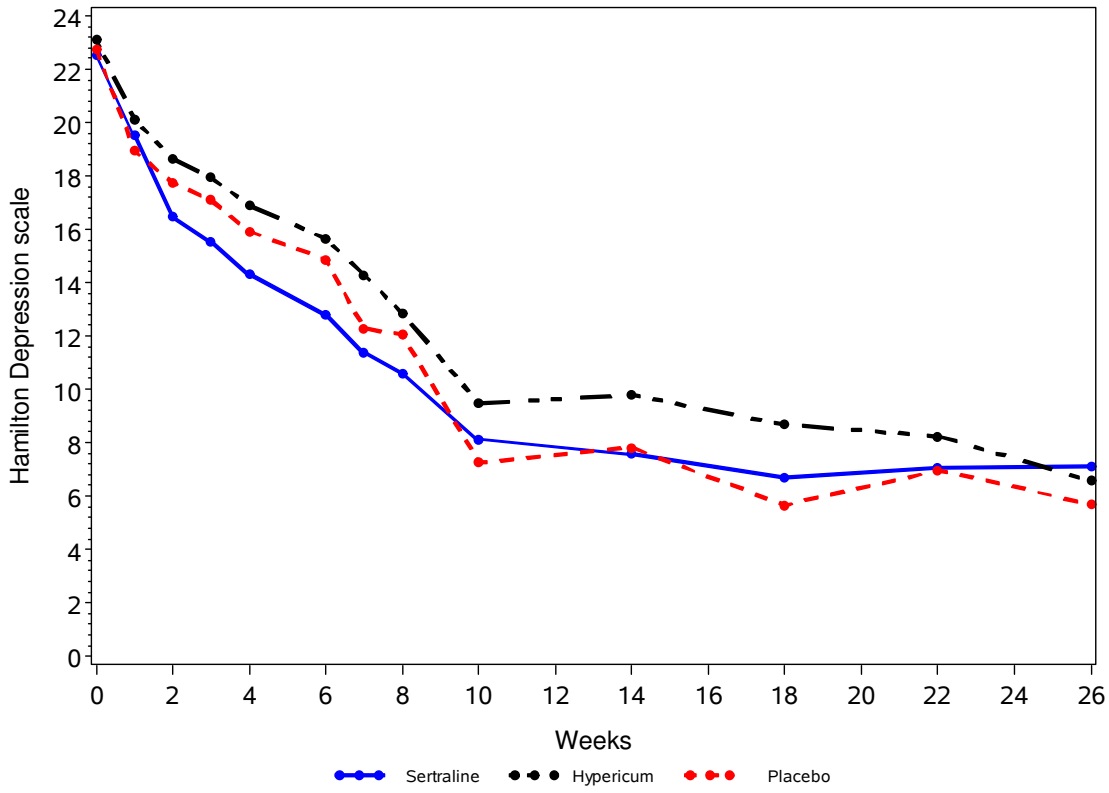



**Figure 6.4: Sample mean of Hamilton depression scale total score, available case analysis**

### 6.3.2 Complete case analysis

In the complete case analysis only participants with no missing data are included and observed HAM-D scores are summarised at all the visits for these participants. In other words, list wise deletion is done. Two separate analyses are done, one including participants who have complete data up to Week 8 and one including participants who have complete data up to Week 26 (Table 6.5, Figure 6.5 and Figure 6.6). The latter analysis includes only a fraction (19.1%) of the enrolled participants and is biased, since it includes only participants deemed responders at Week 8.

The graphical representation of the complete case analysis differs from the available case analysis. As with the available case analysis, we perform a cross-sectional analysis to provide additional information. This cross-sectional analysis can be compared to the cross-sectional analysis done in an available case manner. The intention is not to imply that the cross-sectional analysis fits the same model as the longitudinal analysis also provided. In the cross-sectional analysis of complete data up to Week 8, the treatment arms are only significantly different at baseline. This is in part due to the small number of participants included in this analysis and highlights the effect of analysing only a small, biased subset of the data. Only one, at 8 weeks, of the p-values comparing the means at each time point for the participants with complete data up to Week 26 is significant. In addition to being biased, the complete case analysis is inefficient because it does not utilise all available information.

**Table 6.5: Complete case analysis of Hamilton depression scale**

| | **Complete data up to Week 26** | | | | **Complete data up to Week 8** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Sertraline N = 23** | **Hypericum N = 22** | **Placebo N = 20** | **p-value[$]** | **Sertraline N = 67** | **Hypericum N = 79** | **Placebo N = 68** | **p-value[$]** |
| **Baseline** | | | | 0.54 | | | | 0.05* |
| Mean | 22.2 | 22.9 | 22.9 | | 22.1 | 23.2 | 22.9 | |
| SD | 2.4 | 2.0 | 2.5 | | 2.1 | 2.8 | 2.8 | |
| Standard error | 0.50 | 0.42 | 0.56 | | 0.26 | 0.31 | 0.34 | |
| Median | 22 | 23 | 22.5 | | 21 | 23 | 22 | |
| **Week 1** | | | | 0.20 | | | | 0.83 |
| Mean | 20.0 | 18.3 | 17.9 | | 19.7 | 19.9 | 19.4 | |
| SD | 4.2 | 4.6 | 3.4 | | 4.8 | 5.3 | 4.4 | |
| Standard error | 0.87 | 0.97 | 0.76 | | 0.58 | 0.59 | 0.53 | |
| Median | 20 | 18 | 17 | | 20 | 20 | 20 | |
| **Week 2** | | | | 0.84 | | | | 0.06 |
| Mean | 16.6 | 16.2 | 15.7 | | 16.4 | 18.5 | 18.0 | |
| SD | 4.9 | 5.2 | 5.1 | | 5.0 | 5.9 | 5.0 | |
| Standard error | 1.00 | 1.11 | 1.14 | | 0.61 | 0.67 | 0.61 | |
| Median | 17 | 16 | 16 | | 17 | 18 | 18.5 | |
| **Week 3** | | | | 0.91 | | | | 0.12 |
| Mean | 15.0 | 14.7 | 14.2 | | 15.4 | 17.4 | 17.2 | |
| SD | 5.5 | 6.4 | 6.7 | | 5.6 | 6.4 | 6.5 | |
| Standard error | 1.16 | 1.37 | 1.50 | | 0.69 | 0.72 | 0.78 | |
| Median | 15 | 15 | 15 | | 17 | 18 | 18 | |

**Table 6.5: Complete case analysis of Hamilton depression scale**

| | Complete data up to Week 26 | | | | Complete data up to Week 8 | | | |
|---|---|---|---|---|---|---|---|---|
| | Sertraline N = 23 | Hypericum N = 22 | Placebo N = 20 | p-value[$] | Sertraline N = 67 | Hypericum N = 79 | Placebo N = 68 | p-value[$] |
| **Week 4** | | | | 0.22 | | | | 0.09 |
| Mean | 14.6 | 12.3 | 12.2 | | 14.1 | 16.5 | 15.1 | |
| SD | 4.7 | 5.9 | 5.0 | | 5.9 | 7.0 | 6.4 | |
| Standard error | 0.98 | 1.26 | 1.11 | | 0.72 | 0.79 | 0.78 | |
| Median | 14 | 11.5 | 11 | | 14 | 18 | 15 | |
| **Week 6** | | | | 0.45 | | | | 0.12 |
| Mean | 12.4 | 11.5 | 10.2 | | 13.2 | 15.4 | 14.8 | |
| SD | 5.6 | 6.4 | 5.3 | | 5.6 | 7.0 | 6.6 | |
| Standard error | 1.17 | 1.35 | 1.19 | | 0.69 | 0.79 | 0.80 | |
| Median | 12 | 12 | 9.5 | | 13 | 17 | 15 | |
| **Week 7** | | | | 0.19 | | | | 0.13 |
| Mean | 10.7 | 10.0 | 8.2 | | 11.9 | 14.1 | 12.8 | |
| SD | 5.6 | 4.0 | 3.6 | | 6.0 | 6.5 | 7.1 | |
| Standard error | 1.17 | 0.85 | 0.82 | | 0.73 | 0.73 | 0.86 | |
| Median | 10 | 10 | 8 | | 11 | 14 | 11 | |
| **Week 8** | | | | 0.04 | | | | 0.27 |
| Mean | 8.6 | 7.0 | 6.2 | | 11.1 | 12.7 | 12.7 | |
| SD | 2.5 | 3.6 | 3.2 | | 5.2 | 7.2 | 7.6 | |
| Standard error | 0.52 | 0.77 | 0.72 | | 0.64 | 0.81 | 0.92 | |
| Median | 9 | 7.5 | 6.5 | | 10 | 12 | 11.5 | |
| **Week 10** | | | | 0.35 | | | | 0.35 |
| N | | | | | 42 | 35 | 30 | |
| Mean | 8.4 | 9.9 | 7.9 | | 8.2 | 9.5 | 7.9 | |
| SD | 4.0 | 4.6 | 5.1 | | 4.6 | 5.0 | 4.8 | |
| Standard error | 0.83 | 0.98 | 1.14 | | 0.71 | 0.85 | 0.87 | |
| Median | 8 | 9.5 | 7.5 | | 7 | 9 | 8 | |
| **Week 14** | | | | 0.77 | | | | 0.07 |
| N | | | | | 38 | 33 | 30 | |
| Mean | 7.4 | 8.1 | 8.2 | | 7.2 | 9.8 | 8.1 | |
| SD | 4.4 | 3.6 | 3.8 | | 4.4 | 5.0 | 4.8 | |
| Standard error | 0.92 | 0.77 | 0.85 | | 0.71 | 0.86 | 0.88 | |
| Median | 7 | 8 | 9 | | 7 | 8 | 7.5 | |
| **Week 18** | | | | 0.53 | | | | 0.12 |
| N | | | | | 35 | 27 | 26 | |
| Mean | 7.9 | 8.2 | 6.5 | | 7.0 | 8.7 | 5.7 | |
| SD | 5.1 | 6.5 | 3.6 | | 4.9 | 6.5 | 4.0 | |
| Standard error | 1.07 | 1.39 | 0.81 | | 0.82 | 1.24 | 0.79 | |
| Median | 7 | 5 | 7 | | 6 | 7 | 6.5 | |
| **Week 22** | | | | 0.95 | | | | 0.87 |
| N | | | | | 28 | 25 | 25 | |
| Mean | 8.7 | 8.2 | 8.6 | | 7.9 | 8.2 | 7.4 | |
| SD | 5.7 | 4.3 | 6.1 | | 6.0 | 4.6 | 6.0 | |
| Standard error | 1.20 | 0.91 | 1.36 | | 1.13 | 0.91 | 1.20 | |
| Median | 8 | 7 | 7.5 | | 7 | 7 | 6 | |
| **Week 26** | | | | 0.37 | | | | 0.43 |
| N | | | | | 27 | 24 | 24 | |
| Mean | 8.7 | 6.9 | 6.7 | | 7.8 | 6.6 | 5.9 | |
| SD | 5.2 | 4.5 | 5.6 | | 5.5 | 4.5 | 5.5 | |
| Standard error | 1.08 | 0.96 | 1.26 | | 1.05 | 0.93 | 1.12 | |
| Median | 8 | 5.5 | 5 | | 7 | 5.5 | 4 | |

* Bonferroni pairwise comparison: Sertraline arm significantly different from hypericum arm

$: Analysis of variance (ANOVA)

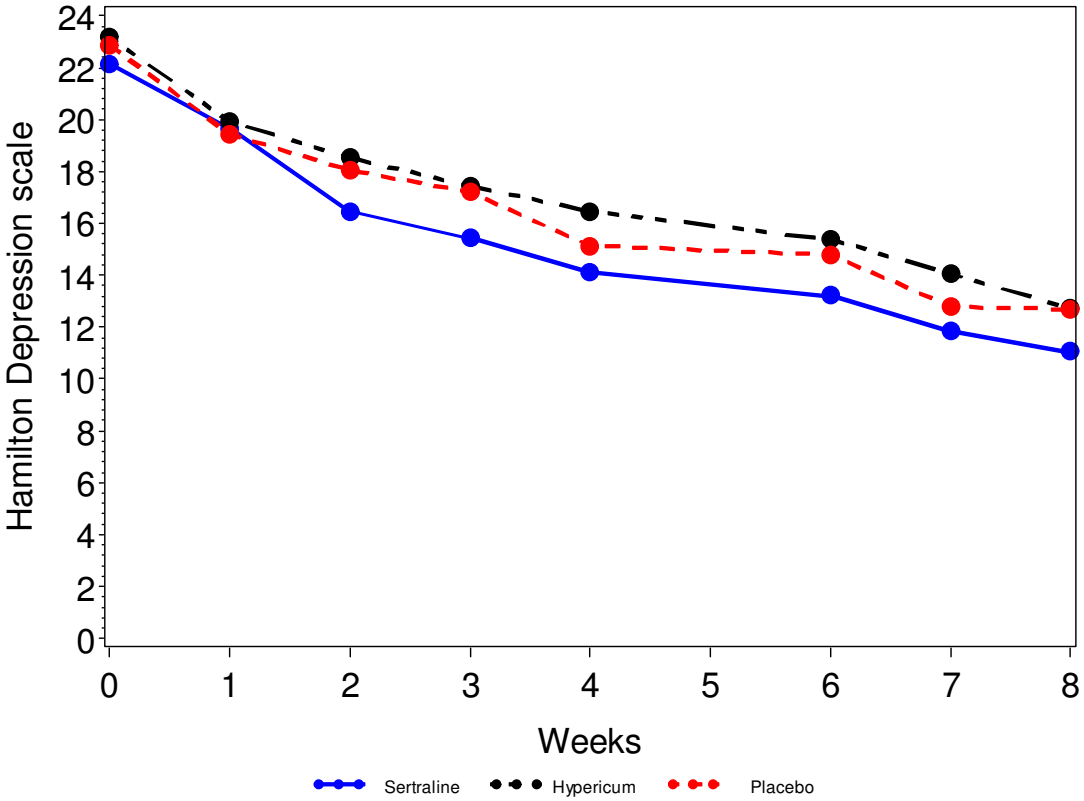



**Figure 6.5: Sample mean of Hamilton depression scale, complete case analysis for participants who have complete data up to Week 8**

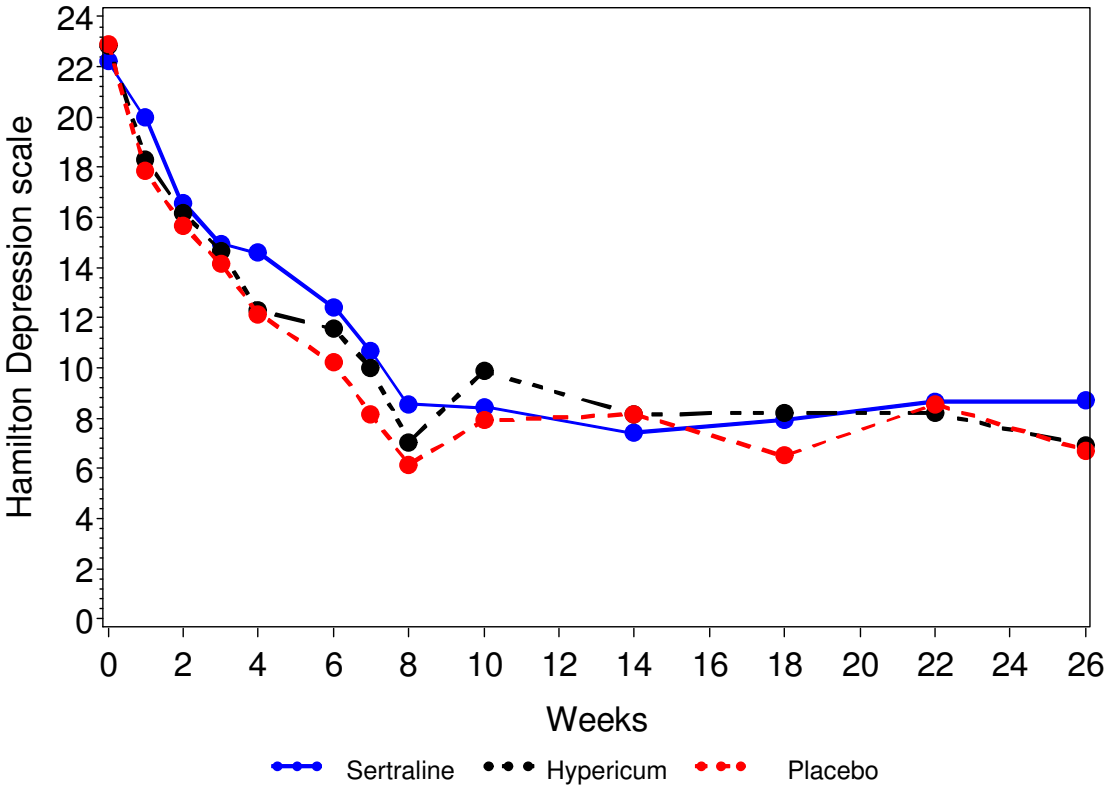



**Figure 6.6: Sample mean of Hamilton depression scale, complete case analysis for participants who have complete data up to Week 26**

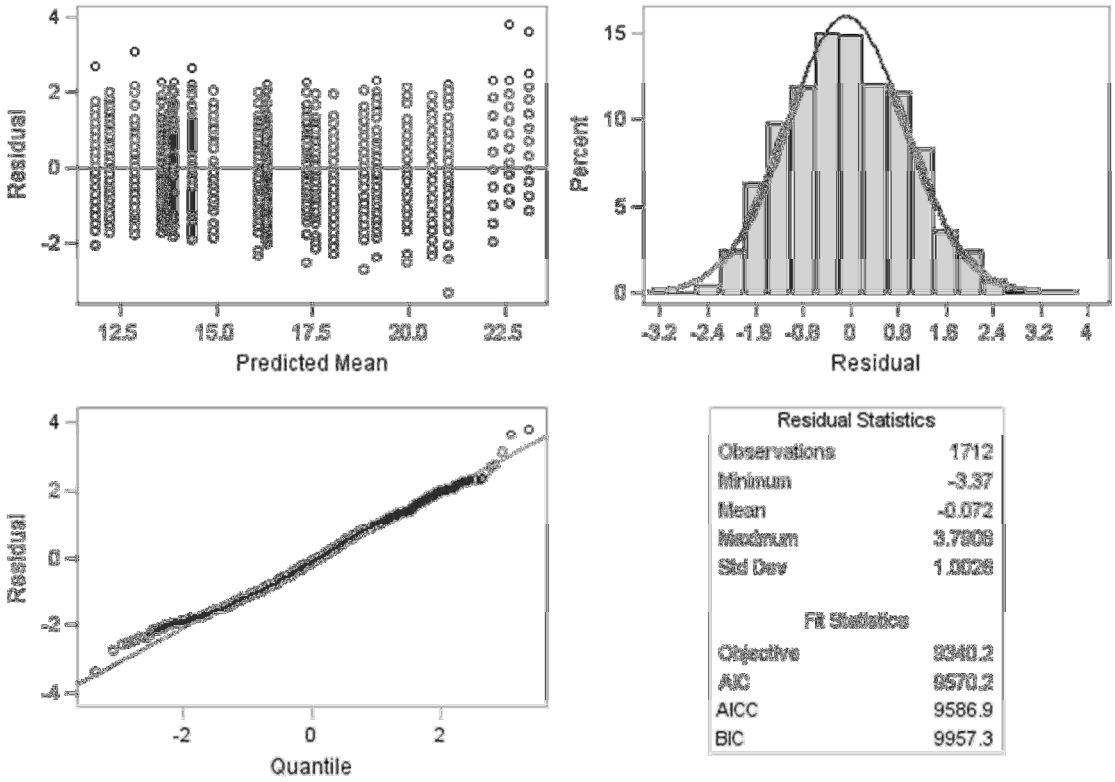



**Figure 6.7: Studentised residuals for the model fitted using SAS procedure MIXED for the complete case analysis for participant with complete data up to Week 8**

The studentised residuals follows a normal distribution in the complete case analysis for participants who had complete data up to Week 8 (Figure 6.7) and up to Week 26 (Figure 6.8).

**Table 6.6: Complete case analysis using SAS procedure MIXED**

| | Complete data up to Week 8 | | | Complete data up to Week 26 | | |
|---|---|---|---|---|---|---|
| **Effect** | **Estimate** | **Standard error** | **p-value** | **Estimate** | **Standard error** | **p-value** |
| Intercept | 22.90 | 0.33 | | 22.98 | 0.60 | |
| Week | -1.07 | 0.11 | <0.001 | -1.99 | 0.08 | <0.001 |
| Week squared | | | | 0.05 | 0.01 | <0.001 |
| Sertraline (ref: placebo) | -0.32 | 0.42 | 0.45 | -2.10 | 0.75 | <0.001 |
| Hypericum (ref: placebo) | 0.33 | 0.47 | 0.48 | -0.20 | 0.75 | 0.79 |
| Interaction week and treatment (sertraline compared to placebo) | -0.26 | 0.13 | 0.05 | 0.11 | 0.06 | 0.07 |
| Interaction week and treatment (hypericum compared to placebo) | -0.11 | 0.15 | 0.44 | 0.08 | 0.06 | 0.24 |

The model for the complete cases was fitted by REML using procedure MIXED in SAS and assuming an unstructured covariance matrix while using the Kenward-Roger method for the degrees of freedom (Kenward & Roger, 1997). A different covariance matrix was estimated for each treatment group.

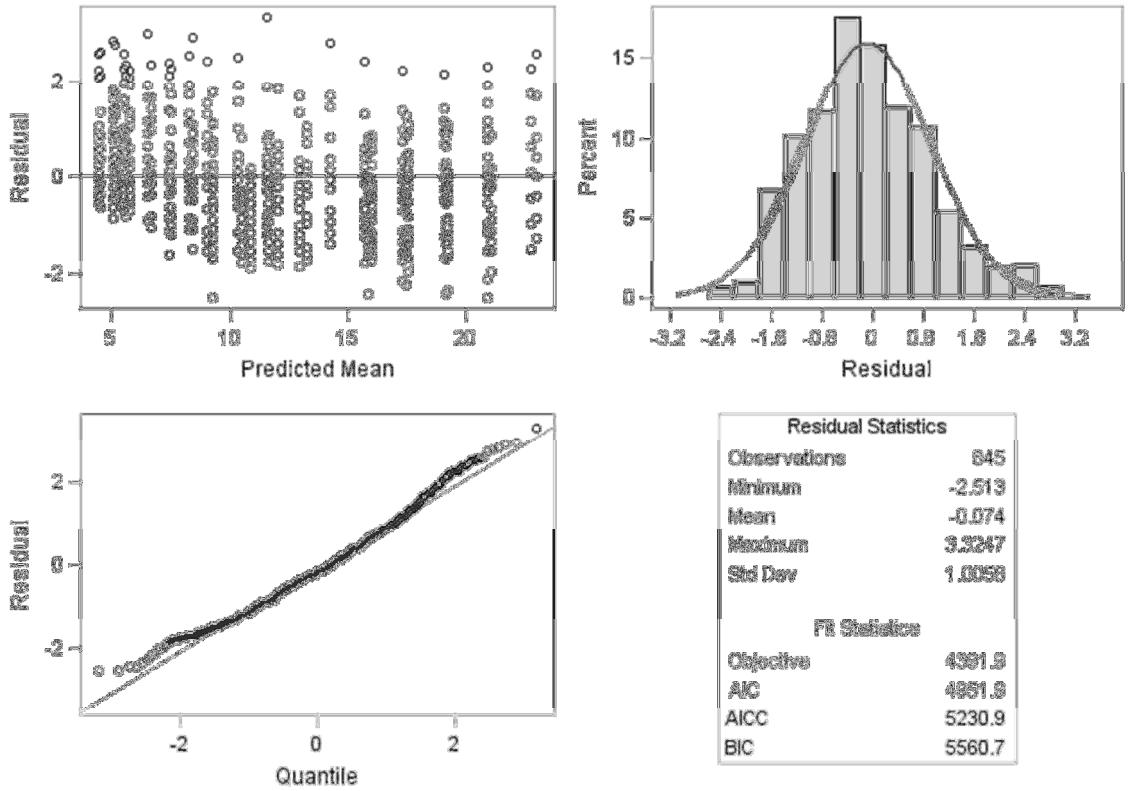



**Figure 6.8: Studentised residuals for the model fitted using SAS procedure MIXED for the complete case analysis Hamilton depression scale for participants with complete data up to Week 26**

We conclude that the decrease in mean HAM-D score over time was significantly different in the sertraline arm compared to the placebo arm for both the analysis up to Week 8 and Week 26 (where the p-value was small, but not <0.05). The interaction between week and the hypericum arm compared to the placebo arm was not significant for either data set (Table 6.6).

### 6.3.3 Conclusions under MCAR assumptions

The MCAR analysis was done for completeness, but since we have evidence that data were not MCAR, the MCAR analysis was not valid. Fitting a mixed model for repeated measures in the complete case analysis we conclude that there is a significant interaction between week and sertraline arm compared to the placebo arm.

These analyses do not lead to the estimand of interest, namely the effect of treatment assignment at the end of the study period according to the ITT principle (Estimand 1). Using these methods, it would also not be possible to estimate Estimands 3 or 6, because analysing all available data or analysing only the participants who completed the study does not lead to estimates of those estimands either.

## 6.4 Analysis under MAR assumptions

MAR assumptions were discussed in Section 2.2.2. Under MAR assumption, we analysed the data using direct likelihood-based approaches (discussed in Section 3.5), multiple imputation (discussed in Section 3.3), direct Bayesian approaches (discussed in Section 3.7) and inverse probability weighting (discussed in Section 3.5).

A model with a linear term for week fitted the data well up to Week 8 and all models up to Week 8 were fitted using the following linear model:

Model up to Week 8: $Y_{ij} = \beta_0 + \beta_1 \mathrm{S}_i + \beta_2 \mathrm{H}_i + \beta_3\, \mathrm{t}_{ij} + (\beta_4 \mathrm{S}_i + \beta_5 \mathrm{H}_i)\, \mathrm{t}_{ij} + \varepsilon_{\mathrm{ij}}$

The statistical notation is given in Section 6.1.4. The profiles from Week 1 to Week 26 did not follow a linear trend over time and for this data we fitted a model including a quadratic term for week.

Model up to Week 26: $Y_{ij} = \beta_0 + \beta_1 \mathrm{S}_i + \beta_2 \mathrm{H}_i + \beta_3\, \mathrm{t}_{ij} + (\beta_4 \mathrm{S}_i + \beta_5 \mathrm{H}_i)\, \mathrm{t}_{ij} + \beta_6 t_{ij}^2 + \varepsilon_{\mathrm{ij}}$

In Figure 6.4, one can see that the first part of the observed profiles follow a linear trend in time, but from Week 8 onwards, the profiles follow a quadratic trend in time.

### 6.4.1 Direct likelihood-based approaches

As mentioned in Section 5.4.1 direct likelihood-based approaches are valid under the MAR assumption and the separability condition. In this analysis there is no interest in the missing data mechanism. Procedure MIXED in SAS was used. The Kenward-Roger correction was applied when calculating degrees of freedom (Kenward & Roger, 1997).

The choice of covariance structure is very important. An unstructured covariance matrix is the most flexible choice, since each variance-covariance component is estimated from the data, therefore making no assumptions that are not supported by the observed data. This covariance matrix has more parameters that need to be estimated than other structures, and is therefore less efficient. We have used the unstructured covariance matrix in this analysis, assuming a different covariance matrix for each of the treatment arms.

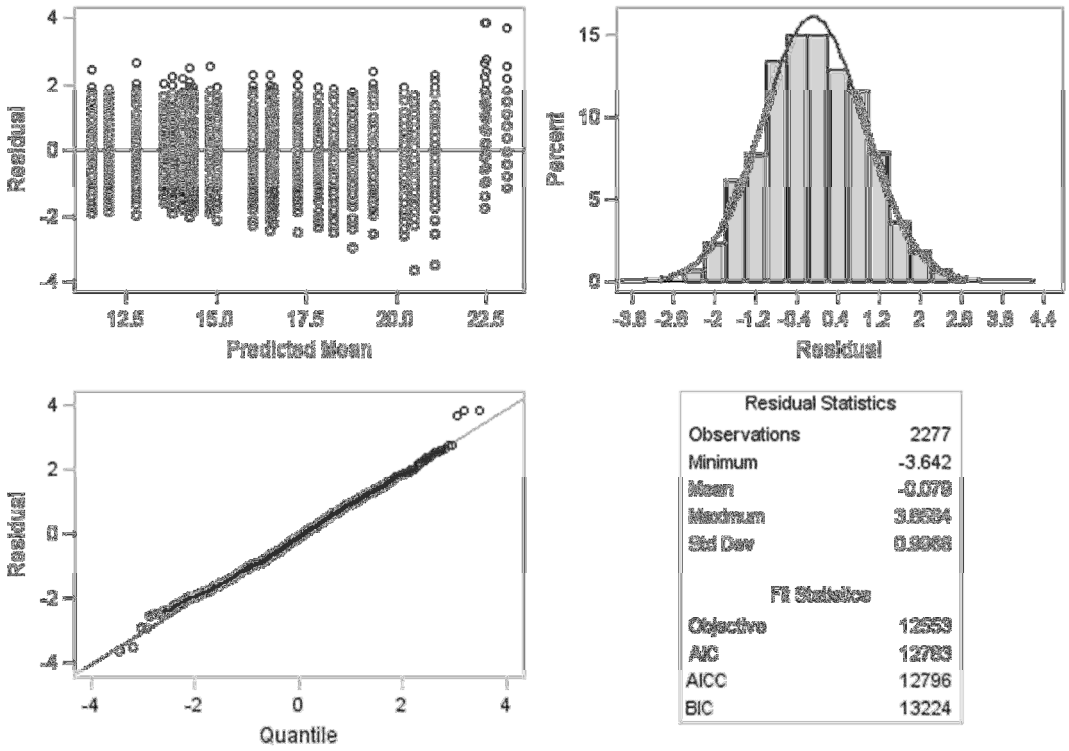



**Figure 6.9: Studentised residuals for the model fitted using procedure MIXED in SAS for the direct likelihood-based analysis using the HAM-D scores up to Week 8**

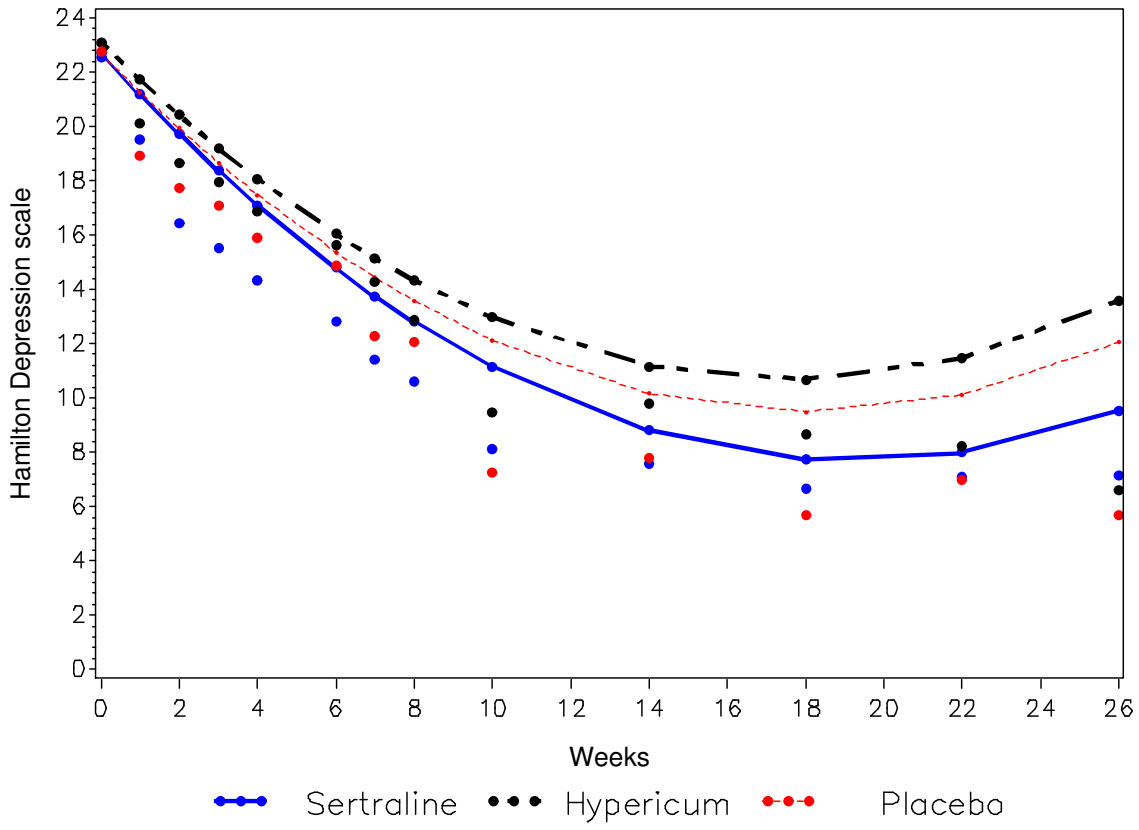



**Figure 6.10: Mean Hamilton depression scale total scores, direct likelihood-based approach (mixed model) analysis up to 26 weeks**

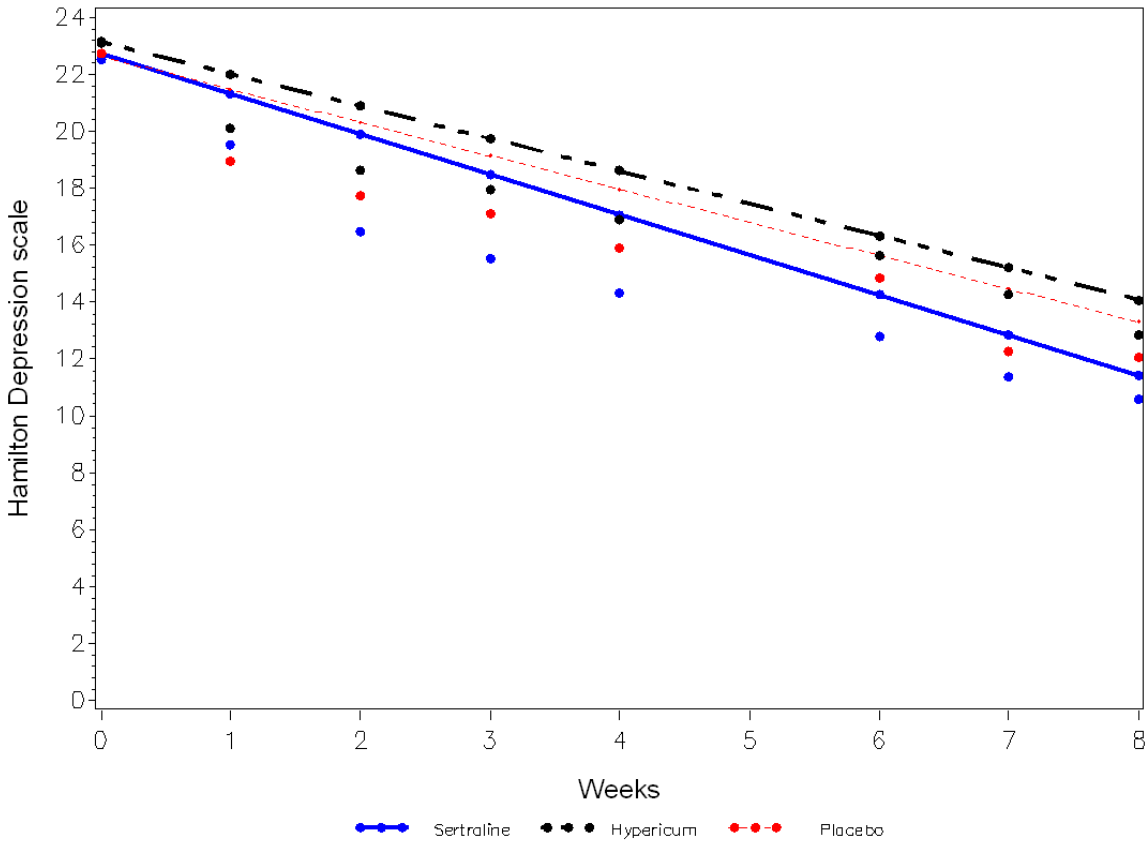



**Figure 6.11: Mean Hamilton depression scale total scores, direct likelihood-based approach (mixed model) analysis up to 8 weeks**

The studentised residuals when the direct likelihood model is fitted followed a normal distribution, (Figure 6.9). The model fitted up to Week 26 is displayed in Figure 6.10 and the model up to Week 8 in Figure 6.11.

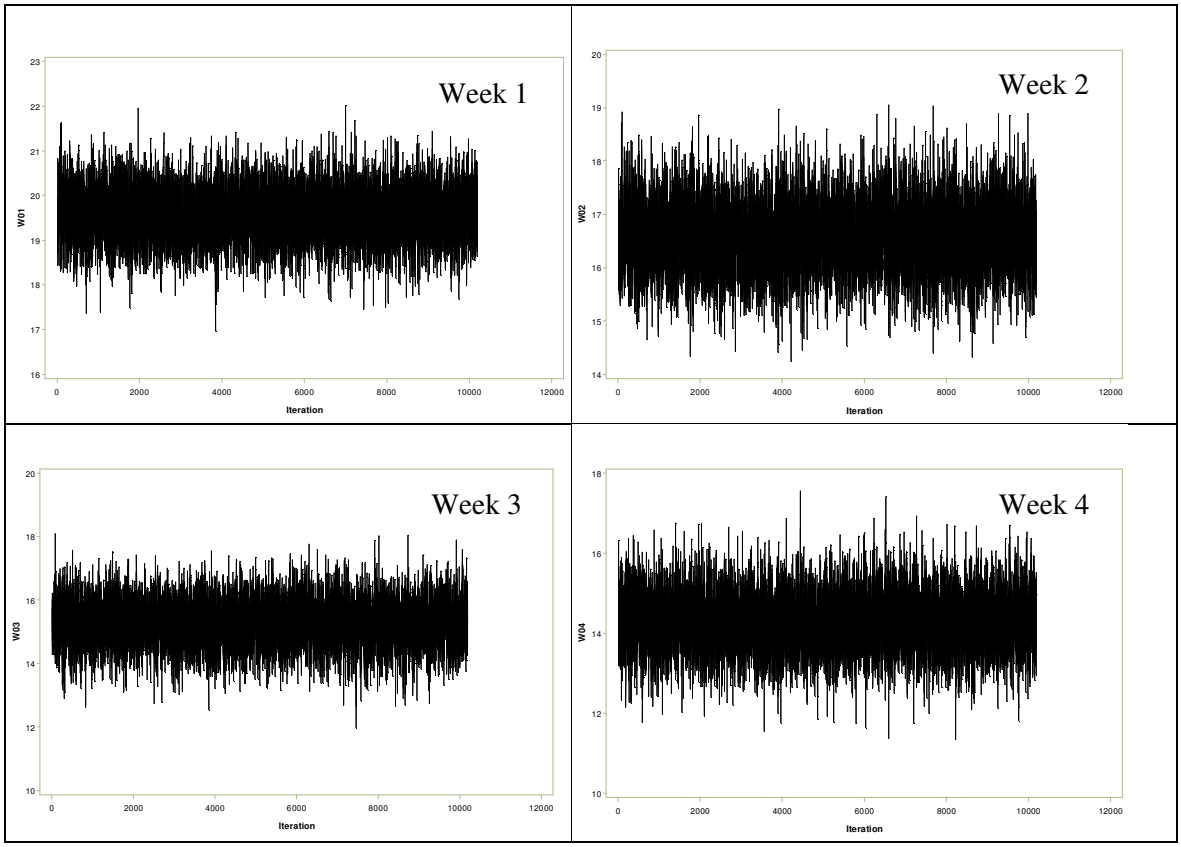
**Table 6.7: Hamilton depression scale. Direct likelihood-based parameter estimates and standard errors when using procedure MIXED in SAS**

| | Data up to Week 26 | | | Data up to Week 8 | | |
|---|---|---|---|---|---|---|
| Effect | **Estimate** | **Standard error** | **p-value** | **Estimate** | **Standard error** | **p-value** |
| Intercept | 22.68 | 0.25 | | 22.64 | 0.25 | |
| Sertraline (ref: placebo) | 0.002 | 0.34 | 0.99 | 0.08 | 0.34 | 0.82 |
| Hypericum (ref: placebo) | 0.41 | 0.38 | 0.28 | 0.50 | 0.36 | 0.17 |
| Week | -1.46 | 0.06 | <0.001 | -1.17 | 0.10 | <0.001 |
| Week squared | 0.04 | 0.002 | <0.001 | | | |
| Interaction week and treatment (sertraline compared to placebo) | -0.10 | 0.05 | 0.07 | -0.24 | 0.12 | 0.05 |
| Interaction week and treatment (hypericum compared to placebo) | -0.04 | 0.06 | 0.48 | 0.03 | 0.13 | 0.79 |

For both the analysis using data up to week 8 and using data up to Week 26, the mean decrease over time was larger in the sertraline arm than in the placebo arm. However, for the hypericum arm the mean decrease over time was not significantly different from the placebo arm.

### 6.4.2 Multiple imputation

The SAS procedure MI was used to generate 100 different imputed data sets and procedure MIANALYZE was used to combine the results of the 100 analyses on these data sets. In the imputer's model the HAM-D score at each of the weekly visits was imputed and baseline HAM-D score, age, gender, race, BDI, duration of depression, GAF, CGI-S and CGI-I scales were included as covariates in this model. These variables were included because it was thought that they could be related to severity of depression over time or to the probability of missingness. Separate imputations were done for each of the three treatment arms. All variables that were to be included in the analyst's model were included in the imputer's model, therefore the imputer's model and the analyst's model were congenial. Since the missing data pattern in the HAM-D scores was not monotone, the MCMC imputation method was used on the full data set. The implementation of multiple imputation in SAS is discussed in more detail in Section 5.4.2.



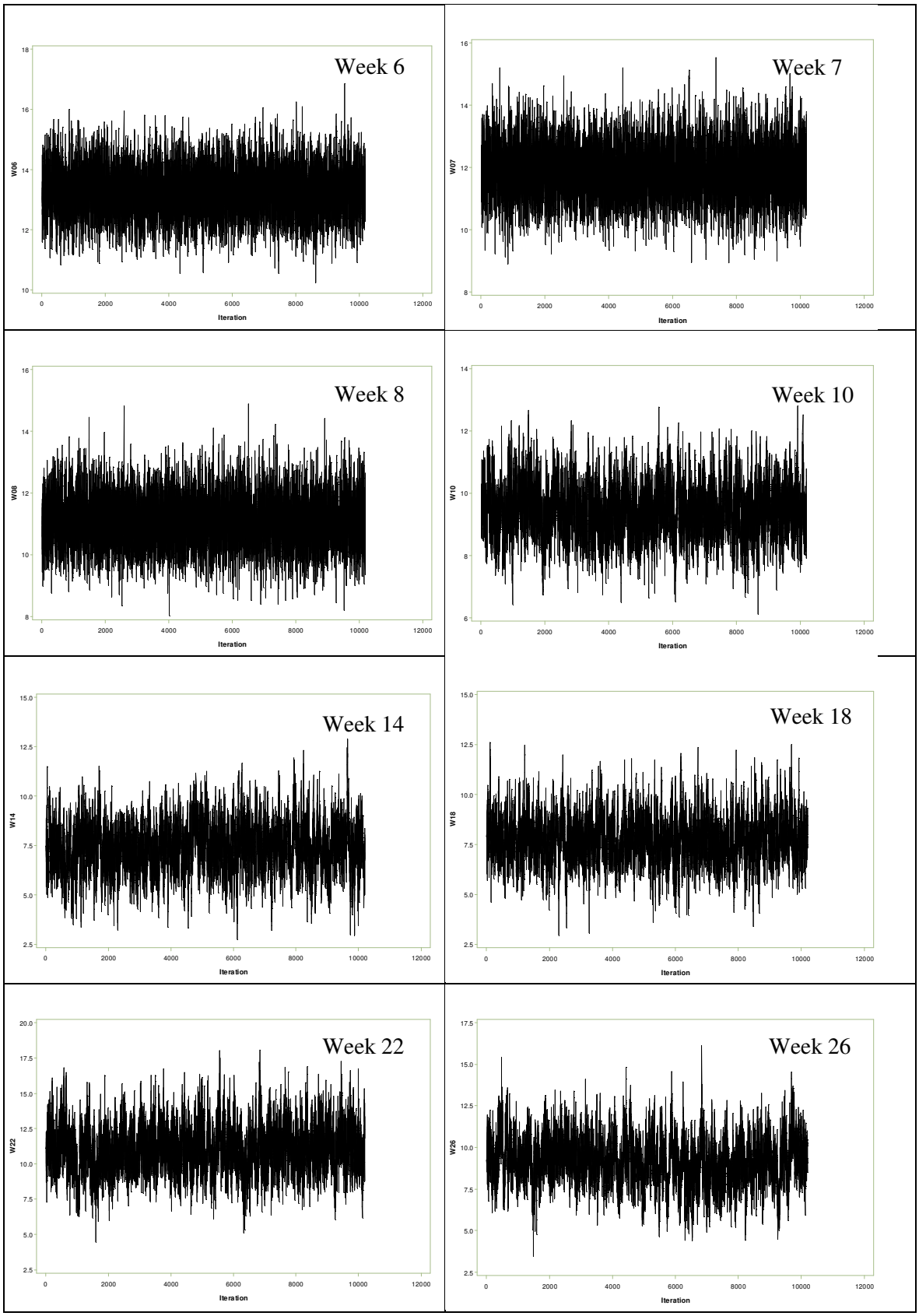
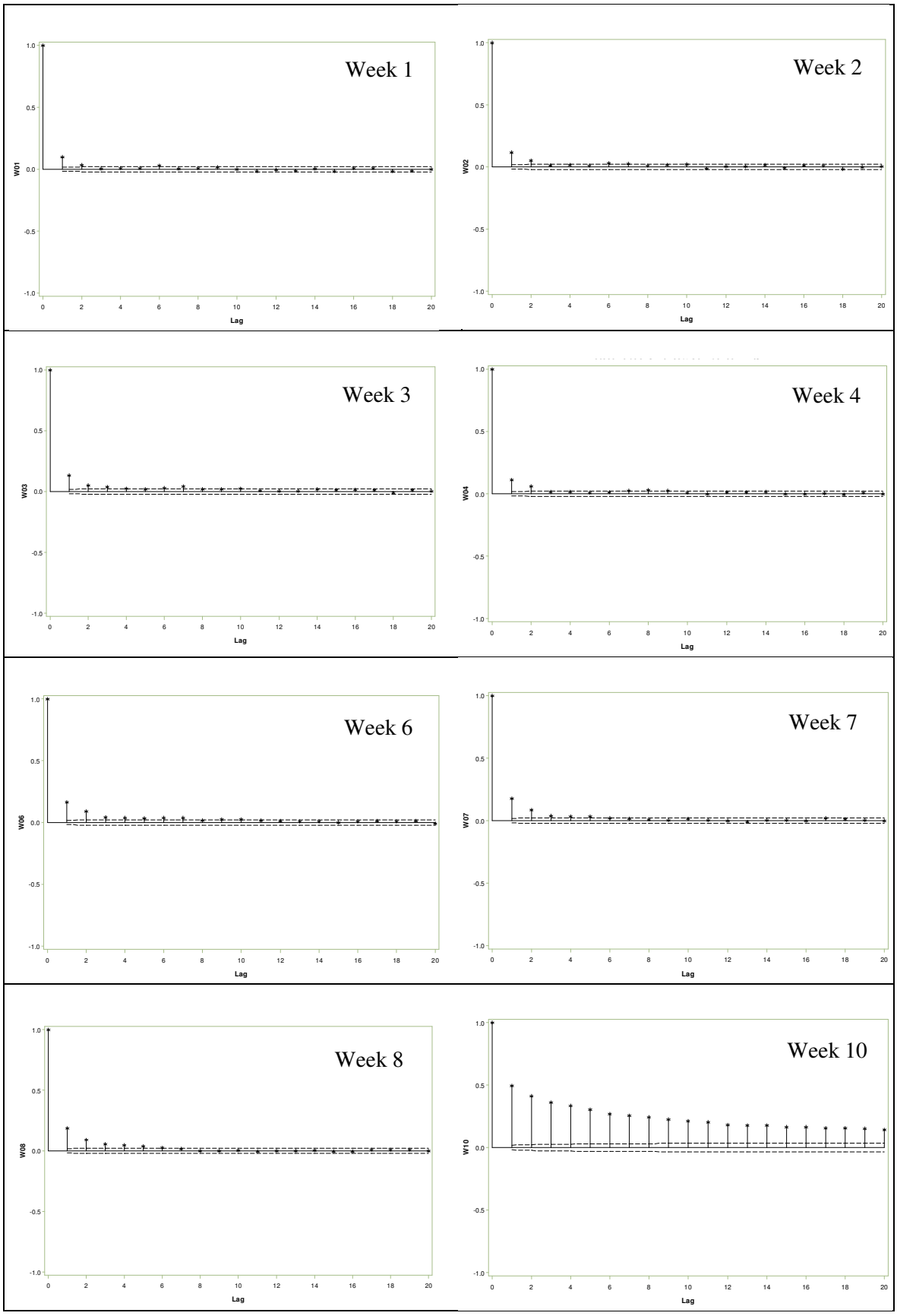



**Figure 6.12: Trace plot for the mean Hamilton depression scale from each week, sertraline arm**


Week 1
W01
Week 2
W02
Week 3
W03
Week 4
W04
Week 6
W06
Week 7
W07
Week 8
W08
Week 10
W10
Lag

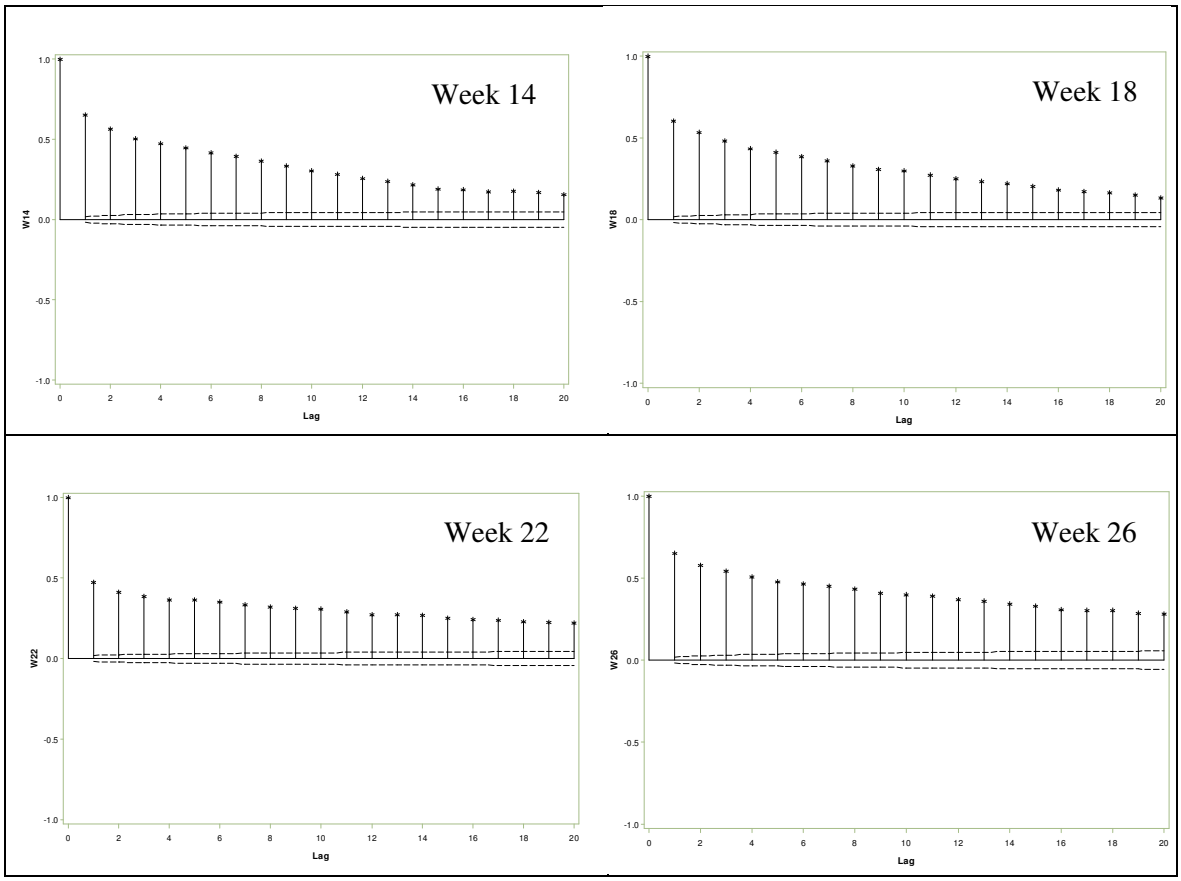



**Figure 6.13: Autocorrelation plot with 95% confidence interval for the mean Hamilton depression scale, sertraline arm**

The trace plots and autocorrelation plots are used to check convergence for the single chain. The trace plots and autocorrelation plots were similar in all three treatment arms, and only the plots for the sertraline arm are given in Figure 6.12 and Figure 6.13. Iterations during the burn-in period are indicated with negative iteration numbers. There is no pattern in the trace plots from Weeks 1 to 8. However from Week 10 onwards a pattern is present. The autocorrelation plots show large positive autocorrelations from Week 10 onwards. The models do not seem to reach a stationary distribution and autocorrelation exist between the imputations. This means that the results from Week 10 onwards should be regarded with some suspicion. This is likely due to the fact that more data are imputed than present. This example illustrates the inadequacy of multiple imputation when a substantial amount of data are incomplete.

Table 6.8 indicates that the major source of variance from Week 1 to Week 8 was the within imputation variance. The between imputation variance added more to the overall variance at the later weeks, where more data were missing than at the earlier weeks. From Week 10 onwards the contribution of the between imputation variation to the total variance is as large or larger as that of the within imputation variance. In the case of the hypericum arm and to a lesser extent the placebo

arm, the contribution of the between imputation variance is much higher than the contribution of the within imputation variance at the later weeks.

**Table 6.8: Multiple imputation variance information**

| | Sertraline | | | | Hypericum | | | | Placebo | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Week** | **Between** | **Within** | **Total** | **df** | **Between** | **Within** | **Total** | **df** | **Between** | **Within** | **Total** | **df** |
| **1** | 0.036 | 0.252 | 0.288 | 93.1 | 0.020 | 0.233 | 0.252 | 100.9 | 0.008 | 0.204 | 0.2119 | 108.7 |
| **2** | 0.060 | 0.293 | 0.353 | 87.2 | 0.023 | 0.289 | 0.312 | 101.3 | 0.011 | 0.275 | 0.286 | 108.6 |
| **3** | 0.070 | 0.375 | 0.446 | 88.7 | 0.038 | 0.379 | 0.418 | 99.1 | 0.045 | 0.377 | 0.422 | 99.7 |
| **4** | 0.084 | 0.377 | 0.462 | 85.7 | 0.042 | 0.446 | 0.488 | 99.8 | 0.029 | 0.376 | 0.406 | 104.4 |
| **6** | 0.119 | 0.420 | 0.540 | 80.7 | 0.077 | 0.469 | 0.547 | 92.6 | 0.080 | 0.446 | 0.526 | 93.6 |
| **7** | 0.133 | 0.468 | 0.602 | 80.6 | 0.077 | 0.412 | 0.490 | 90.5 | 0.094 | 0.446 | 0.542 | 90.5 |
| **8** | 0.132 | 0.444 | 0.577 | 79.6 | 0.114 | 0.487 | 0.603 | 86.2 | 0.134 | 0.511 | 0.647 | 85.9 |
| **10** | 0.369 | 0.274 | 0.647 | 39.7 | 1.460 | 0.519 | 1.993 | 24.7 | 1.873 | 0.576 | 2.468 | 22.8 |
| **14** | 0.860 | 0.425 | 1.294 | 30.6 | 2.711 | 0.442 | 3.181 | 13.7 | 2.032 | 0.624 | 2.677 | 22.8 |
| **18** | 0.646 | 0.391 | 1.044 | 34.9 | 8.100 | 1.400 | 9.582 | 14.4 | 3.118 | 0.291 | 3.440 | 8.8 |
| **22** | 1.088 | 1.126 | 2.226 | 48.2 | 11.395 | 3.130 | 14.639 | 20.5 | 8.973 | 0.667 | 9.730 | 7.3 |
| **26** | 1.282 | 0.541 | 1.835 | 27.4 | 10.308 | 3.277 | 13.688 | 22.8 | 5.056 | 0.862 | 5.969 | 14.6 |

df = degrees of freedom

**Table 6.9: Relative increase in variance and fraction missing information in procedure MI**

| | Sertraline | | Hypericum | | Placebo | |
|---|---|---|---|---|---|---|
| **Week** | **Relative increase in variance** | **Fraction missing information** | **Relative increase in variance** | **Fraction missing information** | **Relative increase in variance** | **Fraction missing information** |
| **1** | 0.143240 | 0.125570 | 0.084527 | 0.078052 | 0.038938 | 0.037506 |
| **2** | 0.206724 | 0.171801 | 0.080002 | 0.074178 | 0.039641 | 0.038158 |
| **3** | 0.189869 | 0.160004 | 0.101472 | 0.092280 | 0.120578 | 0.107812 |
| **4** | 0.224152 | 0.183661 | 0.094138 | 0.086176 | 0.077247 | 0.071804 |
| **6** | 0.285394 | 0.222802 | 0.166407 | 0.143019 | 0.181048 | 0.153696 |
| **7** | 0.286370 | 0.223396 | 0.188143 | 0.158776 | 0.213627 | 0.176539 |
| **8** | 0.299118 | 0.231070 | 0.236749 | 0.192026 | 0.265517 | 0.210511 |
| **10** | 1.360650 | 0.579203 | 2.843796 | 0.742670 | 3.283540 | 0.769271 |
| **14** | 2.044079 | 0.674446 | 6.190996 | 0.862974 | 3.287605 | 0.769491 |
| **18** | 1.669447 | 0.628316 | 5.842185 | 0.855954 | 10.836795 | 0.916913 |
| **22** | 0.975947 | 0.496389 | 3.676539 | 0.788787 | 13.591876 | 0.932639 |
| **26** | 2.393768 | 0.708260 | 3.177220 | 0.763356 | 5.925076 | 0.857686 |

The relative increase in variance showed that the uncertainty caused by the missing data increased the variance more at later visits (Table 6.9). Week 26 was an exception, and the relative increase in variance was smaller than at preceding visits. In the sertraline arm the relative increase in variance reached a high of 2.4 at Week 26, in the hypericum arm it reached a high of 6.2 at Week 14 and in the placebo arm the ratio reached a high of 13.6 at Week 22. The fraction of missing information was above 0.5 at all but one visit post Week 10. The fraction of missing information in the placebo arm was particularly large, in excess of 0.75 after Week 8.

Even though our interest lies in fitting a longitudinal model, we also provide the cross-sectional summaries of mean HAM-D scores at each visit (Table 6.10).

**Table 6.10: Cross sectional summaries of multiple imputation of Hamilton depression scale; mean (standard error)**

| | Sertraline | Hypericum | Placebo | p-value |
|---|---|---|---|---|
| Week 1 | 19.6 (0.53) | 20.1 (0.50) | 18.9 (0.46) | 0.31 |
| Week 2 | 16.6 (0.59) | 18.4 (0.56) | 17.7 (0.53) | 0.18 |
| Week 3 | 15.3 (0.65) | 17.5 (0.64) | 17.3 (0.65) | 0.03 |
| Week 4 | 14.3 (0.67) | 16.7 (0.70) | 15.8 (0.63) | 0.14 |
| Week 6 | 13.3 (0.72) | 15.8 (0.73) | 14.9 (0.72) | 0.13 |
| Week 7 | 11.9 (0.71) | 14.5 (0.69) | 12.8 (0.73) | 0.37 |
| Week 8 | 11.0 (0.68) | 13.1 (0.75) | 12.7 (0.79) | 0.12 |
| Week 10 | 9.5 (0.78) | 13.7 (1.35) | 13.0 (1.54) | 0.05 |
| Week 14 | 7.9 (0.98) | 11.1 (1.61) | 12.9 (1.57) | 0.01 |
| Week 18 | 8.1 (0.98) | 14.1 (2.29) | 7.2 (1.72) | 0.65 |
| Week 22 | 11.0 (1.27) | 10.4 (2.31) | 9.8 (2.73) | 0.69 |
| Week 26 | 9.5 (1.21) | 15.9 (2.20) | 12.0 (2.28) | 0.35 |

Comparing the means and standard errors of the imputed HAM-D scores (Table 6.10) and the results from the available case analysis (Table 6.4) shows that the standard errors and means were higher with multiple imputation, especially after Week 10. The means in the sertraline arm were lower than the means in the other two treatment arms at all time points from Week 2 to Week 14. P-values at Weeks 3, 10 and 14 were significant. The multiple imputation chain did not seem to converge from Week 10 onwards and the results from Week 10 onwards should be interpreted with care. If future missing measurements could not be predicted from the past observed data combined with the covariates, it was unlikely that the multiple imputations are correct.

According to the study design participants who did not respond to treatment by Week 8 were discontinued from the study. Therefore all data in Figure 6.14 to Figure 6.16 from Week 10 onwards are imputed. In the placebo arm data after Week 8 seem to have much less variation than data before Week 8. Values seem to decline from Week 8 to Week 22, but increase again at Week 26. At Week 22, all values are relatively similar (Figure 6.14). In the sertraline arm HAM-D scores seem to decrease from Week 8 to Week 14 and increase again to Weeks 22 and 26. HAM-D scores at Week 26 were similar to HAM-D scores at Week 8 (Figure 6.15). In the hypericum arm there was no specific pattern in the imputed data (Figure 6.16). The pattern of imputed values was different in each of the three treatment arms.

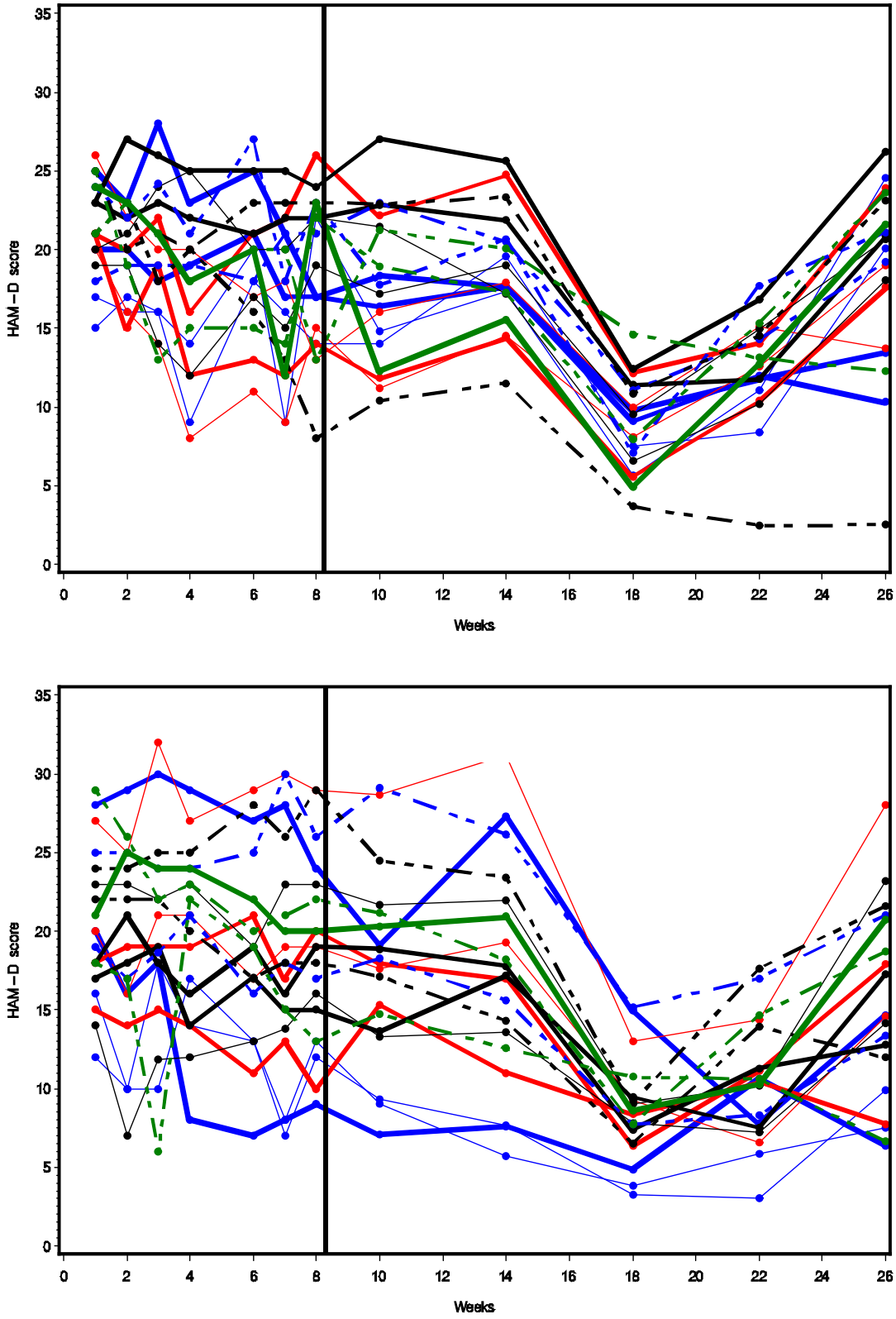



**Figure 6.14: Imputed Hamilton depression scale total scores in the placebo arm for participants who completed the acute phase of treatment (up to Week 8) and did not continue to the continuation phase due to protocol defined non-response at Week 8.**
The vertical black line at 8 weeks indicates the cut-off between observed and imputed data. Each line represents a participant.

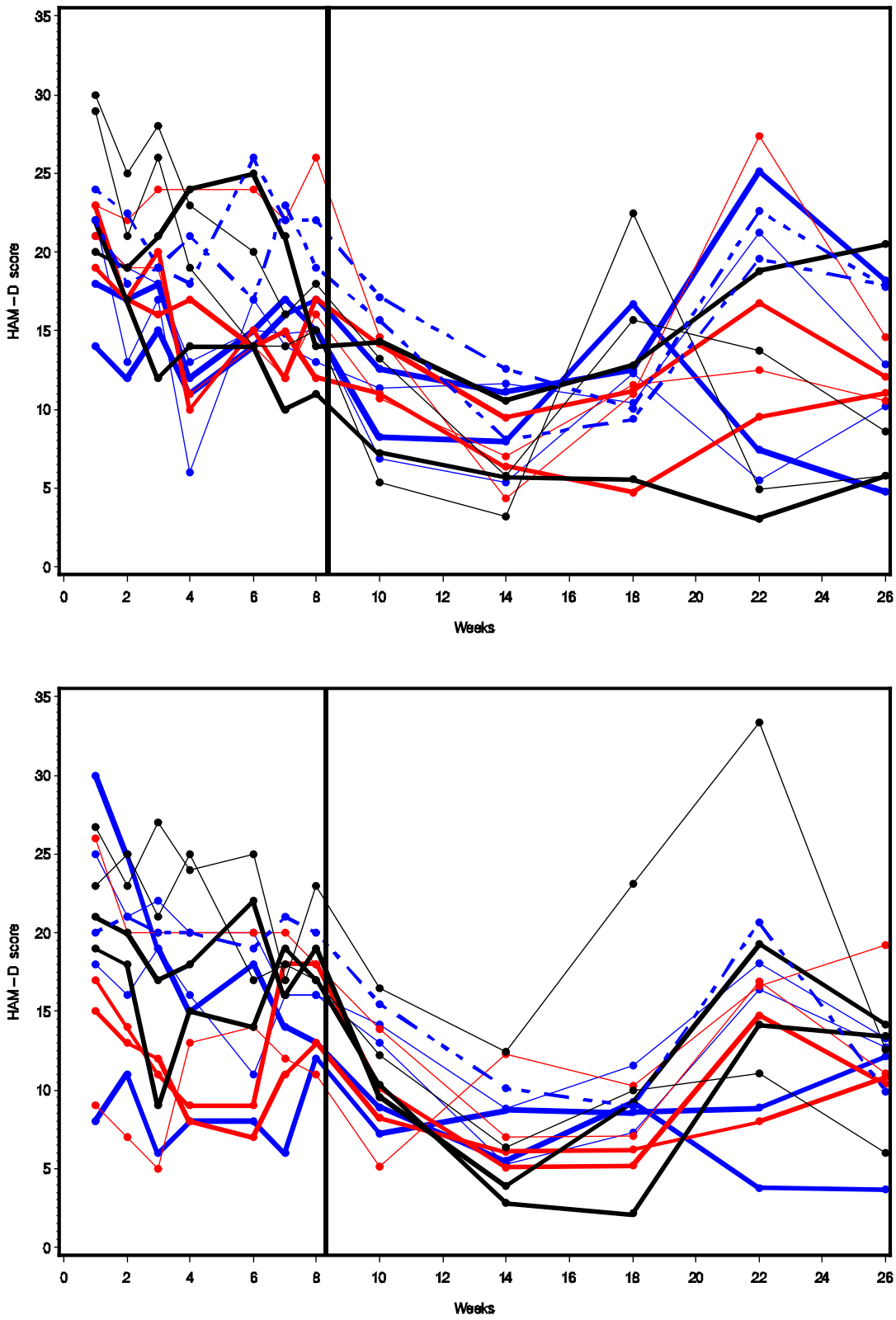



**Figure 6.15: Imputed Hamilton depression scale total score in the sertraline arm for participants who completed the acute phase of treatment (up to Week 8) and did not continue to the continuation phase due to protocol defined non-response at Week 8.**
The vertical black line at 8 weeks indicates the cut-off between observed and imputed data. Each line represents a participant.

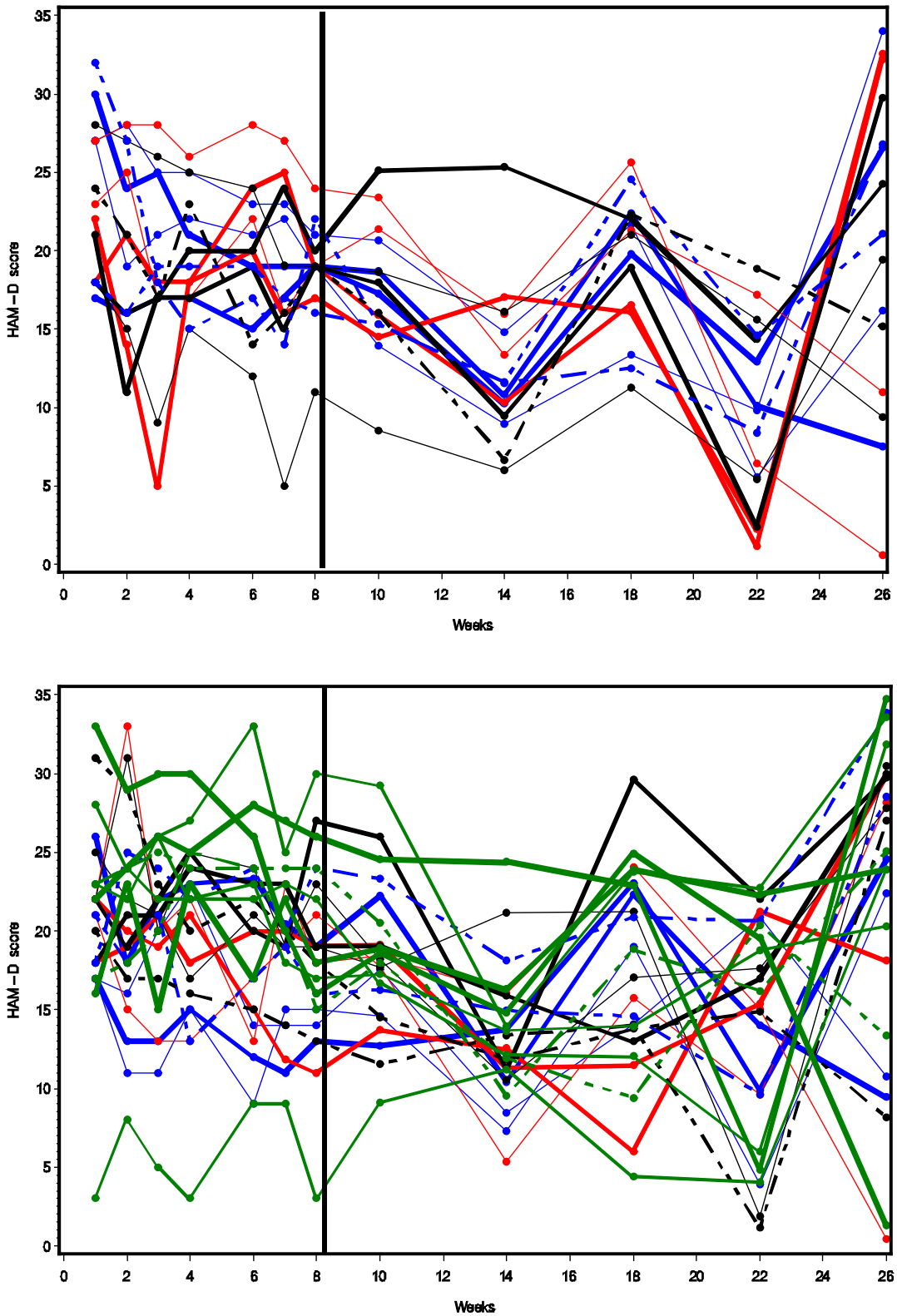



**Figure 6.16: Imputed Hamilton depression scale total score in the hypericum arm for participants who completed the acute phase of treatment (up to Week 8) and did not continue to the continuation phase due to protocol defined non-response at Week 8.**
The vertical black line at 8 weeks indicates the cut-off between observed and imputed data. Each line represents a participant.

All the participants included in Figure 6.14 to Figure 6.16 were terminated from the study at Week 8 because of lack of response. One therefore expects that the HAM-D scores would worsen or stay the same from Week 10 onwards, and not that it would improve. However, in the placebo and sertraline arms an improvement is seen in these imputed values.

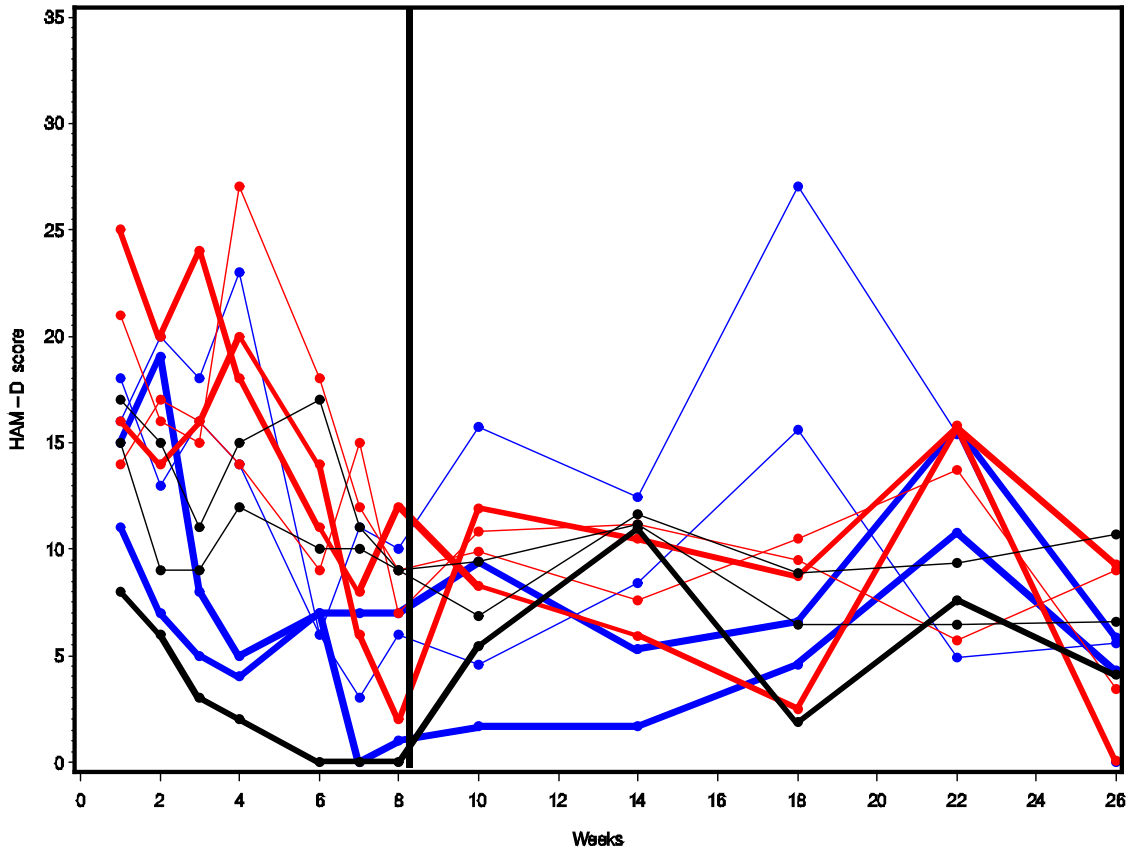



**Figure 6.17: Imputed Hamilton depression scale total score in all arms for participants who completed the acute phase of treatment (up to Week 8) and did not continue to the continuation phase. The participants were responding at Week 8, but chose not to continue with the study.**
The vertical black line at 8 weeks indicates the cut-off between observed and imputed data.

In contrast Figure 6.17 includes participants who responded at Week 8, but did not for some other reason continue into the continuation phase. One would expect these participants to have continued good responses in the imputed data. However, some of the imputed data showed a slight worsening of the depression scores or maintenance of the current score rather than continued improvement.

In Table 6.10 multiple imputation was used to calculate the mean and standard deviation at each of the visits. Multiple imputation can also be used to fit a longitudinal model to the data instead of summarising the data at each of the visits. Table 6.11 gives the results when a mixed model is fitted using the data sets created using multiple imputation. This can be contrasted to the mixed

model fitted in Table 6.7 without using multiple imputation. Both methods are valid under MAR assumptions and give similar results.

**Table 6.11: Hamilton depression scale. Likelihood-based parameter estimates and standard errors when using procedure MIXED in SAS after multiple imputation**

| | Data up to Week 8 | | | Data up to Week 26 | | |
|---|---|---|---|---|---|---|
| **Effect** | **Estimate** | **Standard error** | **p-value** | **Estimate** | **Standard error** | **p-value** |
| Intercept | 22.70 | 0.25 | | 22.55 | 0.31 | |
| Sertraline (ref: placebo) | 0.09 | 0.35 | 0.79 | -0.26 | 0.42 | 0.54 |
| Hypericum (ref: placebo) | 0.42 | 0.35 | 0.23 | 0.42 | 0.42 | 0.33 |
| Week | -1.15 | 0.09 | <0.001 | -1.50 | 0.08 | <0.001 |
| Week squared | | | | 0.05 | 0.003 | <0.001 |
| Interaction week and treatment (sertraline compared to placebo) | -0.23 | 0.13 | 0.08 | -0.04 | 0.08 | 0.59 |
| Interaction week and treatment (hypericum compared to placebo) | 0.01 | 0.13 | 0.95 | 0.10 | 0.09 | 0.24 |

One expects the estimates obtained from likelihood-based methods and multiple imputation to be similar; especially if the imputer's model was the same as the analyst's model. In this case the imputer's model included several covariates, deemed to possibly be associated with HAM-D score, which were not included in the analyst's model or in the likelihood-based model. Differences between these results and the results from the likelihood-based method (Table 6.7) were in part due to the inclusion of covariates in the multiple imputation. Considering the data up to Week 8, we draw similar conclusions; namely that the interaction between week and sertraline compared to placebo had a small p-value (0.05 and 0.08, respectively) and that the interaction between week and hypericum compared to placebo was not significant. The estimated effects were also similar between the two models. We concluded that over the first 8 weeks, the improvement in the sertraline arm is significantly better than the improvement in the placebo arm. The same cannot be said of hypericum.

Figure **6.18** gives the models fitted for all three treatment arms.

If we consider the data up to Week 26, there was no significant week by treatment interaction when multiple imputation was employed (Table 6.11). When the likelihood-based analysis was done, the estimates were different, and the p-value for the week by sertraline treatment arm compared to placebo arm was small (0.07). The multiple imputation results might not be reliable beyond Week 10, where more data were imputed than observed and autocorrelation was present. The results over the entire 26 weeks with multiple imputation should therefore be interpreted with caution.

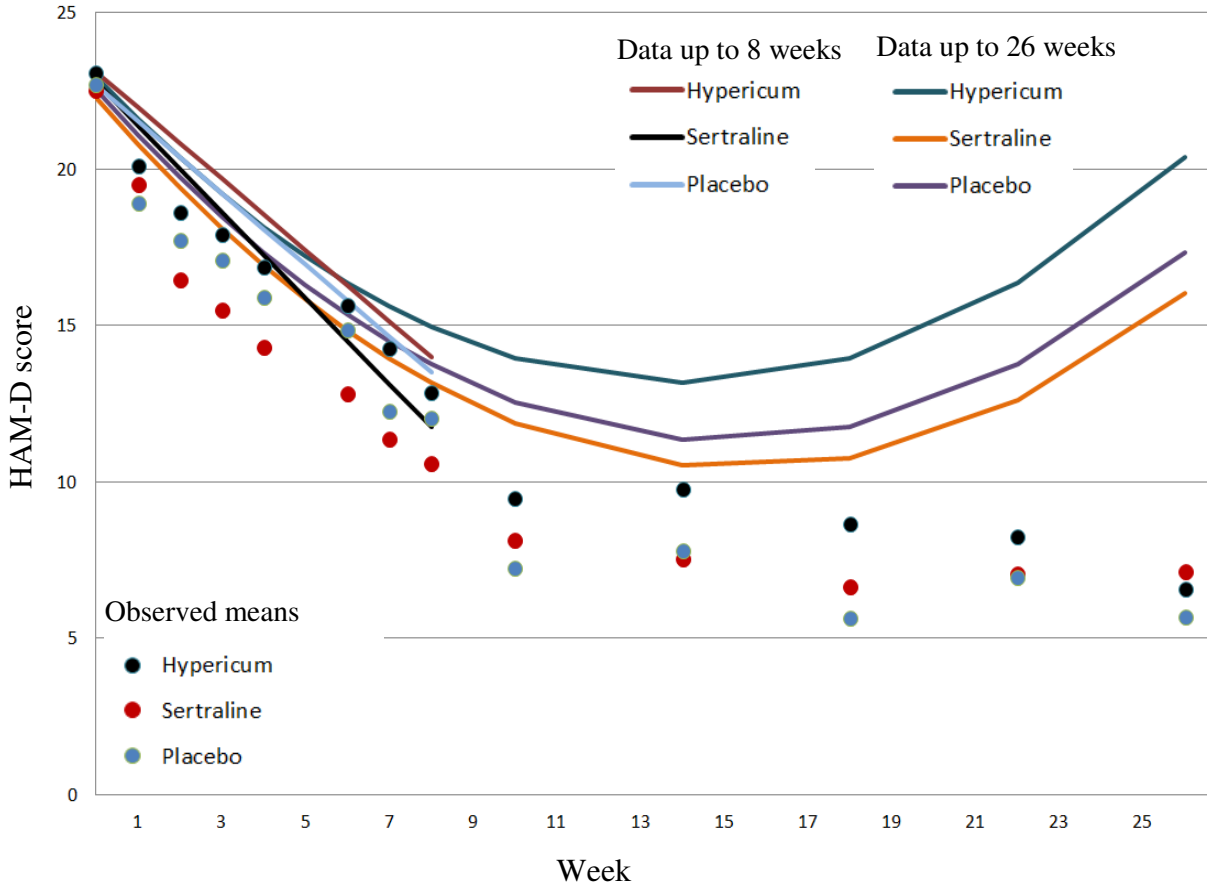



**Figure 6.18: Hamilton depression scale. Mixed model fitted after multiple imputations (Table 6.11)**

### 6.4.3 Bayesian analysis under MAR

A Bayesian model was fitted to the data of all participants, those with and without missing data; Probability distributions were estimated for the missing values and then simulations were drawn from the posterior distribution. One advantage of using Bayesian methods with missing data is that no change is needed to the model specified, i.e. the same model that would be appropriate if there weres no missing data can be specified.

We fitted a longitudinal model with a different unstructured covariance matrix in each of the treatment arms as follows:

$$\boldsymbol{y}_i | Z_i = k \sim \text{Multivariate Normal}(\boldsymbol{\mu}_k, \boldsymbol{\Omega}_k)$$

where $Z_i = k$ indicates that the observation belongs to treatment arm $k$

$$\mu_{ij} = \beta_0 + \beta_1 S_i + \beta_2 H_i + \beta_3 t_{ij} + (\beta_4 S_i + \beta_5 H_i) t_{ij}$$

for the model fitted up to Week 8. The variables are as defined in Section 6.4. The model fitted to the data up to Week 26, was

$$\mu_{ij} = \beta_0 + \beta_1 S_i + \beta_2 H_i + \beta_3 t_{ij} + (\beta_4 S_i + \beta_5 H_i) t_{ij} + \beta_6 t_{ij}^2 .$$

Uninformative priors were chosen for the unknown parameters of the model of interest. Different sets of prior distributions were assigned. In Set 1, the $\beta$ parameters $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$ and $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6)$, respectively are assigned Normal(0; 10,000) priors and the inverse of Ω was assigned a Wishart(I,5) prior, where I is the identity matrix. Prior Set 2 assigned all the $\beta$ parameters a Normal(0, 1000) prior and the inverse of Ω was assigned a Wishart(A,5) prior, where A is a diagonal matrix with 0.1 on the diagonal. Prior Set 3 assigned all the $\beta$ parameters a Normal(0, 10) prior and the inverse of Ω a Wishart(A,5) prior, where A is a diagonal matrix with 10 on the diagonal.

All models were fitted using OpenBugs, which uses MCMC methods and run for 40000 iterations, including 20000 burn-in iterations. Two chains with different starting values were used and convergence was checked by a visual inspection of the trace plots. All the runs discussed converged.

The choice of vague prior did not change the results appreciably in the MAR analysis (Table 6.12). The results were almost identical with prior sets 1 and 2. Of course, many more choices of priors are possible. Using the data up to Week 8, the interaction between week and sertraline arm compared to placebo arm was significant. The interaction between week and the hypericum arm compared to the placebo arm was not significant. The conclusions drawn were the same as with the likelihood-based analysis; although the absolute value of the estimate for the interaction between week and sertraline arm compared to the placebo arm was larger under the Bayesian model.

Using all the data up to Week 26, also leads to the conclusion that the decrease in HAM-D score in the sertraline arm compared to the placebo arm was significant, but the decrease over time in the hypericum arm compared to the placebo arm was not significant. This was the same conclusion as under the likelihood-based analysis.

Differences between the likelihood-based analysis and the Bayesian analysis can be attributed to the addition of prior distributions. Adding prior distributions added additional assumptions to this model, over and above the assumptions made by the substantive model. By choosing non-informative priors, we hope that the results are not sensitive to the choice of the prior. This was illustrated by the fact that the results did not change appreciably when the priors were varied.

**Table 6.12: Hamilton depression score posterior means, standard deviations and credible intervals according to Bayesian analysis under MAR assumptions**

| | Prior set 1<br>$\beta \sim Normal(0, 10\,000)$<br>Ω~Wishart(I,5)<br>I: Identity matrix | | | Prior set 2<br>$\beta \sim Normal(0, 1000)$<br>Ω~Wishart(A,5)<br>A: diag (0.1, 0.1, 0.1, 0.1, 0.1) | | | Prior set 3<br>$\beta s \sim Normal(0, 10)$<br>Ω~Wishart(A,5)<br>A: diag (10, 10, 10, 10, 10) | | |
|---|---|---|---|---|---|---|---|---|---|
| **Effect** | **Mean** | **SD** | **95% credible interval** | **Mean** | **SD** | **95% credible interval** | **Mean** | **SD** | **95% credible interval** |
| **Data up to Week 8** | | | | | | | | | |
| Intercept | 22.53 | 0.27 | | 22.53 | 0.27 | | 22.36 | 0.27 | |
| Sertraline (ref: placebo) | 0.04 | 0.37 | -0.69; 0.77 | 0.04 | 0.37 | -0.69; 0.77 | 0.20 | 0.37 | -0.53; 0.93 |
| Hypericum (ref: placebo) | 0.62 | 0.38 | -0.14; 1.37 | 0.62 | 0.38 | -0.14; 1.37 | 0.77 | 0.38 | 0.03; 1.52 |
| Week | -1.59 | 0.16 | -1.92; -1.28 | -1.59 | 0.16 | -1.92; -1.28 | -1.59 | 0.16 | -1.91; -1.27 |
| Interaction week and treatment (sertraline compared to placebo) | -0.54 | 0.22 | -0.96; -0.11 | -0.54 | 0.22 | -0.96; -0.11 | -0.54 | 0.22 | -0.97; -0.11 |
| Interaction week and treatment (hypericum compared to placebo) | -0.02 | 0.24 | -0.49; 0.44 | -0.02 | 0.24 | -0.49; 0.44 | -0.03 | 0.24 | -0.49; 0.43 |
| **Data up to Week 26** | | | | | | | | | |
| Intercept | 22.45 | 0.26 | | 22.45 | 0.26 | | 22.32 | 0.26 | |
| Sertraline (ref: placebo) | 0.16 | 0.34 | -0.53; 0.83 | 0.16 | 0.34 | -0.53; 0.83 | 0.29 | 0.34 | -0.39; 0.97 |
| Hypericum (ref: placebo) | 0.67 | 0.36 | -0.03; 1.38 | 0.67 | 0.36 | -0.03; 1.38 | 0.80 | 0.36 | 0.10; 1.51 |
| Week | -3.15 | 0.26 | -3.66; -2.63 | -3.15 | 0.26 | -3.66; -2.63 | -3.13 | 0.27 | -3.65; -2.62 |
| Week squared | 0.39 | 0.05 | 0.28; 0.49 | 0.39 | 0.05 | 0.28; 0.49 | 0.39 | 0.05 | 0.28; 0.49 |
| Interaction week and treatment (sertraline compared to placebo) | -0.50 | 0.20 | -0.90;-0.11 | -0.50 | 0.20 | -0.90;-0.11 | -0.51 | 0.20 | -0.90; -0.11 |
| Interaction week and treatment (hypericum compared to placebo) | 0.06 | 0.22 | -0.37; 0.49 | 0.06 | 0.22 | -0.37; 0.49 | 0.06 | 0.22 | -0.37; 0.49 |

SD: Standard deviation
Highlighted sections indicate statistical significance

### 6.4.4 Inverse probability weighting

Inverse probability weighting assumes monotone missingness and we imputed the few non-monotone missing values in order to create a monotone dataset. Inverse probability weighting was done through a two-stage process. A logistic regression model was fitted at each visit to determine the probability of observing the outcome measurement. Baseline covariates and response variables for visits 0 to t-1 were included. For each visit the probability of being observed was conditional on being observed at the previous visit. Any participant who was missing at a visit was not included in the estimation of the probability of being observed at a subsequent visit. The conditional probability of being observed at visit *t* was

$$\lambda_{it} = P(R_{it} = 1 | R_{i,t-1} = 1, X_i, Y_{i1}, \ldots, Y_{i,t-1})$$

The occasion specific unconditional probability of being observed for each visit, $\pi_{it} = P(R_{it} = 1)$, was then estimated as $\hat{\pi}_{it} = \hat{\lambda}_{i1} \times \hat{\lambda}_{i2} \times \ldots \times \hat{\lambda}_{it}$. The inverse of this product of conditional probabilities was used as the occasion specific weight. A linear model of interest was then fitted to the responders using weighted regression, where the weights were the reciprocal of this product, namely $\frac{1}{\hat{\pi}_{it}}$.

GEE is not generally valid unless the missingness process is MCAR, but weighted GEE is valid when the missingness process is MAR. The results of the weighted GEE method is given in Table 6.13, both with a logistic regression model including additional covariates (baseline HAM-D score, age, gender, BDI, duration of depression, GAF, CGI-S and CGI-I scales) and with a logistic regression only including outcome measurements. Including covariates in the calculation of the weights helps improve the accuracy of the weights and therefore improves the results obtained with this method. The weighted GEE analysis used a different estimation method and is not directly comparable with the models given earlier. Standard errors should be calculated using bootstrap methods, because the standard errors calculated using standard methods are not accurate. We give here the incorrectly calculated standard errors calculated through procedure GENMOD applying GEE, because of problems experienced with bootstrap that will be discussed later. We argue that the loss of precision in the calculation of the standard errors is less serious than the possibility of bias introduced when some of the bootstrap samples do not converge. In Chapter 5 standard errors were calculated using bootstrap and the standard errors were similar to the standard errors calculated using standard methods in procedure GENMOD. We therefore believe that the error in the calculation of the standard errors is fairly small.

Up to Week 8 the weighted GEE showed an improvement in mean HAM-D score over time in the sertraline arm compared to the placebo arm. This improvement was small and not statistically significant. The interaction between week and the hypericum arm compared to the placebo arm was almost 0 and not statistically significant. The models with and without covariates in the logistic regression model to calculate the weights gave similar results.

To improve on the inverse probability weighted GEE fitted, a doubly robust method was also used. The application of this method was discussed in more detail in Section 5.4.4. The standard SAS output for the variance of the estimators underestimate the true variance since the response values for the final model are the predicted values from a previous weighted model. Multiple bootstrap samples were taken in an attempt to correctly estimate the variance. Each data set consisted of a sample with replacement. The mean and standard error of the parameter estimates were calculated by taking the mean and standard error of the estimates obtained from each of the bootstrap samples.

**Table 6.13: Hamilton depression score, inverse probability weighting methods, data up to Week 8**

| | Weighted GEE without covariates | | | Weighted GEE with covariates | | |
|---|---|---|---|---|---|---|
| **Effect** | **Estimate** | **Standard error** | **p-value** | **Estimate** | **Standard error** | **p-value** |
| Intercept | 19.69 | 0.53 | | 19.78 | 0.54 | |
| Sertraline (ref: placebo) | -0.62 | 0.81 | 0.45 | -0.73 | 0.84 | 0.38 |
| Hypericum (ref: placebo) | 1.01 | 0.77 | 0.19 | 0.98 | 0.78 | 0.21 |
| Week | -0.90 | 0.11 | <0.001 | -0.90 | 0.11 | <0.001 |
| Interaction week and treatment (sertraline compared to placebo) | -0.12 | 0.15 | 0.41 | -0.13 | 0.16 | 0.41 |
| Interaction week and treatment (hypericum compared to placebo) | -0.01 | 0.15 | 0.98 | -0.04 | 0.15 | 0.80 |
| | **Doubly robust estimator without covariates** | | | **Doubly robust estimator with covariates** | | |
| | **Estimate** | **Standard error** | **p-value** | **Estimate** | **Standard error** | **p-value** |
| Intercept | 19.17 | 0.63 | | 19.12 | 0.67 | |
| Sertraline (ref: placebo) | -0.13 | 0.72 | 0.86 | -0.24 | 0.77 | 0.76 |
| Hypericum (ref: placebo) | 0.98 | 0.84 | 0.25 | 0.74 | 0.79 | 0.36 |
| Week | -0.74 | 0.19 | <0.001 | -0.68 | 0.25 | 0.01 |
| Interaction week and treatment (sertraline compared to placebo) | -0.30 | 0.14 | 0.04 | -0.30 | 0.15 | 0.05 |
| Interaction week and treatment (hypericum compared to placebo) | -0.06 | 0.15 | 0.68 | -0.12 | 0.18 | 0.51 |

With covariates and without covariates refers to whether auxiliary covariates were included in the logistic regression model used to calculate the probability of being observed. Treatment arm was included as a covariate in both models.

The logistic regression in step 1 used to calculate the probability of being observed had many convergence problems, because the bootstrapped samples included duplicate records compared to the original data set. The convergence problems were more extreme when more covariates were included in the models. If the proportion of non-converging samples is high, the estimates may be biased. If auxiliary covariates were included only 34 of 200 bootstrap samples converged. If auxiliary covariates were not included only 61 of 200 bootstrap samples converged. The resulting estimates could therefore be biased. Despite the possibility of bias introduced, the results of the doubly robust estimator (without covariates) are similar to the results found with the likelihood based, multiple imputation and Bayesian methods. From the doubly robust analysis we conclude that mean HAM-D score in the sertraline arm improved more over time than mean HAM-D score in the placebo arm, and this difference was statistically significant. The same could not be said about the hypericum arm compared to the placebo arm.

**Table 6.14: Hamilton depression score, inverse probability weighting methods, data up to Week 26**

| Effect | Weighted GEE without covariates Estimate | Standard error | p-value | Weighted GEE with covariates Estimate | Standard error | p-value |
|---|---|---|---|---|---|---|
| Intercept | 19.20 | 0.73 | | 19.16 | 0.78 | |
| Sertraline (ref: placebo) | 0.18 | 0.95 | 0.85 | 0.04 | 0.90 | 0.97 |
| Hypericum (ref: placebo) | 1.72 | 0.94 | 0.07 | 0.96 | 0.92 | 0.30 |
| Week | -0.96 | 0.13 | <0.001 | -0.91 | 0.16 | <0.001 |
| Interaction week and treatment (sertraline compared to placebo) | -0.20 | 0.12 | 0.08 | -0.16 | 0.07 | 0.02 |
| Interaction week and treatment (hypericum compared to placebo) | -0.12 | 0.11 | 0.26 | 0.01 | 0.08 | 0.94 |
| Week squared | 0.02 | 0.004 | <0.001 | 0.02 | 0.01 | 0.01 |
| | **Doubly robust estimator without covariates** | | | **Doubly robust estimator with covariates** | | |
| | **Estimate** | **Standard error** | **p-value** | **Estimate** | **Standard error** | **p-value** |
| Intercept | 19.85 | 2.00 | | | | |
| Sertraline (ref: placebo) | -1.61 | 3.08 | 0.61 | | | |
| Hypericum (ref: placebo) | 0.95 | 1.68 | 0.58 | | | |
| Week | -1.16 | 0.21 | <0.001 | | | |
| Interaction week and treatment (sertraline compared to placebo) | -0.05 | 0.13 | 0.70 | | | |
| Interaction week and treatment (hypericum compared to placebo) | -0.04 | 0.14 | 0.79 | | | |
| Week squared | 0.03 | 0.01 | <0.001 | | | |

With covariates and without covariates refers to whether auxiliary covariates were included in the logistic regression model used to calculate the probability of being observed. Treatment arm was included as a covariate in both models.

For the data up to Week 26, the weighted GEE with and without covariates lead to similar conclusions. We conclude that there is a significant interaction between week and the sertraline arm compared to the placebo arm (the p-value is small and almost significant for the weighted GEE without covariates). The interaction between week and the hypericum arm compared to the placebo arm was not significant. Only 4 samples converged in the doubly robust estimator that included covariates in the missing data model. These results are not presented. The logistic regression without covariates had slightly better convergence and 19 of the 200 samples converged. The results of the doubly robust estimator could be biased because such a large number of bootstrap samples are excluded and we did not consider it further here.

### 6.4.5 Conclusion of MAR analysis

It was reasonable to assume that the data might be MAR; thus the multiple imputation, likelihood-based analysis, Bayesian analyses, pattern-mixture model using ACMV identifying restrictions (discussed in Section 6.5.1.3 with the MNAR models) and inverse probability weighting performed under MAR assumptions could be valid or reasonable approaches (Table 6.15). Up to Week 8, where most data were observed, the multiple imputation, likelihood-based analysis, Bayesian analysis and inverse probability weighting gave similar results. We concluded from this that the

interaction between week and sertraline compared to placebo was significant, and the interaction between week and hypericum compared to placebo was not.

Analysing data up to Week 26, where the majority of the data were missing, the multiple imputation results differed from the results with likelihood-based and Bayesian methods. The multiple imputation chain did not reach a stationary distribution and autocorrelation existed between the imputations from Week 10 onwards. For this reason, we believed that the results of the multiple imputation up to Week 26 should be interpreted with caution. If we only considered the likelihood-based, Bayesian and weighted GEE methods, we concluded that there was a significant interaction between week and sertraline compared to placebo, but not between week and hypericum compared to placebo.

**Table 6.15: Summary of findings in all MAR sensitivity analyses**

| | Interaction week and sertraline arm | | Interaction week and hypericum arm | |
|---|---|---|---|---|
| | **Estimate** | **p-value** | **Estimate** | **p-value** |
| **Data up to 8 weeks** | | | | |
| Likelihood-based methods | -0.24 | 0.05 | 0.03 | 0.79 |
| Multiple imputation | -0.23 | 0.08 | 0.01 | 0.95 |
| Bayesian analysis (Prior set 1) | -0.54 | Significant | -0.02 | Not significant |
| Inverse probability weighting (doubly robust without covariates) | -0.30 | 0.04 | -0.06 | 0.68 |
| Weighted GEE with covariates | -0.13 | 0.41 | -0.04 | 0.80 |
| **Data up to 26 weeks** | | | | |
| Likelihood-based methods | -0.10 | 0.07 | -0.04 | 0.48 |
| Multiple imputation | -0.04 | 0.59 | 0.10 | 0.24 |
| Bayesian analysis (Prior set 1) | -0.50 | Significant | 0.06 | Not significant |
| Weighted GEE with covariates | -0.16 | 0.02 | 0.01 | 0.94 |

We concluded that the sertraline arm was superior to the placebo arm in treating depression. The same was not observed for the hypericum arm. We could thus conclude that hypericum was not superior to placebo, in a clinical trial where a standard anti-depressive (sertraline) was found to be superior to placebo. This changed the interpretation of the clinical trial substantially from what was concluded previously, where missing data were ignored in the analysis (Hypericum Depression Trial Study Group, 2002).

We were interested in Estimand 1, as described by the National Research Council (2010). The methods used attempted to impute or otherwise include data up to Week 8 for all participants randomised, therefore enabling Estimand 1. However, the assumption was made that participants who dropped out continued to adhere to treatment in the same way as they did while on treatment. This assumption was not completely unrealistic, since these treatments were available and participants could continue to obtain the medication outside of the study setting. It was also

possible that participants who discontinued study visits could obtain other rescue medication or take no medication whatsoever, which would invalidate this assumption.

If the aim was to estimate drug efficacy, one is also interested in Estimand 3. It is unlikely that the data collected in this trial would enable one to estimate Estimand 3, the difference in outcome improvement if all participants tolerated and adhered to treatment. There was simply too much non-adherence due to drop-out. An estimate for this estimand can be calculated by only including data while participants were adhering to treatment and then inferring future profiles using this data. This was not done. Estimand 6 can more realistically be obtained than Estimand 3, in which case, the methods discussed in Carpenter et al. (2013) would be useful. Estimand 6 is estimated using a weighted average of the treatment effect at endpoint in those who adhered to study medication and the treatment effect in those who discontinued study medication. The placebo arm is then used to estimate the treatment effect in participants who discontinued study medication in the active treatment arm.

The analysis including all data up to Week 26 does not answer any well-defined research question or address any estimand of interest. Data are collected up to Week 8 in all participants who do not drop out, and from Week 10 data are collected only for participants who responded to treatment. During the analysis data are imputed or simulated for the participants who did not respond according to the patterns observed up to that time point. The results depend more on the treatment of the missing data than on what was observed. A more useful analysis may be to only analyse the data from Week 8 to Week 26 in those participants who responded at Week 8. This would then be a conditional analysis; given that a participant responded at Week 8, what is the continued treatment effect from Week 10 to Week 26. This is what was done in Sarris et al. (2012), who handled missing data after Week 10 by LOCF and not by appropriate methods. This is a randomised comparison any more. This conditional analysis needs to be done taking missing data into account, since only a small number of participants who entered the continuation phase completed the entire study to Week 26; 28 (57.1%) in the sertraline arm, 24 (63.2%) in the hypericum arm and 27 (64.3%) in the placebo arm.

This data set highlights a major problem with the missing data techniques. Even though the methods exist to analyse data under MAR assumptions and these methods can easily be implemented in standard software, the methods do not perform well, and might not converge or reach stationary distributions when the majority of data are missing. Sparse dropout patterns also create problems during the analysis of data.

# 6.5 Analysis of HAM-D score under MNAR assumptions

## 6.5.1 Pattern-mixture models

The patterns of missing data are given in Table 6.2. The first challenge when using pattern-mixture models is to determine how the various dropout patterns should be grouped. Three different dropout variables were defined. In the first model dropout was a binary variable created to indicate whether a participant completed the study or not. Dropout was assigned a value of zero if the study was completed and a value of one if the study was not completed (Model 1 and Model 3). In Model 1 data up to Week 8 only were considered and in Model 3 data up to Week 26 were considered. In addition, a second dropout variable was created with three categories, completing the entire study, being regarded as a treatment failure at Week 8 and discontinued from the study and dropped out of the study for any other reason. This model (Model 2) was fitted using data up to Week 26. Table 6.16 gives the number of participants in each of the categories of missing data.

Only 24.1% of participants completed the study up to 26 weeks, while 72.1% of participants completed the first 8 weeks of the study. In Model 2 the largest group was the participants who dropped out of the study (Table 6.16).

**Table 6.16: Number of participants in each of the categories of missing data**

| | Sertraline, n = 111 | Hypericum, n = 113 | Placebo, n = 116 |
|---|---|---|---|
| **Model 1: Binary dropout variable for study up to 8 weeks** | | | |
| Completed | 79 (71.2%) | 82 (72.6%) | 84 (72.4%) |
| Dropped out | 32 (28.8%) | 31 (27.4%) | 32 (27.6%) |
| **Model 2: Three dropout categories (up to 26 weeks)** | | | |
| Completed study | 31 (27.9%) | 24 (21.2%) | 27 (23.3%) |
| Treatment failure at Week 8 | 27 (24.3%) | 40 (35.4%) | 38 (32.8%) |
| Dropout | 53 (47.7%) | 49 (43.4%) | 51 (44.0%) |
| **Model 3: Binary dropout variable for entire study (up to 26 weeks)** | | | |
| Completed | 31 (27.9%) | 24 (21.2%) | 27 (23.3%) |
| Dropped out | 80 (72.1%) | 89 (78.8%) | 89 (76.7%) |

Table 6.17 gives the estimated β-coefficients for each of the patterns of missing data. We assumed an unstructured covariance matrix in all cases. With some of the models, especially the models with dropout, the models did not converge because the Hessian matrix was not positive definite. A heterogeneous compound symmetry covariance matrix was assumed for the dropouts in Model 1 and Model 3. The heterogeneous compound symmetry assumes the same correlations between observations at any two pairs of times, while allowing unequal variances at the different time points. This is a relatively simple covariance structure and is constructed as follows:

$$\begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho & \sigma_1\sigma_4\rho & \sigma_1\sigma_5\rho & \sigma_1\sigma_6\rho & \sigma_1\sigma_7\rho & \sigma_1\sigma_8\rho \\ \sigma_2\sigma_1\rho & \sigma_2^2 & \sigma_2\sigma_3\rho & \sigma_2\sigma_4\rho & \sigma_2\sigma_5\rho & \sigma_2\sigma_6\rho & \sigma_2\sigma_7\rho & \sigma_2\sigma_8\rho \\ \sigma_3\sigma_1\rho & \sigma_3\sigma_2\rho & \sigma_3^2 & \sigma_3\sigma_4\rho & \sigma_3\sigma_5\rho & \sigma_3\sigma_6\rho & \sigma_3\sigma_7\rho & \sigma_3\sigma_8\rho \\ \sigma_4\sigma_1\rho & \sigma_4\sigma_2\rho & \sigma_4\sigma_3\rho & \sigma_4^2 & \sigma_4\sigma_5\rho & \sigma_4\sigma_6\rho & \sigma_4\sigma_7\rho & \sigma_4\sigma_8\rho \\ \sigma_5\sigma_1\rho & \sigma_5\sigma_2\rho & \sigma_5\sigma_3\rho & \sigma_5\sigma_4\rho & \sigma_5^2 & \sigma_5\sigma_6\rho & \sigma_5\sigma_7\rho & \sigma_5\sigma_8\rho \\ \sigma_6\sigma_1\rho & \sigma_6\sigma_2\rho & \sigma_6\sigma_3\rho & \sigma_6\sigma_4\rho & \sigma_6\sigma_5\rho & \sigma_6^2 & \sigma_6\sigma_7\rho & \sigma_6\sigma_8\rho \\ \sigma_7\sigma_1\rho & \sigma_7\sigma_2\rho & \sigma_7\sigma_3\rho & \sigma_7\sigma_4\rho & \sigma_7\sigma_5\rho & \sigma_7\sigma_6\rho & \sigma_7^2 & \sigma_7\sigma_8\rho \\ \sigma_8\sigma_1\rho & \sigma_8\sigma_2\rho & \sigma_8\sigma_3\rho & \sigma_8\sigma_4\rho & \sigma_8\sigma_5\rho & \sigma_8\sigma_6\rho & \sigma_8\sigma_7\rho & \sigma_8^2 \end{bmatrix}$$

**Table 6.17: Estimates and standard errors of the different patterns of missing data**

| | Model 1 (data up to 8 weeks) | | | | Model 3 (data up to week 26) | | | |
|---|---|---|---|---|---|---|---|---|
| | Completed | | Dropped out | | Completed | | Dropped out | |
| **Effect** | **Estimate** | **SE** | **Estimate** | **SE** | **Estimate** | **SE** | **Estimate** | **SE** |
| Intercept | 22.86 | 0.29 | 22.18 | 0.50 | 21.53 | 0.53 | 22.70 | 0.34 |
| Sertraline (ref: placebo) | 0.42 | 0.44 | 0.54 | 0.68 | 0.25 | 0.67 | 0.41 | 0.45 |
| Hypericum (ref: placebo) | -0.23 | 0.37 | 0.82 | 0.75 | 0.88 | 0.68 | 0.46 | 0.47 |
| Week | -1.18 | 0.10 | -0.94 | 0.29 | -1.91 | 0.06 | -1.53 | 0.09 |
| Interaction week and treatment (sertraline compared to placebo) | -0.20 | 0.13 | -0.63 | 0.44 | 0.09 | 0.05 | -0.19 | 0.08 |
| Interaction week and treatment (hypericum compared to placebo) | -0.01 | 0.13 | 0.02 | 0.41 | 0.06 | 0.05 | 0.05 | 0.08 |
| Week squared | | | | | 0.05 | 0.002 | 0.07 | 0.01 |

| | Model 2: Three dropout categories (up to 26 weeks) | | | | | |
|---|---|---|---|---|---|---|
| | Completed study | | Treatment failure at Week 8 | | Dropout | |
| **Effect** | **Estimate** | **SE** | **Estimate** | **SE** | **Estimate** | **SE** |
| Intercept | 21.53 | 0.53 | 23.06 | 0.50 | 22.11 | 0.70 |
| Sertraline (ref: placebo) | 0.25 | 0.67 | -0.01 | 0.62 | -0.51 | 0.96 |
| Hypericum (ref: placebo) | 0.88 | 0.68 | 0.23 | 0.77 | 0.12 | 0.96 |
| Week | -1.91 | 0.06 | -1.39 | 0.21 | -2.18 | 0.16 |
| Interaction week and treatment (sertraline compared to placebo) | 0.09 | 0.05 | -0.24 | 0.14 | -0.14 | 0.14 |
| Interaction week and treatment (hypericum compared to placebo) | 0.06 | 0.05 | -0.07 | 0.14 | 0.26 | 0.16 |
| Week squared | 0.05 | 0.002 | 0.11 | 0.02 | 0.09 | 0.01 |

SE: Standard error

In Model 3 the Hessian matrix was not positive definite when either an unstructured covariance matrix or a heterogeneous compound symmetry matrix was assumed for the participants who dropped out. This was despite this pattern containing almost half the participants, therefore the convergence problems were not caused by small sample size. This model was fitted using an autoregressive matrix, which was appropriate, since the visits were spaced at equal intervals. MAR models assume the slope of the HAM-D scores over time is the same across patterns, however the slopes differed by dropout pattern for all the models fitted.

#### 6.5.1.1 Pattern-mixture models using random-effects mixed models

Procedure MIXED in SAS was used to fit the model with HAM-D score as the dependent variable. Independent variables included the fixed, categorical effects of treatment, dropout, and the dropout by treatment interaction, as well as the continuous effect of time. The time by treatment and

dropout by time and dropout by treatment interaction and the three level interaction of dropout, treatment and time were included. Adding a term for time squared improved the fit of the model since it was not a linear pattern. An unstructured covariance matrix was used to model the within participant errors. Parameters were estimated using REML with the Newton–Raphson algorithm. Asymptotically exact standard errors were obtained. Denominator degrees of freedom were estimated using the Kenward-Roger correction (Kenward & Roger, 1997).

**The regression equation for the pattern-mixture model with a binary dropout variable (Model 1 and Model 3)**

$$Y_{ij} = \beta_0 + \beta_1 H_i + \beta_2 S_i + \beta_3 t_{ij} + (\beta_{41} H_i + \beta_{42} S_i) t_{ij} + \beta_5 D_i + \beta_6 t_{ij} D_i + \beta_7 t_{ij}^2 + \beta_{81} D_i H_i + \beta_{82} D_i S_i + (\beta_{91} H_i + \beta_{92} S_i) t_{ij} D_i + \varepsilon_{ij}$$

**Regression equation for pattern-mixture model with three dropout categories (Model 2)**

$$Y_{ij} = \beta_0 + \beta_1 H_i + \beta_2 S_i + \beta_3 t_{ij} + (\beta_{41} H_i + \beta_{42} S_i) t_{ij} + \beta_{51} C_i + \beta_{52} W_i + (\beta_{61} C_i + \beta_{62} W_i) t_{ij} + \beta_7 t_{ij}^2 + \beta_{811} C_i H_i + \beta_{812} W_i H_i + \beta_{821} C_i S_i + \beta_{822} W_i S_i + (\beta_{911} C_i H_i + \beta_{912} W_i H_i + \beta_{921} C_i S_i + \beta_{922} W_i S_i) t_{ij} + \varepsilon_{ij}$$

where $C_i$= 1 if completed, 0 otherwise; $W_i$= 1 if participant was a treatment failure at Week 8 and terminated from the study, 0 otherwise

For Models 1 and 3, a binary variable for dropout was added to a longitudinal model as a main effect and as interactions with the variables treatment and week. This allowed one to examine the degree to which dropouts differed from completers in terms of the outcome variable. The main effect of dropout was not significant in Model 1, but significant in Model 3. The degree to which the dropout pattern moderated the influence of other model terms was investigated by looking at interactions with the dropout pattern. The interaction of dropout and week was significant, but the interaction of treatment and dropout was not significant in either Model 1 or 3.

In Model 2, the missing data variable was not a binary variable, but consisted of three patterns. The effect of the missing data pattern was significant. The interaction of dropout and week was significant, but the interaction of treatment and dropout was not significant in any of the models. In Model 1 the only significant interaction was the interaction between week and dropout. Although the pattern-mixture model fitted can be simplified by excluding some non-significant interactions from the model, we did not exclude any interactions. The interaction between treatment and dropout was retained, since the p-value was less than 0.10, even though it was not below 0.05. The interaction between week and treatment was also retained, since this is the primary comparison we are interested in (Table 6.18). These fitted models are given in Table 6.19.

**Table 6.18: F-test of fixed effects for models including all two way interactions and a three way interaction between treatment arm, dropout and week, adjusted for all other variables and interactions in the model**

| Effect | Degrees of freedom | | F-value | p-value |
|---|---|---|---|---|
| | Numerator | Denominator | | |
| **Model 1: Binary dropout variable, data up to 8 weeks** | | | | |
| Treatment | 2 | 205 | 3.85 | 0.02 |
| Week | 1 | 292 | 286.6 | <0.001 |
| Week$^2$ | 1 | 280 | 72.3 | <0.001 |
| Dropout | 1 | 348 | 0.22 | 0.64 |
| Interaction week and treatment | 2 | 165 | 1.14 | 0.32 |
| Interaction between week and dropout | 1 | 418 | 11.3 | 0.001 |
| Interaction between treatment and dropout | 2 | 235 | 2.52 | 0.08 |
| Interaction between week, treatment and dropout | 2 | 293 | 1.93 | 0.15 |
| **Model 2: Three dropout categories (up to 26 weeks)** | | | | |
| Treatment | 2 | 186 | 0.04 | 0.96 |
| Week | 1 | 187 | 1473 | <0.001 |
| Week$^2$ | 1 | 77 | 704 | <0.001 |
| Dropout | 2 | 239 | 4.71 | 0.01 |
| Interaction week and treatment | 2 | 204 | 2.88 | 0.06 |
| Interaction between week and dropout | 2 | 215 | 85 | <0.001 |
| Interaction between treatment and dropout | 4 | 179 | 1.02 | 0.40 |
| Interaction between week, treatment and dropout | 4 | 184 | 2.55 | 0.04 |
| **Model 3: Binary dropout variable for entire study (up to 26 weeks)** | | | | |
| Treatment | 2 | 112 | 0.19 | 0.83 |
| Week | 1 | 105 | 1221 | <0.001 |
| Week$^2$ | 1 | 74 | 528 | <0.001 |
| Dropout | 1 | 165 | 4.90 | 0.03 |
| Interaction week and treatment | 2 | 138 | 4.16 | 0.02 |
| Interaction between week and dropout | 1 | 145 | 51.4 | <0.001 |
| Interaction between treatment and dropout | 2 | 111 | 1.92 | 0.15 |
| Interaction between week, treatment and dropout | 2 | 137 | 4.48 | 0.01 |

The trajectory for mean HAM-D scores over time was different for participants who dropped out and participants who completed the study. For participants who completed the study there was little difference between the treatment arms. Amongst dropouts the sertraline arm showed continued improvement, while the placebo and hypericum arms did not. A visual examination of the observed and fitted means indicate that the models fit well in all treatment arms for the completers. For those who dropped out, the model did not fit well for the sertraline arm (Figure 6.19).

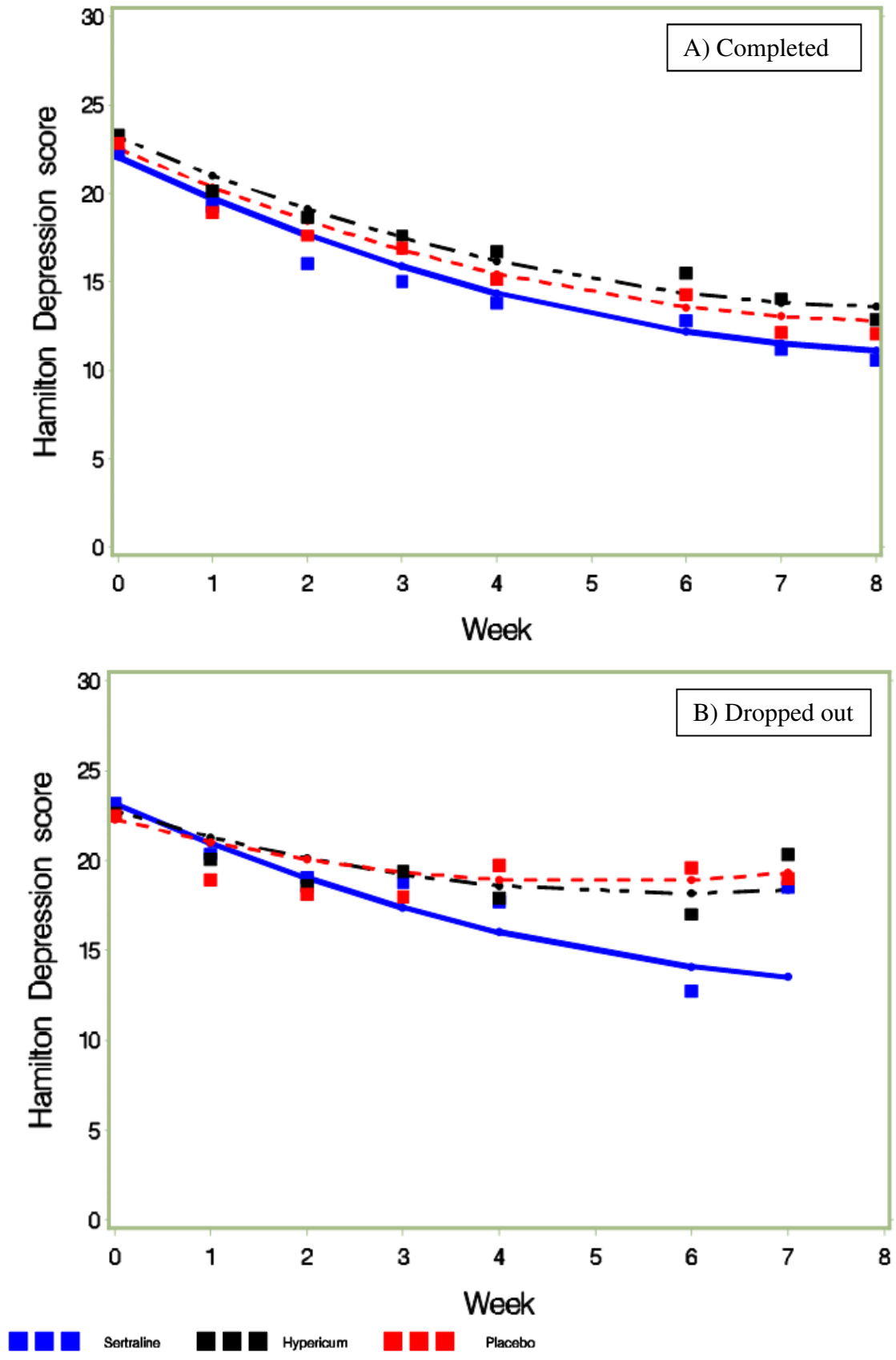



**Figure 6.19: Hamilton depression scale by dropout pattern, Model 1 with data up to Week 8. Squares indicate observed data and lines indicate predicted data.**

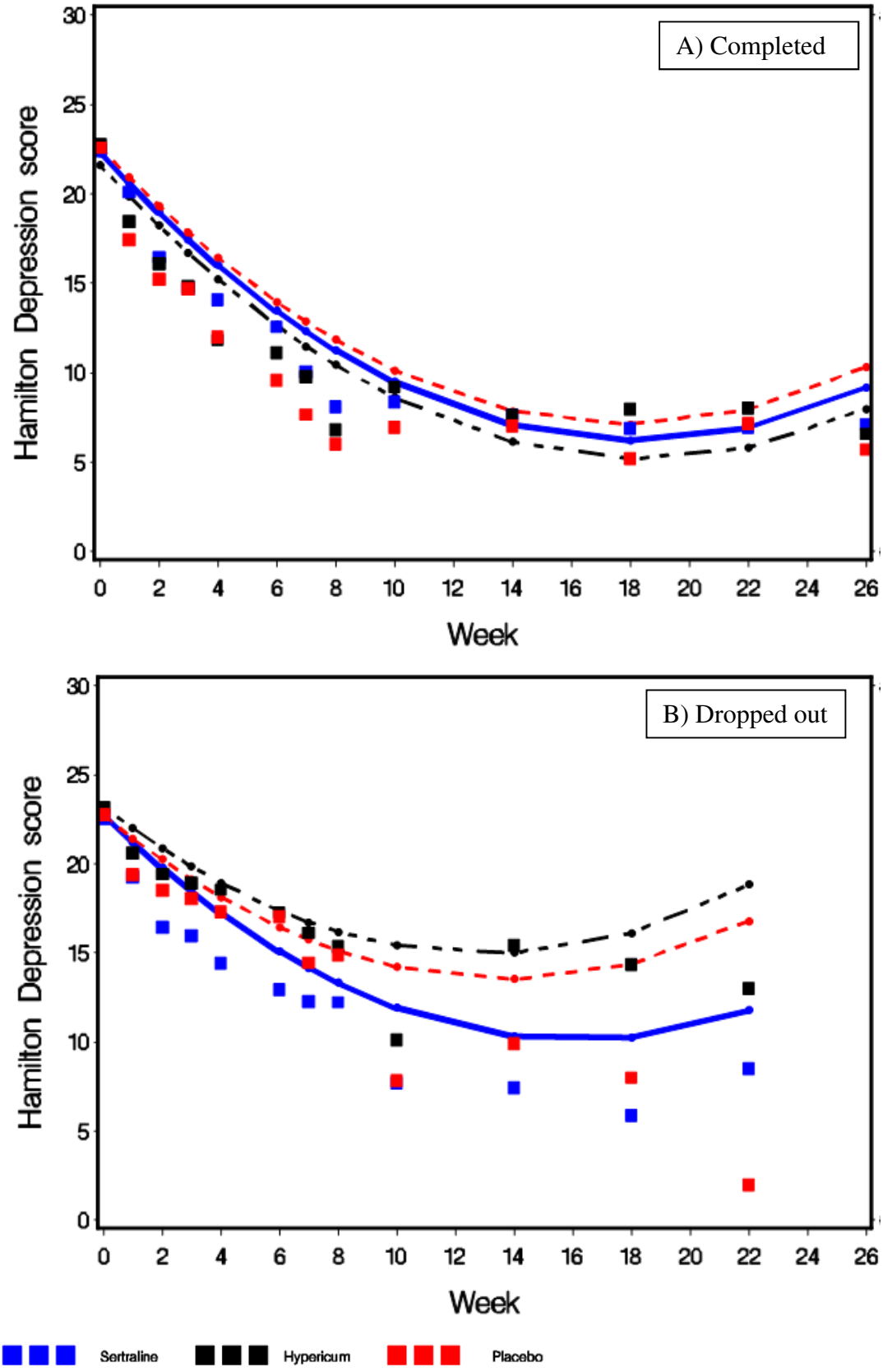



**Figure 6.20: Hamilton depression scale by dropout pattern, Model 3 with data up to Week 26. Squares indicate observed data and lines indicate predicted data.**

**Table 6.19: Hamilton depression scale, fitted as pattern-mixture model**

| Effect | Model 1 Binary dropout variable, 8 weeks | | | Model 2 Three dropout categories, 26 weeks | | | Model 3 Binary dropout variable, 26 weeks | | |
|---|---|---|---|---|---|---|---|---|---|
| | Estimate | SE | p-value | Estimate | SE | p-value | Estimate | SE | p-value |
| Intercept | 22.54 | 0.29 | | 22.26 | 0.37 | | 22.62 | 0.50 | |
| Sertraline (ref: placebo) | -0.48 | 0.39 | 0.22 | 0.15 | 0.50 | 0.77 | -0.32 | 0.67 | 0.63 |
| Hypericum (ref: placebo) | 0.65 | 0.42 | 0.13 | 0.67 | 0.55 | 0.23 | -1.01 | 0.70 | 0.15 |
| Week | -2.33 | 0.16 | <0.001 | -1.77 | 0.08 | <0.001 | -1.74 | 0.07 | <0.001 |
| Interaction week and treatment (hypericum compared to placebo) | 0.02 | 0.13 | 0.88 | 0.22 | 0.11 | 0.045 | -0.05 | 0.06 | 0.42 |
| Interaction week and treatment (sertraline compared to placebo) | -0.15 | 0.13 | 0.25 | -0.10 | 0.10 | 0.36 | -0.03 | 0.06 | 0.59 |
| Dropout | -0.25 | 0.56 | 0.66 | | | | 0.09 | 0.57 | 0.87 |
| Complete compared to dropout | | | | 0.30 | 0.62 | 0.63 | | | |
| Failure compared to dropout | | | | 0.86 | 0.57 | 0.13 | | | |
| Interaction week, dropout | 0.93 | 0.29 | 0.001 | | | | 0.40 | 0.09 | <0.001 |
| Complete compared to dropout | | | | -0.25 | 0.09 | 0.01 | | | |
| Failure compared to dropout | | | | 0.84 | 0.13 | <0.001 | | | |
| Week squared | 0.14 | 0.02 | <0.001 | 0.06 | 0.002 | <0.001 | 0.05 | 0.002 | <0.001 |
| Interaction hypericum, dropout | -0.17 | 0.80 | 0.83 | | | | 1.51 | 0.81 | 0.06 |
| Complete compared to dropout for hypericum compared to placebo | | | | -1.68 | 0.89 | 0.06 | | | |
| Failure compared to dropout for hypericum compared to placebo | | | | -0.29 | 0.83 | 0.73 | | | |
| Interaction sertraline, dropout | 1.36 | 0.75 | 0.07 | | | | 0.34 | 0.77 | 0.66 |
| Complete compared to dropout for sertraline compared to placebo | | | | -0.49 | 0.82 | 0.55 | | | |
| Failure compared to dropout for sertraline compared to placebo | | | | -0.19 | 0.80 | 0.81 | | | |
| Interaction hypericum, week, dropout | -0.22 | 0.41 | 0.59 | | | | 0.12 | 0.12 | 0.31 |
| Complete compared to dropout for hypericum compared to placebo | | | | -0.17 | 0.12 | 0.18 | | | |
| Failure compared to dropout for hypericum compared to placebo | | | | -0.30 | 0.19 | 0.11 | | | |
| Interaction sertraline, week, dropout | -0.80 | 0.42 | 0.06 | | | | -0.20 | 0.11 | 0.09 |
| Complete compared to dropout for sertraline compared to placebo | | | | 0.15 | 0.12 | 0.19 | | | |
| Failure compared to dropout for sertraline compared to placebo | | | | -0.12 | 0.18 | 0.51 | | | |

SE: Standard error

Figure 6.20 shows the fitted curves and the observed data for Model 3. Because values decrease sharply from baseline to Week 8 and level out after that, the quadratic model overestimates values slightly prior to Week 10 and fits well after Week 10 for completers. The fit for dropouts is not good, especially in the placebo arm. This did not improve if an interaction term between week squared and treatment was included.

Population estimates, averaging over the missing data patterns, were calculated for the fixed effects using the following formula: $\hat{\bar{\beta}} = \sum_{i=1}^{k} \hat{b}_i \hat{\pi}_i$ In Model 1 the sample proportions, $\hat{\pi}_i$, of completers and dropouts were 0.72 and 0.28, respectively. In Model 2 the sample proportion of completers was 0.24, the proportion of failures at Week 8 was 0.31 and the proportion of dropouts was 0.38. In Model 3 the sample proportions, $\hat{\pi}_i$, of completers and dropouts were 0.76 and 0.24, respectively. The average estimates for $\hat{\bar{\beta}}$ is given in Table 6.20. The corresponding standard errors were calculated using the methodology from section 3.3.2.1.

**Table 6.20: Population averaged estimates and standard errors of pattern mixture model using random-effects mixed model for Hamilton depression scale**

| | **Model 1 Binary dropout variable, 8 weeks** | | | **Model 2 Three dropout categories, 26 weeks** | | | **Model 3 Binary dropout variable for entire study, 26 weeks** | | |
|---|---|---|---|---|---|---|---|---|---|
| **Effect** | $\hat{\bar{\beta}}$ | **Standard error** | **p-value** | $\hat{\bar{\beta}}$ | **Standard error** | **p-value** | $\hat{\bar{\beta}}$ | **Standard error** | **p-value** |
| Intercept | 22.66 | 0.25 | <0.001 | 22.60 | 0.25 | <0.001 | 22.65 | 0.38 | <0.001 |
| Sertraline (ref: placebo) | -0.24 | 0.36 | 0.51 | -0.03 | 0.34 | 0.94 | -0.28 | 0.52 | 0.58 |
| Hypericum (ref: placebo) | 0.43 | 0.38 | 0.27 | 0.04 | 0.36 | 0.91 | -0.61 | 0.54 | 0.27 |
| Week | -1.38 | 0.19 | <0.001 | -1.70 | 0.15 | <0.001 | -1.66 | 0.08 | <0.001 |
| Interaction week and treatment (sertraline compared to placebo) | -0.38 | 0.15 | 0.01 | -0.10 | 0.07 | 0.17 | -0.08 | 0.05 | 0.12 |
| Interaction week and treatment (hypericum compared to placebo) | -0.002 | 0.15 | 0.99 | 0.09 | 0.07 | 0.18 | -0.02 | 0.05 | 0.68 |
| Week squared | | | | 0.06 | 0.002 | <0.001 | 0.05 | 0.002 | <0.001 |

According to Model 1, which uses data up to Week 8, the improvement in the sertraline arm was significantly different from the improvement in the placebo arm over time, whereas the improvement in the hypericum arm over time was not statistically significantly different from the improvement in the placebo arm. Models 2 and 3, which used data up to Week 26 did not find either treatment arm to lead to improved depression scores over time compared to placebo. In this analysis, HAM-D scores after dropout were extrapolated. These extrapolations and criticism against these are discussed in Section 5.5.1.1.

### 6.5.1.2 Pattern-mixture models using multiple imputation

We implemented pattern-mixture models using multiple imputations as described in Section 5.5.1.2. Different values of $\delta$, which indicates the association between HAM-D score and dropout, were assumed for participants who dropped out and those who completed the study. Imputations were restricted to be between 0 and 33, which are valid scores on the HAM-D questionnaire.

**Table 6.21: Multiple imputation of HAM-D scores using a pattern-mixture approach, data up to Week 8**

| | Sertraline arm | | Hypericum arm | | Placebo arm | | p-value* |
|---|---|---|---|---|---|---|---|
| | **Mean** | **Standard error** | **Mean** | **Standard error** | **Mean** | **Standard error** | |
| **Assuming decrease in HAM-D score for participants who dropped out** | | | | | | | |
| **Model 1: δ for dropout = 0.5, σ = 1** | | | | | | | |
| **Week 1** | 19.5 | 0.53 | 20.0 | 0.51 | 18.9 | 0.47 | 0.34 |
| **Week 2** | 16.2 | 0.74 | 18.2 | 0.67 | 17.6 | 0.57 | 0.15 |
| **Week 3** | 14.5 | 1.47 | 17.0 | 1.11 | 17.2 | 0.87 | 0.14 |
| **Week 4** | 12.7 | 2.51 | 15.6 | 1.93 | 15.3 | 1.52 | 0.40 |
| **Week 6** | 10.5 | 4.33 | 13.8 | 3.26 | 13.9 | 2.79 | 0.53 |
| **Week 7** | 7.4 | 6.84 | 11.2 | 5.36 | 11.1 | 4.99 | 0.68 |
| **Week 8** | 4.5 | 10.09 | 8.1 | 8.17 | 10.0 | 7.82 | 0.40 |
| **Model 2: δ for dropout = 0.25, σ = 1** | | | | | | | |
| **Week 1** | 19.6 | 0.53 | 20.0 | 0.51 | 18.9 | 0.46 | 0.32 |
| **Week 2** | 16.4 | 0.74 | 18.3 | 0.68 | 17.6 | 0.56 | 0.20 |
| **Week 3** | 15.1 | 1.55 | 17.3 | 1.16 | 17.2 | 0.87 | 0.25 |
| **Week 4** | 13.8 | 2.66 | 16.2 | 2.01 | 15.5 | 1.50 | 0.60 |
| **Week 6** | 12.4 | 4.53 | 14.8 | 3.39 | 14.5 | 2.78 | 0.72 |
| **Week 7** | 10.4 | 7.23 | 12.8 | 5.55 | 12.0 | 4.93 | 0.87 |
| **Week 8** | 9.0 | 10.67 | 10.5 | 8.55 | 11.6 | 7.85 | 0.60 |
| **Assuming increase in HAM-D score for participants who dropped out** | | | | | | | |
| **Model 3: δ for dropout = -0.5, σ = 1** | | | | | | | |
| **Week 1** | 19.6 | 0.53 | 20.1 | 0.51 | 18.9 | 0.47 | 0.30 |
| **Week 2** | 16.7 | 0.74 | 18.6 | 0.67 | 17.8 | 0.56 | 0.29 |
| **Week 3** | 15.9 | 1.55 | 18.0 | 1.15 | 17.6 | 0.73 | 0.38 |
| **Week 4** | 15.5 | 2.67 | 17.5 | 2.01 | 16.2 | 0.95 | 0.78 |
| **Week 6** | 15.5 | 4.55 | 17.1 | 3.40 | 15.6 | 1.40 | 0.88 |
| **Week 7** | 15.3 | 7.26 | 16.7 | 5.55 | 13.9 | 2.20 | 0.99 |
| **Week 8** | 16.2 | 10.71 | 16.5 | 8.56 | 14.4 | 3.10 | 0.78 |
| **Model 4: δ for dropout = -0.25, σ = 1** | | | | | | | |
| **Week 1** | 19.6 | 0.53 | 20.1 | 0.51 | 18.9 | 0.46 | 0.31 |
| **Week 2** | 16.6 | 0.74 | 18.5 | 0.67 | 17.7 | 0.56 | 0.26 |
| **Week 3** | 15.6 | 1.55 | 17.8 | 1.15 | 17.5 | 0.87 | 0.33 |
| **Week 4** | 14.9 | 2.66 | 17.1 | 2.0 | 16.1 | 1.50 | 0.72 |
| **Week 6** | 14.5 | 4.54 | 16.3 | 3.39 | 15.8 | 2.78 | 0.83 |
| **Week 7** | 13.7 | 7.24 | 15.4 | 5.54 | 14.3 | 4.94 | 0.95 |
| **Week 8** | 13.8 | 10.68 | 14.5 | 8.54 | 15.2 | 7.86 | 0.72 |

Means and standard errors calculated from the multiple imputed values for each of these models are given using data up to Week 8 (Table 6.21) and using data up to Week 26 (Table 6.22). The cross-sectional results are not to be compared to the longitudinal model, but the means over time are helpful in describing the implications of the different assumptions made to non-statisticians. The cross-sectional summaries allow one to illustrate whether an assumption implies increasing or decreasing means. Models 1 and 2 assume a decrease in HAM-D score for participants who

dropped out, while Models 3 and 4 assume an increase in HAM-D score for participants who dropped out.

None of the four models fitted with data up to Week 8 showed a significant effect of either treatment arm over placebo at any of the weeks. Neither assuming that participants who dropped out improved or that participants who dropped out worsened, led to the conclusion that there is a significant effect of treatment at any visit.

When the data up to Week 26 were fitted, this method used a lot of computing power and was slow. The multiple imputations were fast to do, but fitting each of the mixed models took time. Since 50 models had to be fitted, it took almost 2 hours to run each of the models on a standard computer. The models fitted with the data up to Week 8 required less computer resources to be fitted.

The models fitted using all data up to Week 26 were more problematic than the models fitted using the data up to Week 8 (Table 6.22). The models fitted lead to negative estimates of HAM-D scores, which are not possible, from Week 14, onwards. This was corrected by truncating the estimated values at 0 and 33.

The problem with this method of handling MNAR missing data, is that the decrease in HAM-D score assumed after drop out accumulates over time and therefore the assumption made about $\delta$ influences the later weeks post dropout much more than the earlier weeks. From Week 10 onwards more values are missing than observed, meaning that the assumption made about the value of $\delta$ plays a larger role in the results than the actual data observed for the later visits.

Models 5 and 6 assumed that HAM-D scores decreased after dropout. This is a fairly illogical assumption to make, since a substantial number of the dropouts at Week 8 are because of lack of response. If it is not plausible to assume that HAM-D scores will decrease after dropout, it is not helpful to fit such a model. This model is included here for comparison purposes, but should not be used to analyse this data. The standard errors are also very large for the later weeks. This means that variability is increased over time, making all estimates less reliable. This is due to the fact that so much data are missing.

**Table 6.22: Multiple imputation of HAM-D scores using a pattern-mixture approach, data up to Week 26**

| | Sertraline arm | | Hypericum arm | | Placebo arm | | p-value* |
|---|---|---|---|---|---|---|---|
| | Mean | Standard error | Mean | Standard error | Mean | Standard error | |
| **Assuming decrease (improvement) in HAM-D score for participants who dropped out** | | | | | | | |
| **Model 5: δ for dropout = 0.1, σ = 1** | | | | | | | |
| **Week 1** | 19.5 | 0.52 | 20.0 | 0.51 | 18.9 | 0.46 | 0.32 |
| **Week 2** | 16.5 | 0.71 | 18.4 | 0.68 | 17.6 | 0.56 | 0.23 |
| **Week 3** | 15.2 | 1.28 | 17.5 | 1.01 | 17.3 | 0.83 | 0.20 |
| **Week 4** | 14.3 | 1.91 | 16.6 | 1.42 | 15.8 | 1.21 | 0.53 |
| **Week 6** | 13.4 | 2.52 | 15.7 | 1.88 | 15.2 | 1.87 | 0.61 |
| **Week 7** | 12.1 | 3.18 | 14.5 | 2.57 | 13.3 | 2.75 | 0.81 |
| **Week 8** | 11.6 | 3.68 | 13.3 | 3.20 | 13.4 | 3.52 | 0.76 |
| **Week 10** | 10.6 | 5.89 | 13.9 | 6.47 | 13.7 | 6.67 | 0.76 |
| **Week 14** | 10.4 | 7.36 | 12.3 | 8.27 | 14.1 | 8.59 | 0.77 |
| **Week 18** | 10.9 | 8.55 | 13.8 | 9.84 | 12.5 | 9.99 | 0.92 |
| **Week 22** | 12.5 | 9.50 | 13.2 | 10.61 | 14.1 | 10.79 | 0.92 |
| **Week 26** | 12.3 | 10.24 | 14.1 | 11.09 | 14.8 | 11.26 | 0.89 |
| **Model 6: δ for dropout = 0.25, σ = 1** | | | | | | | |
| **Week 1** | 19.5 | 0.52 | 20.0 | 0.51 | 18.8 | 0.46 | 0.32 |
| **Week 2** | 16.5 | 0.71 | 18.3 | 0.68 | 17.6 | 0.56 | 0.21 |
| **Week 3** | 15.1 | 1.27 | 17.4 | 1.00 | 17.2 | 0.83 | 0.18 |
| **Week 4** | 14.0 | 1.89 | 16.4 | 1.38 | 15.7 | 1.19 | 0.50 |
| **Week 6** | 13.0 | 2.49 | 15.4 | 1.82 | 14.9 | 1.83 | 0.58 |
| **Week 7** | 11.7 | 3.13 | 14.1 | 2.47 | 13.0 | 2.69 | 0.79 |
| **Week 8** | 11.1 | 3.64 | 12.8 | 3.07 | 13.0 | 3.48 | 0.75 |
| **Week 10** | 9.8 | 5.79 | 12.9 | 6.29 | 12.9 | 6.60 | 0.76 |
| **Week 14** | 9.5 | 7.19 | 11.0 | 7.86 | 13.0 | 8.49 | 0.78 |
| **Week 18** | 9.8 | 8.40 | 12.5 | 9.55 | 11.3 | 9.80 | 0.92 |
| **Week 22** | 11.2 | 9.38 | 11.6 | 10.26 | 12.9 | 10.8 | 0.91 |
| **Week 26** | 10.9 | 10.13 | 12.4 | 10.73 | 13.4 | 11.25 | 0.88 |
| **Assuming increase (worsening) in HAM-D score for participants who dropped out** | | | | | | | |
| **Model 7: δ for dropout = -0.05, σ = 0.5** | | | | | | | |
| **Week 1** | 19.5 | 0.52 | 20.1 | 0.50 | 18.9 | 0.46 | 0.32 |
| **Week 2** | 16.6 | 0.60 | 18.4 | 0.59 | 17.7 | 0.54 | 0.19 |
| **Week 3** | 15.4 | 0.86 | 17.6 | 0.78 | 17.3 | 0.70 | 0.08 |
| **Week 4** | 14.4 | 1.34 | 16.7 | 1.09 | 15.9 | 0.87 | 0.40 |
| **Week 6** | 13.6 | 1.90 | 15.9 | 1.42 | 15.2 | 1.34 | 0.51 |
| **Week 7** | 12.4 | 2.49 | 14.7 | 1.97 | 13.4 | 1.98 | 0.79 |
| **Week 8** | 11.9 | 3.04 | 13.6 | 2.55 | 13.6 | 2.71 | 0.72 |
| **Week 10** | 11.0 | 4.63 | 14.8 | 4.69 | 14.2 | 4.92 | 0.68 |
| **Week 14** | 10.5 | 5.92 | 12.8 | 6.59 | 14.6 | 6.78 | 0.68 |
| **Week 18** | 11.4 | 7.32 | 15.1 | 8.17 | 12.1 | 8.33 | 0.96 |
| **Week 22** | 13.6 | 8.28 | 14.0 | 9.38 | 14.7 | 9.44 | 0.94 |
| **Week 26** | 13.5 | 9.25 | 15.6 | 9.86 | 15.9 | 10.2 | 0.88 |
| **Model 8: δ for dropout = -0.25, σ = 0.5** | | | | | | | |
| **Week 1** | 19.6 | 0.52 | 20.1 | 0.50 | 18.9 | 0.46 | 0.31 |
| **Week 2** | 16.7 | 0.60 | 18.5 | 0.59 | 17.7 | 0.54 | 0.22 |
| **Week 3** | 15.6 | 0.86 | 17.8 | 0.77 | 17.4 | 0.70 | 0.10 |
| **Week 4** | 14.8 | 1.33 | 17.1 | 1.09 | 16.1 | 0.87 | 0.45 |
| **Week 6** | 14.2 | 1.91 | 16.4 | 1.45 | 15.6 | 1.35 | 0.57 |
| **Week 7** | 13.3 | 2.50 | 15.5 | 2.01 | 14.0 | 2.00 | 0.83 |
| **Week 8** | 13.0 | 3.05 | 14.5 | 2.62 | 14.4 | 2.70 | 0.75 |
| **Week 10** | 12.7 | 4.63 | 16.5 | 4.69 | 15.7 | 4.87 | 0.69 |
| **Week 14** | 12.8 | 6.01 | 15.3 | 6.85 | 16.7 | 6.64 | 0.68 |
| **Week 18** | 14.2 | 7.34 | 17.9 | 7.93 | 15.0 | 8.40 | 0.95 |
| **Week 22** | 16.6 | 8.17 | 17.4 | 9.44 | 17.6 | 9.23 | 0.94 |
| **Week 26** | 17.0 | 9.16 | 18.8 | 9.46 | 18.8 | 9.57 | 0.90 |

**Table 6.23: Multiple imputation of HAM-D scores using a pattern-mixture approach, results of mixed model**

| | Data up to 8 weeks Model 1 δ for dropout = 0.5 | | | Data up to 8 weeks Model 2 δ for dropout = 0.25 | | | Data up to 8 weeks Model 3 δ for dropout = -0.5 | | |
|---|---|---|---|---|---|---|---|---|---|
| **Parameter** | **Estimate** | **SE** | **p-value** | **Estimate** | **SE** | **p-value** | **Estimate** | **SE** | **p-value** |
| Intercept | 22.48 | 0.25 | | 22.58 | 0.25 | | 22.74 | 0.28 | |
| Sertraline (ref: placebo) | -0.11 | 0.33 | 0.75 | 0.46 | 0.34 | 0.18 | -0.12 | 0.43 | 0.78 |
| Hypericum (ref: placebo) | 0.77 | 0.37 | 0.04 | 0.73 | 0.36 | 0.05 | 0.51 | 0.39 | 0.19 |
| Week | -1.77 | 0.18 | <0.001 | -1.98 | 0.21 | <0.001 | -1.33 | 0.29 | <0.001 |
| Interaction week and treatment (sertraline compared to placebo) | -0.44 | 0.26 | 0.09 | -0.63 | 0.23 | 0.01 | -0.30 | 0.52 | 0.56 |
| Interaction week and treatment (hypericum compared to placebo) | 0.15 | 0.23 | 0.51 | 0.34 | 0.28 | 0.23 | 0.08 | 0.41 | 0.85 |

| | Data up to 8 weeks Model 4 δ for dropout = -0.25 | | | Data up to 26 weeks Model 5 δ for dropout = 0.1 | | | Data up to 26 weeks Model 6 δ for dropout = 0.25 | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Estimate** | **SE** | **p-value** | **Estimate** | **SE** | **p-value** | **Estimate** | **SE** | **p-value** |
| Intercept | 22.72 | 0.28 | | 22.43 | 0.54 | | 22.41 | 0.53 | |
| Sertraline (ref: placebo) | -0.11 | 0.44 | 0.80 | -0.01 | 0.72 | 0.98 | -0.01 | 0.68 | 0.99 |
| Hypericum (ref: placebo) | 0.52 | 0.39 | 0.19 | 0.50 | 0.77 | 0.52 | 0.48 | 0.81 | 0.55 |
| Week | -1.33 | 0.29 | <0.001 | -1.42 | 0.44 | 0.002 | -1.47 | 0.42 | 0.001 |
| Interaction week and treatment (sertraline compared to placebo) | -0.32 | 0.54 | 0.55 | -0.07 | 0.53 | 0.90 | -0.08 | 0.52 | 0.88 |
| Interaction week and treatment (hypericum compared to placebo) | 0.07 | 0.41 | 0.86 | -0.004 | 0.45 | 0.99 | -0.02 | 0.47 | 0.96 |
| Week squared | | | | 0.04 | 0.01 | <0.001 | 0.04 | 0.005 | <0.001 |

| | Data up to 26 weeks Model 7 δ for dropout = -0.05 σ = 0.5 | | | Data up to 26 weeks Model 8 δ for dropout = -0.25 σ = 0.5 | | |
|---|---|---|---|---|---|---|
| | **Estimate** | **SE** | **p-value** | **Estimate** | **SE** | **p-value** |
| Intercept | 22.46 | 0.57 | | 22.51 | 0.59 | |
| Sertraline (ref: placebo) | -0.01 | 0.67 | 0.98 | -0.02 | 0.64 | 0.98 |
| Hypericum (ref: placebo) | 0.26 | 0.84 | 0.76 | 0.30 | 0.86 | 0.73 |
| Week | -1.42 | 0.39 | 0.001 | -1.24 | 0.44 | 0.01 |
| Interaction week and treatment (sertraline compared to placebo) | -0.06 | 0.51 | 0.90 | -0.05 | 0.46 | 0.91 |
| Interaction week and treatment (hypericum compared to placebo) | 0.05 | 0.43 | 0.90 | 0.07 | 0.39 | 0.87 |
| Week squared | 0.04 | 0.01 | <0.001 | 0.04 | 0.01 | <0.001 |

SE: Standard error

None of the models fitted up to Week 8, neither those assuming an improvement in HAM-D scores post drop-out nor those assuming a worsening in HAM-D scores post drop, had a significant interaction between week and the hypericum treatment arm compared to the placebo arm. Model 2, which assumed a small decrease in HAM-D score in participants who dropped out had a significant interaction between week and the sertraline arm compared to the placebo arm. All the models fitted up to Week 26 found no significant effect of treatment arm.

These models used multiple imputation to impute missing HAM-D scores. The imputed values were changed according to the pattern assumed. Differences between these results and the models fitted under multiple imputation can be attributed to the assumptions made about the unobserved data.

#### 6.5.1.3 Pattern-mixture models using identifying restrictions

Multiple imputation was used to create a data set that included only monotone missingness before the identifying restrictions were utilised to impute the missing data. We imputed 70 data sets using each of the identifying restrictions. A model was fitted to the pattern-specific identifiable densities. Estimates were calculated over all patterns where the required components were observed, and the conditional distributions of the unobserved outcomes given the observed ones were calculated. This is described in more detail in Section 5.5.1.3.

Imputed data were truncated at 0 and 33 respectively to ensure that only valid HAM-D scores were generated. If the truncation was not done some of the HAM-D scores imputed were less than 0, especially at the later weeks. Only a small fraction of the imputed values were affected by this. After the imputations using identifying restrictions, the data were analysed using the same methods as with multiple imputation.

**Table 6.24: The number of participants in each pattern**

| Week | Sertraline | | Hypericum | | Placebo | |
|---|---|---|---|---|---|---|
| | Up to Week 8 | Up to Week 26 | Up to Week 8 | Up to Week 26 | Up to Week 8 | Up to Week 26 |
| Baseline | 7 | 7 | 9 | 9 | 4 | 4 |
| 1 | 10 | 10 | 2 | 2 | 1 | 1 |
| 2 | 2 | 2 | 2 | 2 | 6 | 6 |
| 3 | 2 | 2 | 3 | 3 | 5 | 5 |
| 4 | 7 | 7 | 5 | 5 | 6 | 6 |
| 6 | 4 | 4 | 10 | 10 | 10 | 10 |
| 8 | 79 | 35 | 82 | 49 | 84 | 47 |
| 14 | | 5 | | 6 | | 5 |
| 18 | | 8 | | 3 | | 5 |
| 26 | | 31 | | 24 | | 27 |

Small numbers of participants had a last visit at Weeks 7, 10 and 22. This meant that those patterns were sparse and the models could not be fitted for those weeks. This was circumvented in the analysis by combining these time points with other time points. This means for example that a participant who had a last visit at Week 7 was combined in the same pattern with participants who had a last visit at Week 6 (Table 6.24).

This model differed from the model using multiple imputation in Section 6.4.2 in the variables included during imputation. In the previous models, several baseline covariates were included in the imputers' model. In this imputer's model only the observed outcomes are included. Because the analysis and imputation is done by pattern, it is more elaborate than previous models. Where patterns are sparse, the imputation and modeling in some of the patterns may not be efficient.

**Table 6.25: HAM-D scores over time, pattern-mixture model with identifying restrictions using data up to Week 8**

| | ACMV | | | CCMV | | | NCMV | | |
|---|---|---|---|---|---|---|---|---|---|
| **Effect** | **Estimate** | **SE** | **p-value** | **Estimate** | **SE** | **p-value** | **Estimate** | **SE** | **p-value** |
| Intercept | 22.28 | 0.18 | | 22.26 | 0.18 | | 22.31 | 0.19 | HA |
| Sertraline (ref: placebo) | 0.08 | 0.26 | 0.75 | 0.11 | 0.26 | 0.67 | 0.11 | 0.27 | 0.68 |
| Hypericum (ref: placebo) | 0.31 | 0.25 | 0.22 | 0.31 | 0.25 | 0.22 | 0.34 | 0.26 | 0.18 |
| Week | -1.15 | 0.09 | <0.001 | -1.14 | 0.08 | <0.001 | -1.09 | 0.14 | <0.001 |
| Interaction week and treatment (sertraline compared to placebo) | 0.08 | 0.12 | 0.53 | 0.08 | 0.12 | 0.47 | -0.16 | 0.18 | 0.38 |
| Interaction week and treatment (hypericum compared to placebo) | 0.13 | 0.11 | 0.24 | 0.13 | 0.10 | 0.21 | 0.03 | 0.14 | 0.82 |

SE: Standard error

Under both the analysis up to Week 8 and Week 26 we concluded that there was no significant interaction of week and either treatment arm compared to placebo using the ACMV model. The pattern-mixture model using ACMV identifying restrictions gave different results from the other MAR methods; for both the analysis using data up to Week 8 and using data up to Week 26. This method gave less reliable results because some of the patterns were sparse, as can be seen in Table 6.24, and we based our conclusions on the other methods used.

The ACMV model, which is valid under MAR assumptions, and the CCMV model, which is valid under MNAR assumptions, had similar results. If most participants completed, as is the case here, then borrowing information from the completers is not very different from borrowing information from the available cases. The NCMV identifying restriction gave different results (Table 6.25), because information was borrowed from the closest neighbour. Comparing the ACMV results to the CCMV and NCMV results indicates how results differ if MNAR assumptions rather than MAR assumptions are made. As can be seen from Figure 6.21, all the treatment arms followed a similar pattern over time. The NCMV model estimated a lower HAM-D score for each of the treatment arms than the other identifying restrictions, while the models using the ACMV and CCMV identifying restrictions were often superimposed on one another.

With both the MNAR models, namely the CCMV and the NCMV models, we conclude that there is no difference between the sertraline and placebo arm or between the hypericum and placebo arm in the change in mean HAM-D score over time.

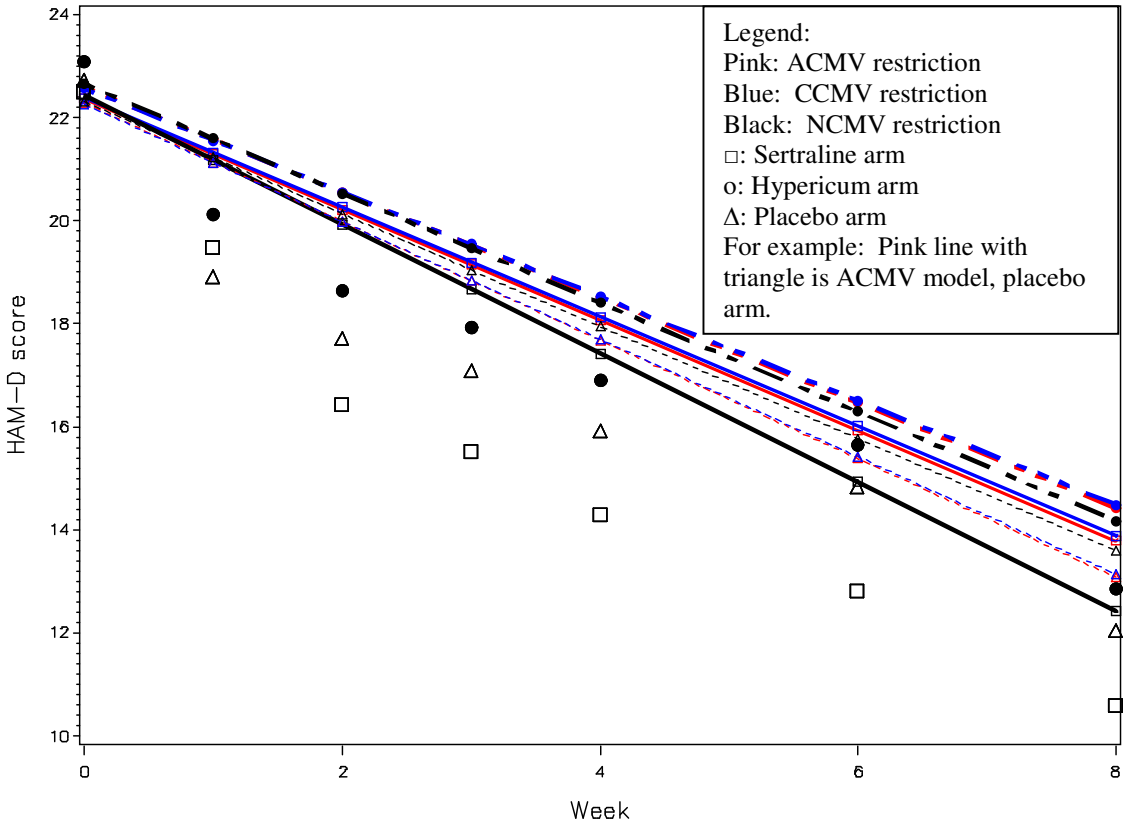



**Figure 6.21: Models fitted using identifying restrictions, HAM-D scores over the first 8 weeks**

**Table 6.26: HAM-D score over time, pattern-mixture model with identifying restrictions using data up to Week 26**

| | ACMV | | | CCMV | | | NCMV | | |
|---|---|---|---|---|---|---|---|---|---|
| **Effect** | **Estimate** | **SE** | **p-value** | **Estimate** | **SE** | **p-value** | **Estimate** | **SE** | **p-value** |
| Intercept | 22.10 | 0.40 | | 22.20 | 0.40 | | 22.19 | 0.35 | |
| Sertraline (ref: placebo) | 0.49 | 0.63 | 0.44 | 0.52 | 0.63 | 0.42 | 0.32 | 0.49 | 0.51 |
| Hypericum (ref: placebo) | 0.45 | 0.44 | 0.31 | 0.41 | 0.43 | 0.35 | 0.34 | 0.43 | 0.44 |
| Week | -1.11 | 0.20 | <0.001 | -1.11 | 0.18 | <0.001 | -1.28 | 0.31 | <0.001 |
| Interaction week and treatment (sertraline compared to placebo) | 0.18 | 0.13 | 0.17 | 0.16 | 0.11 | 0.16 | 0.05 | 0.19 | 0.78 |
| Interaction week and treatment (hypericum compared to placebo) | 0.12 | 0.10 | 0.25 | 0.11 | 0.09 | 0.24 | -0.01 | 0.14 | 0.94 |
| Week squared | 0.02 | 0.01 | 0.01 | 0.02 | 0.01 | 0.007 | 0.03 | 0.01 | 0.003 |

SE: Standard error

Using all the data up to Week 26, also lead to the conclusion that there was no significant interaction between week and either the sertraline or hypericum treatment arm compared to placebo (Table 6.26). For the sertraline and hypericum arms lower mean HAM-D scores were obtained

with the NCMV identifying restrictions than with the other restrictions. In the placebo arm the three different models had similar mean HAM-D scores until Week 18, after which the NCMV identifying restriction model had higher mean HAM-D scores than the other two models (Figure 6.22).

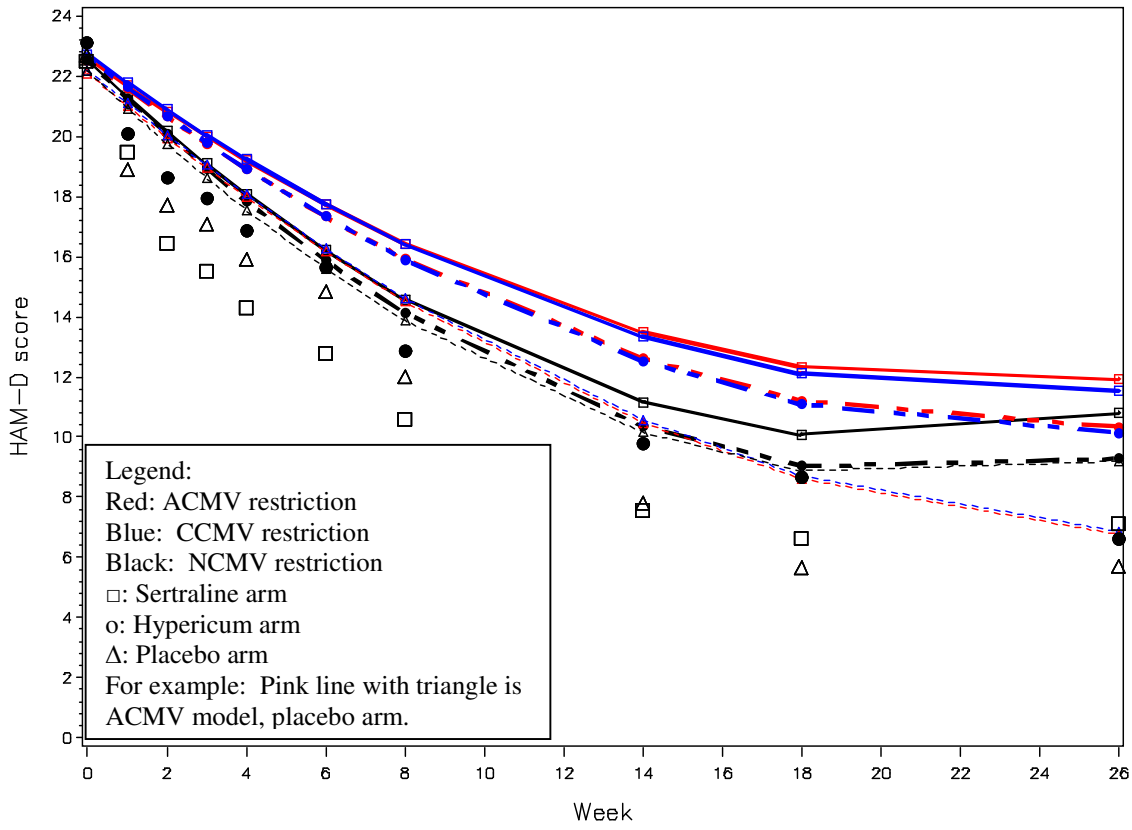



**Figure 6.22: Models fitted using identifying restrictions, HAM-D scores over 26 weeks**

## 6.5.2 Selection models

Selection models were fitted in a Bayesian framework under several different assumptions about the missing data mechanism. The reason for using Bayesian methods to fit the selection models was discussed in Section 5.5.2. The model for the observed data was the same as the model fitted under MAR. The same uninformative prior distributions were used as in the MAR case. The model and prior distributions are given in Section 6.4.3 and are summarised in Table 6.27.

In addition, a model for missingness was added, taking the form

$$m_{ij} \sim \text{Bernoulli}(p_{ij}),$$

$$\text{logit}(p_{ij}) = \theta_0 + \Delta y_{ij}$$

where $m_{ij}$ is a binary missing value indicator for $y_{ij}$. This model allowed the missingness to depend on the value that would have been observed. A flat prior was specified on the scale of $p_{ij}$ by specifying a logistic(0, 1) prior for $\theta_0$ in all sets. The prior distributions specified for $\Delta$ were:

Prior Set 1: $\Delta$ ~ Normal(0, 10 000) distribution

Prior Set 2: $\Delta$ ~ Normal(1, 100)

Prior Set 3: $\Delta$ ~ Normal(1, 10).

A second MNAR model was fitted where the model of missingness had the following form:

$$logit\, p_{ij} = \theta_0 + \theta_1 y_{i,j-1} + \theta_2 (y_{ij} - y_{i,j-1}).$$

This model allowed the missingness to depend on the previous HAM-D score and the change in HAM-D score from the previous visit to the one where the missing data occurred. The prior distributions assigned were:

Prior Set 1: $\theta_0$ ~ logistic(0,1) and $\theta_1, \theta_2$ ~ Normal(0, 10 000)

Prior Set 2: $\theta_0, \theta_1, \theta_2$ ~ Normal(0, 100)

Prior Set 3: $\theta_0, \theta_1, \theta_2$ ~Normal(0, 10)

A third MNAR model was fitted where the model of missingness had the following form:

$$logit\, p_{ij} = \theta_0 + \theta_1 y_{ij} + \theta_2 y_{ij} S_{ij} + \theta_3 y_{ij} H_{ij}$$

This model allowed the missingness to depend on the HAM-D score that would have been observed, while allowing for a different mechanism in each treatment arm by including the HAM-D score by treatment arm interaction. The prior sets were:

Prior Set 1: $\theta_0$~ logistic(0,1), $\theta_1$, $\theta_2$ , $\theta_3$ ~ Normal(0, 10 000)

Prior Set 2: $\theta_0$, $\theta_1$, $\theta_2$ , $\theta_3$ ~ Normal(1, 100)

Prior Set 3: $\theta_0$, $\theta_1$, $\theta_2$ , $\theta_3$ ~ Normal(1, 10)

Many other models for non-random missingness could be fitted, depending on the assumptions made regarding the missing data mechanism.

**Table 6.27: Hamilton depression score posterior means, standard deviations and credible intervals according to Bayesian analysis for data up to Week 8**

| | **Substantive model: Prior Set 1** β ~ Normal(0, 10 000), Ω~Wishart(I,5) I: Identity matrix, | | | **Substantive model: Prior Set 2** β ~ Normal(0, 1000), Ω~Wishart(A,5) A: diag(0.1, 0.1, 0.1, 0.1, 0.1 | | | **Substantive model: Prior Set 3** β ~ Normal(0, 10), Ω~Wishart(A,5) A: diag(10, 10, 10, 10, 10), | | |
|---|---|---|---|---|---|---|---|---|---|
| **MNAR model 1:** $m_{ij} \sim$ Bernoulli$(p_{ij})$, logit$(p_{ij}) = \theta_0 + \Delta y_{ij}$ | | | | | | | | | |
| | **MNAR model 1: Prior Set 1** $\theta_0$ ~ logistic(0,1), Δ ~ Normal(0, 10 000) | | | **MNAR model 1: Prior Set 2** $\theta_0$ ~ logistic(0,1), Δ ~ Normal(0, 100) | | | **MNAR model 1: Prior Set 3** $\theta_0$ ~ logistic(0,1), Δ ~ Normal(0, 10) | | |
| **Effect** | **Mean** | **SD** | **95% credible interval** | **Mean** | **SD** | **95% credible interval** | **Mean** | **SD** | **95% credible interval** |
| $\beta_0$ | 22.55 | 0.29 | 21.98; 23.13 | 22.55 | 0.29 | 21.98; 23.13 | 22.36 | 0.29 | 21,78; 22.93 |
| $\beta_1$ | 0.20 | 0.42 | -0.62; 1.02 | 0.20 | 0.42 | -0.62; 1.02 | 0.38 | 0.42 | -0.46; 1.21 |
| $\beta_2$ | 0.57 | 0.41 | -0.22;1.37 | 0.57 | 0.41 | -0.22;1.37 | 0.75 | 0.40 | -0.03; 1.54 |
| $\beta_3$ | -1.13 | 0.11 | -1.35; -0.90 | -1.12 | 0.11 | -1.35; -0.90 | -1.13 | 0.11 | -1.36; -0.90 |
| $\beta_4$ | -0.28 | 0.15 | -0.58; 0.02 | -0.28 | 0.15 | -0.58; 0.02 | -0.27 | 0.15 | -0.57;0.03 |
| $\beta_5$ | -0.01 | 0.15 | -0.30; 0.29 | -0.01 | 0.15 | -0.30; 0.29 | -0.004 | 0.15 | -0.30; 0.29 |
| Δ | 0.03 | 0.03 | -0.02; 0.08 | 0.03 | 0.03 | -0.02; 0.08 | 0.03 | 0.03 | -0.02; 0.08 |
| $\theta_0$ | -4.41 | 0.51 | -5.47; -3.46 | -4.41 | 0.51 | -5.47; -3.46 | -4.40 | 0.51 | -5.45; -3.45 |
| **MNAR model 2:** $m_{ij} \sim$ Bernoulli$(p_{ij})$, $logit\ p_{ij} = \theta_0 + \theta_1 y_{i,j-1} + \theta_2(y_{ij} - y_{i,j-1})$ | | | | | | | | | |
| | **MNAR model 2: Prior Set1** $\theta_0$ ~ logistic(0,1), $\theta_1, \theta_2$ ~ Normal(0, 10 000) | | | **MNAR model 2: Prior Set 2** $\theta_0, \theta_1, \theta_2$ ~ Normal(0, 100) | | | **MNAR model 2: Prior Set 3** $\theta_0, \theta_1, \theta_2$ ~Normal(0, 10) | | |
| $\beta_0$ | 22.6 | 0.31 | 22.0; 23.2 | 22.6 | 0.31 | 22.0; 23.2 | 22.4 | 0.30 | 21.8; 22.98 |
| $\beta_1$ | 0.15 | 0.42 | -0.68; 0.97 | 0.15 | 0.42 | -0.68; 0.97 | 0.33 | 0.42 | -0.48; 1.17 |
| $\beta_2$ | 0.49 | 0.41 | -0.32; 1.29 | 0.49 | 0.41 | -0.32; 1.30 | 0.67 | 0.41 | -0.13; 1.46 |
| $\beta_3$ | -1.10 | 0.12 | -1.33; -0.88 | -1.10 | 0.12 | -1.33; -0.88 | -1.1 | 0.12 | -1.33; -0.87 |
| $\beta_4$ | -0.33 | 0.15 | -0.63;-0.03 | -0.33 | 0.15 | -0.63;-0.03 | -0.33 | 0.15 | -0.63;-0.03 |
| $\beta_5$ | -0.03 | 0.16 | -0.33; 0.28 | -0.02 | 0.15 | -0.32; 0.27 | -0.02 | 0.15 | -0.32; 0.28 |
| $\theta_0$ | -3.96 | 0.42 | -4.82; -3.15 | -4.12 | 0.43 | -4.99; -3.30 | -4.05 | 0.43 | -4.93; -3.23 |
| $\theta_1$ | 0.04 | 0.02 | 0.003; 0.09 | 0.05 | 0.02 | 0.01; 0.10 | 0.05 | 0.02 | 0.01; 0.09 |
| $\theta_2$ | 0.01 | 0.03 | -0.05; 0.06 | 0.01 | 0.03 | -0.05; 0.06 | 0.01 | 0.03 | -0.04; 0.06 |
| **MNAR model 3:** $logit\ p_{ij} = \theta_0 + \theta_1 y_{ij} + \theta_2 y_{ij} S_{ij} + \theta_3 y_{ij} H_{ij}$ | | | | | | | | | |
| | **MNAR model 3: Prior Set1** $\theta_0$ ~ logistic(0,1), $\theta_1, \theta_2, \theta_3$~Normal(0,10 000) | | | **MNAR model 3: Prior Set 2** $\theta_0$ ~ logistic(0,1), $\theta_1, \theta_2, \theta_3$~Normal(0, 100) | | | **MNAR model 3: Prior Set 3** $\theta_0$ ~ logistic(0,1), $\theta_1, \theta_2, \theta_3$ ~ Normal(0, 10) | | |
| $\beta_0$ | 22.59 | 0.30 | 21.99; 23.19 | 22.58 | 0.30 | 21.99; 23.15 | 22.38 | 0.29 | 21.8; 22.95 |
| $\beta_1$ | 0.14 | 0.43 | -0.70; 0.99 | 0.16 | 0.42 | -0.65; 0.98 | 0.33 | 0.41 | -0.47; 1.13 |
| $\beta_2$ | 0.49 | 0.41 | -0.31; 1.30 | 0.51 | 0.40 | -0.27; 1.31 | 0.68 | 0.40 | -0.10; 1.48 |
| $\beta_3$ | -1.12 | 0.12 | -1.35; -0.88 | -1.11 | 0.12 | -1.35; -0.88 | -1.12 | 0.12 | -1.35; -0.88 |
| $\beta_4$ | -0.31 | 0.16 | -0.63; -0.01 | -0.32 | 0.16 | -0.62; -0.01 | -0.31 | 0.16 | -0.62; 0.004 |
| $\beta_5$ | -0.01 | 0.15 | -0.32; 0.29 | -0.01 | 0.16 | -0.32; 0.30 | -0.01 | 0.15 | -0.31; 0.29 |
| $\theta_0$ | -4.19 | 0.46 | -5.14; -3.36 | -4.26 | 0.48 | -5.25; -3.39 | -4.27 | 0.49 | -5.31; -3.36 |
| $\theta_1$ | 0.03 | 0.03 | -0.02; 0.08 | 0.03 | 0.03 | -0.01; 0.09 | 0.04 | 0.03 | -0.02; 0.09 |
| $\theta_2$ | -0.01 | 0.02 | -0.04; 0.03 | -0.01 | 0.02 | -0.04; 0.03 | -0.01 | 0.02 | -0.05; 0.03 |
| $\theta_3$ | -0.04 | 0.02 | -0.09; 0.001 | -0.04 | 0.02 | -0.10; 0.001 | -0.04 | 0.02 | -0.09; -<0.001 |

$\beta_0$ Intercept; $\beta_1$ Sertraline arm (ref: placebo arm); $\beta_2$ Hypericum arm (ref: placebo arm); $\beta_3$ Week,
$\beta_4$ Interaction between week and treatment arm (sertraline compared to placebo)
$\beta_5$ Interaction between week and treatment arm (hypericum compared to placebo)
MNAR: Missing not at random, SD: Standard deviation

Using the data up to Week 8, the results of the three Bayesian MNAR models differed slightly from the results of the MAR Bayesian model. The absolute size of the estimated coefficient for the interaction between week and sertraline compared to placebo was smaller under the MNAR models

than under the MAR model. In Models 2 and 3 this was statistically significant. In Model 1, the upper boundary of the 95% credible interval was close to 0, indicating that the interaction had a small p-value. In all three MNAR models, the interaction between week and the hypericum arm compared to placebo was not significant using any of the sets of prior distributions and the estimated coefficient was close to 0. The choice of prior set also did not influence the results appreciably (Table 6.27).

The same MNAR models were also fitted using the same sets of prior distributions with all the data up to Week 26. These models were difficult to fit using some of the other methods because the majority of the data were missing. When using all the data up to Week 26, the effect of sertraline over time was not statistically significant and we could not conclude that participants in the sertraline arm improved more than participants in the placebo arm. The effect of hypericum over time was also not statistically significant (Table 6.28). The size of the interaction coefficients differed a lot between the various models. Prior Set 2 had different results from the other prior distributions assumed, thus showing that these models were sensitive to the prior distributions assumed.

**Table 6.28: Hamilton depression score posterior means, standard deviations and credible intervals according to Bayesian analysis for data up to Week 26**

| | **Substantive model: Prior Set 1** β ~ Normal(0, 10 000), Ω~Wishart(I,5) I: Identity matrix, | | | **Substantive model: Prior Set 2** β ~ Normal(0, 1000), Ω~Wishart(A,5) A: diag (0.1, 0.1, 0.1, 0.1, 0.1 | | | **Substantive model: Prior Set 3** β ~ Normal(0, 10), Ω~Wishart(A,5) A: diag (10, 10, 10, 10, 10), | | |
|---|---|---|---|---|---|---|---|---|---|
| **Effect** | **Mean** | **SD** | **95% credible interval** | **Mean** | **SD** | **95% credible interval** | **Mean** | **SD** | **95% credible interval** |
| **MNAR model 1:** $m_{ij} \sim Bernoulli(p_{ij}), logit(p_{ij}) = \theta_0 + \Delta y_{ij}$ | | | | | | | | | |
| | **MNAR model 1: Prior Set 1** $\theta_0, \Delta$ ~ Normal(0, 10 000) | | | **MNAR model 1: Prior Set 2** $\theta_0$ ~ logistic(0,1), $\Delta$ ~ Normal(0, 1000) | | | **MNAR model 1: Prior Set 3** $\theta_0$ ~ logistic(0,1), $\Delta$ ~ Normal(0, 10) | | |
| $\beta_0$ | 21.87 | 0.36 | 21.19; 22.6 | 22.24 | 0.40 | 21.4; 22.97 | 22.86 | 0.35 | 21.18; 22.58 |
| $\beta_1$ | 0.70 | 0.45 | -0.20; 1.56 | 0.34 | 0.48 | -0.57; 1.32 | 0.66 | 0.44 | -0.20; 1.51 |
| $\beta_2$ | 0.86 | 0.44 | 0.01; 1.75 | 0.72 | 0.64 | -0.42; 2.01 | 0.88 | 0.41 | 0.09; 1.70 |
| $\beta_3$ | -1.88 | 0.21 | -2.20; -1.38 | -1.60 | 0.12 | -1.83; -1.37 | -1.98 | 0.50 | -2.84; -1.39 |
| $\beta_4$ | 0.27 | 0.18 | -0.12; 0.58 | -0.05 | 0.11 | -0.26; 0.17 | -0.14 | 0.14 | -0.41; 0.14 |
| $\beta_5$ | 0.28 | 0.20 | -0.12; 0.61 | -0.11 | 0.13 | -0.37; 0.14 | -0.01 | 0.13 | -0.23; 0.26 |
| $\beta_6$ | 0.04 | 0.003 | 0.04; 0.05 | 0.05 | 0.004 | 0.04; 0.05 | 0.10 | 0.06 | 0.04; 0.20 |
| $\Delta$ | 0.03 | 0.02 | -0.01; 0.07 | 0.03 | 0.02 | -0.004; 0.08 | 0.03 | 0.02 | -0.01; 0.08 |
| $\theta_0$ | -4.41 | 0.41 | -5.30; -3.68 | -4.53 | 0.40 | -5.45; -3.79 | -4.50 | 0.41 | -5.34; -3.74 |
| **MNAR model 2:** $m_{ij} \sim Bernoulli(p_{ij}), logit\ p_{ij} = \theta_0 + \theta_1 y_{i,j-1} + \theta_2(y_{ij} - y_{i,j-1})$ | | | | | | | | | |
| | **MNAR model 2: Prior Set 1** $\theta_0$ ~ logistic(0,1), $\theta_1, \theta_2$ ~ Normal(0, 10 000) | | | **MNAR model 2: Prior Set 2** $\theta_0, \theta_1, \theta_2$ ~ Normal(0, 100) | | | **MNAR model 2: Prior Set 3** $\theta_0, \theta_1, \theta_2$ ~ Normal(0, 10) | | |
| **Effect** | **Mean** | **SD** | **95% credible interval** | **Mean** | **SD** | **95% credible interval** | **Mean** | **SD** | **95% credible interval** |
| $\beta_0$ | 22.39 | 0.45 | 21.35; 23.12 | 22.59 | 0.29 | 22.02; 23.16 | 21.99 | 0.51 | 21.06; 22.89 |
| $\beta_1$ | 0.16 | 0.54 | -0.78; 1.34 | -0.05 | 0.40 | -0.83; 0.74 | 0.53 | 0.58 | -0.55; 1.62 |
| $\beta_2$ | 0.46 | 0.53 | -0.52; 1.53 | 0.41 | 0.39 | -0.35; 1.19 | 0.87 | 0.72 | -0.44; 2.12 |
| $\beta_3$ | -1.65 | 0.33 | -2.67; -1.33 | -1.51 | 0.11 | -1.76; -1.31 | -1.57 | 0.13 | -1.82; -1.30 |
| $\beta_4$ | -0.02 | 0.13 | -0.28; 0.28 | -0.03 | 0.10 | -0.21; 0.18 | -0.03 | 1.11 | -0.26; 0.20 |
| $\beta_5$ | 0.04 | 0.14 | -0.23; 0.37 | 0.05 | 0.11 | -0.14; 0.30 | 0.03 | 0.15 | -0.26; 0.31 |
| $\beta_6$ | 0.06 | 0.03 | 0.04; 0.16 | 0.04 | 0.003 | 0.04; 0.05 | 0.04 | 0.004 | 0.04; 0.05 |
| $\theta_0$ | -2.27 | 0.60 | -2.79; -0.24 | -2.53 | 0.17 | -2.87; -2.19 | -2.49 | 0.18 | -2.85; -2.13 |
| $\theta_1$ | 0.01 | 0.03 | -0.10; 0.04 | 0.02 | 0.01 | 0.003; 0.04 | 0.02 | 0.01 | -0.001; 0.04 |
| $\theta_2$ | 0.06 | 0.02 | -0.003; 0.10 | 0.06 | 0.02 | 0.03; 0.10 | 0.05 | 0.02 | 0.004; 0.09 |
| **MNAR model 3:** $logit\ p_{ij} = \theta_0 + \theta_1 y_{ij} + \theta_2 y_{ij} S_{ij} + \theta_3 y_{ij} H_{ij}$ | | | | | | | | | |
| | **MNAR model 3: Prior Set 1** $\theta_0$ ~ logistic(0,1), $\theta_1, \theta_2, \theta_3$~Normal(0,10 000) | | | **MNAR model 3: Prior Set 2** $\theta_0, \theta_1, \theta_2, \theta_3$~Normal(0,1000) | | | **MNAR model 3: Prior Set 3** $\theta_0, \theta_1, \theta_2, \theta_3$ ~ Normal(0, 10) | | |
| **Effect** | **Mean** | **SD** | **95% credible interval** | **Mean** | **SD** | **95% credible interval** | **Mean** | **SD** | **95% credible interval** |
| $\beta_0$ | 22.34 | 0.39 | 21.53; 23.03 | 21.93 | 0.4 | 21.11; 22.73 | 22.06 | 0.38 | 21.3; 22.78 |
| $\beta_1$ | 0.21 | 0.48 | -0.7; 1.19 | 0.61 | 0.49 | .-0.36; 1.59 | 0.49 | 0.47 | -0.41; 1.41 |
| $\beta_2$ | 0.67 | 0.48 | -0.25; 1.63 | 1.18 | 0.55 | 0.10; 2.26 | 0.95 | 0.46 | 0.05; 1.87 |
| $\beta_3$ | -1.55 | 0.13 | -1.79; -1.31 | -2.01 | 0.32 | .-2.73; -1.44 | -1.58 | 0.13 | -1.84; -1.35 |
| $\beta_4$ | -0.05 | 0.14 | -0.35; 0.19 | 0.24 | 0.23 | .-0.16; 0.59 | -0.02 | 0.11 | -0.23; 0.22 |
| $\beta_5$ | 0.07 | 0.15 | -0.23; 0.33 | 0.35 | 0.27 | .-0.13; 0.77 | 0.10 | 0.13 | -0.14; 0.36 |
| $\beta_6$ | 0.04 | 0.003 | 0.04; 0.05 | 0.06 | 0.04 | 0.04; 0.18 | 0.04 | 0.003 | 0.04; 0.05 |
| $\theta_0$ | -4.25 | 0.39 | .-5.05; -3.53 | -4.3 | 0.42 | .-5.20; -3.55 | -4.42 | 0.41 | -5.26; -3.66 |
| $\theta_1$ | 0.04 | 0.02 | .-0.003; 0.09 | 0.04 | 0.03 | .-0.002; 0.10 | 0.05 | 0.02 | 0.01; 0.10 |
| $\theta_2$ | -0.02 | 0.02 | .-0.06; 0.02 | -0.02 | 0.02 | .-0.06; 0.03 | -0.02 | 0.02 | -0.06; 0.02 |
| $\theta_3$ | -0.09 | 0.03 | .-0.16; -0.03 | -0.08 | 0.03 | .-0.16; -0.03 | -0.09 | 0.03 | -0.15; -0.03 |

$\beta_0$ Intercept; $\beta_1$ Sertraline arm (ref: placebo arm); $\beta_2$ Hypericum arm (ref: placebo arm), $\beta_3$ Week,
$\beta_4$ Interaction between week and treatment arm (sertraline compared to placebo)
$\beta_5$ Interaction between week and treatment arm (hypericum compared to placebo); $\beta_6$: Week squared
MNAR: Missing not at random, SD: Standard deviation

### 6.5.3 Shared-parameter models

Longitudinal HAM-D score and time to dropout were jointly modelled, with shared random effects, conditional upon which they are independent. It is assumed that an underlying individual characteristic, such as disease process, influences both changes in HAM-D score and whether a participant drops out of the study. The measurement error terms for the longitudinal model are assumed to be independent and to follow a normal distribution with mean 0 and variance $\sigma^2$. Dropout is modelled through a linear effect of the treatment variables, intercept and a random disturbance term in a time to event model. The baseline hazard function follows a Weibull distribution. The connection between HAM-D score and time to dropout is accomplished via the shared random effects U and V, which have mean 0 and the parameters $a_3$ and $a_4$ rescale their variances to account for the different scales of HAM-D score and the time variable. A Cholesky parameterisation of the random effects covariance matrix was used to ensure positive definiteness. Initial values of the parameters play an important role in whether the model converges and reasonable estimates of the initial values were estimated by fitting the longitudinal model using the SAS procedure MIXED and the survival model using the SAS procedure LIFEREG separately.

Model with shared random intercept, $\mathrm{U}_i$, and random slope, $\mathrm{V}_i$, between the survival and the longitudinal sub-models:

Survival sub-model: $h_i(t) = h_0(\mathrm{t})\exp[a_0 + a_1\mathrm{S_i} + a_2\mathrm{H_i} + a_3\mathrm{U}_i + a_4\mathrm{V}_i]$

where $h_0(\mathrm{t})$ is the baseline hazard function and follows a Weibull distribution.

Longitudinal sub-model:

$$Y_{ij} = \beta_1 + \mathrm{U}_i + \beta_2 t_{ij} + \mathrm{V}_i t_{ij} + \beta_3\mathrm{S_i} + \beta_4\mathrm{H_i} + \beta_5\mathrm{S_i}t_{ij} + \beta_6\mathrm{H_i}t_{ij} + \varepsilon_i$$

for data up to Week 8. The data up to Week 26 were fitted using the same models, with the addition of $\beta_7 t_{ij}^2$ in the longitudinal sub-model.

Model with shared random intercept, $\mathrm{U}_i$, between the survival and the longitudinal sub-models:

Survival sub-model: $h_i(t) = h_0(\mathrm{t})\exp[a_0 + a_1\mathrm{S_i} + a_2\mathrm{H_i} + a_3\mathrm{U}_i]$

where $h_0(\mathrm{t})$ is the baseline hazard function and follows a Weibull distribution

Longitudinal sub-model: $Y_{ij} = \beta_1 + \mathrm{U}_i + \beta_2 t_{ij} + \beta_3\mathrm{S_i} + \beta_4\mathrm{H_i} + \beta_5\mathrm{S_i}t_{ij} + \beta_6\mathrm{H_i}t_{ij} + \varepsilon_i$

for data up to Week 8. The data up to Week 26 were fitted using the same models, with the addition of $\beta_7 t_{ij}^2$ in the longitudinal sub-model.

**Table 6.29: Joint model for HAM-D score over time and time to dropout: Data up to Week 8**

| | | Intercept linkage | | | Intercept and linear slope linkage | | |
|---|---|---|---|---|---|---|---|
| **Effect** | **Parameter** | **Estimate** | **Standard error** | **p-value** | **Estimate** | **Standard error** | **p-value** |
| **Longitudinal sub-model** | | | | | | | |
| Intercept | $\beta_1$ | 20.98 | 0.47 | <0.001 | 20.95 | 0.37 | <0.001 |
| Week | $\beta_2$ | -1.17 | 0.05 | <0.001 | -1.15 | 0.08 | <0.001 |
| Sertraline (ref: placebo) | $\beta_3$ | -0.22 | 0.67 | 0.741 | -0.18 | 0.53 | 0.737 |
| Hypericum (ref: placebo) | $\beta_4$ | 0.81 | 0.66 | 0.223 | 0.78 | 0.52 | 0.137 |
| Interaction week and treatment (sertraline compared to placebo) | $\beta_5$ | -0.16 | 0.077 | 0.042 | -0.17 | 0.12 | 0.154 |
| Interaction week and treatment (hypericum compared to placebo) | $\beta_6$ | 0.07 | 0.076 | 0.380 | 0.06 | 0.12 | 0.600 |
| **Time to dropout sub-model** | | | | | | | |
| Intercept survival model only | $a_0$ | 5.53 | 13.10 | 0.673 | 6.18 | 15.04 | 0.681 |
| Sertraline (ref: placebo) | $a_1$ | 0.70 | 19.97 | 0.972 | 0.95 | 23.41 | 0.968 |
| Hypericum (ref: placebo) | $a_2$ | 0.73 | 20.58 | 0.972 | 0.98 | 22.78 | 0.966 |
| Intercept shared | $a_3$ | 0.002 | 1.22 | 0.999 | 0.02 | 1.64 | 0.990 |
| Slope shared | $a_4$ | | | | 0.58 | 10.53 | 0.956 |

From the joint model with intercept linkage we concluded that the interaction between week and the sertraline arm compared to the placebo arm was statistically significant. The interaction between week and the hypericum arm compared to the placebo arm was not significant. In the joint model with shared intercept and slope neither of the interactions between treatment arm and week was significant. Although the interaction between week and the hypericum arm was significant in the joint model with random intercept and not in the joint model with random intercept and slope the actual estimates were similar in the two models. The standard error was larger in the model with random intercept and slope. This model also used more degrees of freedom to estimate all the parameters (Table 6.29).

The results for the survival sub-model were similar in the two models. The shared intercept ($a_3$) estimate was close to 0 in both joint models and not significant in both models, indicating that baseline HAM-D score did not influence dropout time. The positive shared slope indicated that participants with larger linear increase in HAM-D score also had an increase in dropout time, but this was not statistically significant.

The substantive model fitted in these joint models is not the same as the substantive model fitted in the other sections because SAS cannot fit such a joint model. The joint model that can be fitted with current SAS capabilities was fitted. In order to compare the joint model under MNAR with the appropriate substantive model, and highlight the differences between fitting a joint model and just fitting a longitudinal model, a MAR model was fitted with the same formulation as the

longitudinal sub-model in the joint model. These models incorporate the repeated measures of each participant by including a participant specific random intercept and a participant specific random intercept and slope, respectively. These models were fitted with a compound symmetry covariance matrix and are given in Table 6.30.

**Table 6.30: MAR longitudinal model with random subject specific effects: comparable to the longitudinal sub-model in Table 6.29: Data up to Week 8**

| | | Intercept linkage | | | Intercept and linear slope linkage | | |
|---|---|---|---|---|---|---|---|
| **Effect** | **Parameter** | **Estimate** | **Standard error** | **p-value** | **Estimate** | **Standard error** | **p-value** |
| **Longitudinal sub-model** | | | | | | | |
| Intercept | $\beta_1$ | 21.00 | 0.47 | <0.001 | 20.94 | 0.24 | <0.001 |
| Week | $\beta_2$ | -1.17 | 0.05 | <0.001 | -1.16 | 0.10 | <0.001 |
| Sertraline (ref: placebo) | $\beta_3$ | -0.21 | 0.67 | 0.757 | -0.27 | 0.35 | 0.447 |
| Hypericum (ref: placebo) | $\beta_4$ | 0.76 | 0.67 | 0.254 | 0.76 | 0.35 | 0.030 |
| Interaction week and treatment (sertraline compared to placebo) | $\beta_5$ | -0.16 | 0.08 | 0.041 | -0.15 | 0.15 | 0.332 |
| Interaction week and treatment (hypericum compared to placebo) | $\beta_6$ | 0.07 | 0.08 | 0.363 | 0.07 | 0.15 | 0.658 |

There is very little difference between the β-estimates and p-values between the joint model and the longitudinal model only (Table 6.29 and Table 6.30). The same conclusions are drawn. The deviations from MAR assumed by the joint model do not change the conclusions.

The joint model using data up to Week 26 found that the interaction between week and the sertraline arm compared to the placebo arm was significant in the model with intercept and slope linkage. It was not significant in the model with intercept linkage only. Neither of the models had a significant interaction between week and hypericum compared to placebo. None of the parameters in the survival sub-model was significant (Table 6.31).

**Table 6.31: Joint model for HAM-D score over time and time to dropout: Data up to Week 26**

| | | Intercept linkage | | | Intercept and linear slope linkage | | |
|---|---|---|---|---|---|---|---|
| **Effect** | **Parameter** | **Estimate** | **Standard error** | **p-value** | **Estimate** | **Standard error** | **p-value** |
| **Longitudinal sub-model** | | | | | | | |
| Intercept | $\beta_1$ | 20.92 | 0.44 | <0.001 | 20.97 | 0.39 | <0.001 |
| Week | $\beta_2$ | -1.42 | 0.04 | <0.001 | -1.41 | 0.06 | <0.001 |
| Sertraline (ref: placebo) | $\beta_3$ | -0.22 | 0.62 | 0.718 | -0.07 | 0.55 | 0.903 |
| Hypericum (ref: placebo) | $\beta_4$ | 1.17 | 0.62 | 0.057 | 0.87 | 0.55 | 0.117 |
| Interaction week and treatment (sertraline compared to placebo) | $\beta_5$ | -0.05 | 0.03 | 0.118 | -0.17 | 0.08 | 0.039 |
| Interaction week and treatment (hypericum compared to placebo) | $\beta_6$ | 0.02 | 0.04 | 0.520 | 0.03 | 0.08 | 0.673 |
| Week squared | $\beta_7$ | 0.04 | 0.002 | <0.001 | 0.05 | 0.002 | <0.001 |
| **Time to dropout sub-model** | | | | | | | |
| Intercept survival model only | $a_0$ | 7.42 | 4.95 | 0.135 | 7.30 | 7.26 | 0.316 |
| Sertraline (ref: placebo) | $a_1$ | -0.80 | 4.76 | 0.867 | 1.05 | 7.81 | 0.894 |
| Hypericum (ref: placebo) | $a_2$ | -0.84 | 4.92 | 0.865 | 1.06 | 7.86 | 0.893 |
| Intercept shared | $a_3$ | -0.11 | 0.39 | 0.782 | -0.07 | 0.90 | 0.938 |
| Slope shared | $a_4$ | | | | 0.91 | 12.22 | 0.941 |

**Table 6.32: MAR longitudinal model with random subject specific effects: comparable to the longitudinal sub-model in Table 6.29: Data up to Week 26**

| | | Intercept linkage | | | Intercept and linear slope linkage | | |
|---|---|---|---|---|---|---|---|
| **Effect** | **Parameter** | **Estimate** | **Standard error** | **p-value** | **Estimate** | **Standard error** | **p-value** |
| **Longitudinal sub-model** | | | | | | | |
| Intercept | $\beta_1$ | 21.19 | 0.44 | <0.001 | 20.82 | 0.24 | <0.001 |
| Week | $\beta_2$ | -1.43 | 0.04 | <0.001 | -1.38 | 0.10 | <0.001 |
| Sertraline (ref: placebo) | $\beta_3$ | -0.53 | 0.62 | 0.392 | 0.08 | 0.32 | 0.795 |
| Hypericum (ref: placebo) | $\beta_4$ | 0.90 | 0.62 | 0.146 | 0.91 | 0.33 | 0.006 |
| Interaction week and treatment (sertraline compared to placebo) | $\beta_5$ | -0.05 | 0.03 | 0.152 | -0.26 | 0.14 | 0.066 |
| Interaction week and treatment (hypericum compared to placebo) | $\beta_6$ | 0.03 | 0.04 | 0.444 | 0.02 | 0.14 | 0.915 |
| Week squared | $\beta_7$ | 0.04 | 0.002 | <0.001 | 0.06 | 0.002 | <0.001 |

As with the data up to Week 8, we also fitted the longitudinal model only in order to compare the joint model to a MAR model with the same substantive model (Table 6.32). The results do not seem to differ substantially between the joint model and the similar longitudinal model only. The only noteworthy difference is for the interaction between week and hypericum treatment compared to placebo, where the p-value changes from 0.039 to 0.066.

The joint models deviate from the MAR models by allowing the longitudinal profiles and the missingness process to share random effects. The models need to be interpreted conditional on these random effects. Shared-random effects models relate a participant's propensity to show an increase or decrease in HAM-D score with his or her propensity to miss a visit, while selection models directly model the probability of a missed visit as a function of HAM-D score.

### 6.5.4 MNAR conclusions

No unequivocal conclusion can be drawn from the MNAR models fitted, because many MNAR models could be fitted and the data holds no indication as to which of the models are more appropriate.

Using data up to Week 8, the sertraline arm showed a significant improvement over the placebo arm according to a pattern-mixture model fitted using random effects mixed models for each pattern of missing data. The multiple imputation pattern-mixture models assuming an improvement in HAM-D score after dropout also showed reduced depression scores over time in the sertraline arm compared to the placebo arm, but models assuming a decrease in depression scores after drop-out did not find a significant difference between the sertraline and placebo arms. Pattern-mixture models fitted using CCMV and NCMV identifying restrictions did not show significant improvement over time in the sertraline arm. Two of the three MNAR Bayesian selection models found significant improvement over time in the sertraline arm compared to the placebo arm. A joint model with shared intercept also had a significant result for the interaction between week and the sertaline arm compared to the placebo arm. In most models the size of the estimated interaction coefficient was similar to the size of the coefficient under the MAR models. None of the MNAR models fitted, found a significant interaction between week and hypericum compared to placebo over the first 8 weeks. The interaction coefficient was close to 0, as in the MAR case (Table 6.33).

Fitting models using all the data up to Week 26 was problematic. From week 10 and onwards there were more missing data than observed data. Many models fitted poorly to this data or had problems converging. Using all the data up to Week 26, only one of the MNAR models found a significant interaction between week and the sertraline arm compared to placebo, this was the joint model with shared intercept and slope linkage. No significant interaction between week and the hypericum arm compared to placebo was found (Table 6.34).

**Table 6.33: Summary of findings in MNAR sensitivity analyses up to Week 8**

| | Interaction between week and sertraline arm | | Interaction between week and hypericum arm | |
|---|---|---|---|---|
| | **Estimate** | **p-value** | **Estimate** | **p-value** |
| **Pattern-mixture model** | | | | |
| Pattern-mixture model using random-effects mixed models for each of the patterns of missing data (Model 1) | -0.38 | 0.01 | -0.002 | 0.99 |
| Pattern-mixture models using multiple imputation | | | | |
| Model 1: δ for dropout = 0.5 | -0.44 | 0.09 | 0.15 | 0.51 |
| Model 2: δ for dropout = 1 | -0.63 | 0.01 | 0.34 | 0.23 |
| Model 3: δ for dropout = -0.5 | -0.30 | 0.56 | 0.08 | 0.85 |
| Model 4: δ for dropout = -1 | -0.32 | 0.55 | 0.07 | 0.86 |
| Pattern-mixture models using identifying restrictions | | | | |
| CCMV | 0.08 | 0.47 | 0.13 | 0.21 |
| NCMV | -0.16 | 0.38 | 0.03 | 0.82 |
| **Selection models** | | | | |
| Bayesian MNAR model 1 | -0.28 | NS | -0.01 | NS |
| Bayesian MNAR model 2 | -0.33 | Sign | -0.03 | NS |
| Bayesian MNAR model 3 | -0.31 | Sign | -0.01 | NS |
| **Shared parameter models** | | | | |
| Joint longitudinal and time to event model with intercept linkage | -0.16 | 0.04 | 0.07 | 0.38 |
| Joint longitudinal and time to event model with intercept and slope linkage | -0.17 | 0.15 | 0.06 | 0.60 |

Sign = significant, NS = not significant

Using data up to 26 weeks, the majority of data were missing and the effect of the assumption made about missing data compounded with each subsequent missing value. This means that at the later time points the assumptions made about the MNAR pattern of the missing data had a larger effect on the model fitted than the data observed.

As with the MAR analysis, we were interested in Estimand 1. To estimate this estimand data are needed on all participants at the planned end of the trial. All the MNAR methods made assumptions about the missing data process in order to extrapolate or impute, by using for example multiple imputations, the data that was not observed. If we wanted to estimate drug efficacy, we could have estimated Estimands 3 and 6, as discussed in Section 6.4.5.

**Table 6.34: Summary of findings in MNAR sensitivity analyses up to Week 26**

| | Interaction between week and sertraline arm | | Interaction between week and hypericum arm | |
|---|---|---|---|---|
| | Estimate | p-value | Estimate | p-value |
| **Pattern-mixture model** | | | | |
| Pattern-mixture model using random- effects mixed models for each of the patterns of missing data | | | | |
| Model 2: Three dropout categories | -0.10 | 0.17 | 0.09 | 0.18 |
| Model 3: Binary dropout | -0.08 | 0.12 | -0.02 | 0.68 |
| Pattern-mixture models using multiple imputation | | | | |
| Model 5: δ for dropout = 0.5 | -0.07 | 0.90 | -0.004 | 0.99 |
| Model 6: δ for dropout = 1 | -0.08 | 0.88 | -0.02 | 0.96 |
| Model 7: δ for dropout = -0.5 | -0.06 | 0.90 | 0.05 | 0.90 |
| Model 8: δ for dropout = -1 | -0.05 | 0.91 | 0.07 | 0.87 |
| Pattern-mixture models using identifying restrictions | | | | |
| CCMV | 0.16 | 0.16 | 0.11 | 0.24 |
| NCMV | 0.05 | 0.78 | -0.01 | 0.94 |
| **Bayesian models** | | | | |
| Bayesian MNAR model 1 | 0.27 | NS | 0.28 | NS |
| Bayesian MNAR model 2 | -0.02 | NS | 0.04 | NS |
| Bayesian MNAR model 3 | -0.05 | NS | 0.07 | NS |
| **Shared parameter models** | | | | |
| Joint longitudinal and time to event model with intercept linkage | -0.05 | 0.12 | 0.02 | 0.52 |
| Joint longitudinal and time to event model with intercept and slope linkage | -0.17 | 0.04 | 0.03 | 0.67 |

Sign = significant, NS = not significant

As mentioned previously, because so much data were missing, it might have been better to analyse the data from Week 10 to Week 26 in a conditional way. In such a conditional analysis, the question addressed is; given that participants responded to treatment at Week 8, what was the continued effect of treatment arm up to Week 26. In this instance, there was still a lot of missing data and this analysis would need to take missing data into account. As it is currently presented, these analyses should be interpreted with extreme caution and it does not address any well-defined estimand of interest.

## 6.6 Conclusion

In contrast to the SAPiT study discussed in Chapter 5, this study had missing data by design. Participants who did not show adequate response by Week 8 were not enrolled in the continuation phase of the study. This type of study design was encouraged by Carpenter et al. (2002). However, this study illustrates that this design should be used with caution. Because of the large amount of missing data introduced by design the standard statistical methods break down and the analysis could not be performed validly.

The multiple imputation analysis did not reach a stationary distribution and autocorrelation existed between the imputations from Week 10 onwards, making the analysis invalid. Valid analyses, if

the unverifiable assumptions about missing data were correct, were the direct likelihood analysis (assuming MAR) and the pattern-mixture analysis (assuming MNAR) and the Bayesian analyses (assuming MAR and MNAR).

In this study, the goal was to determine the short term effect of the intervention at Week 8, and to confirm that a positive response at Week 8 is maintained over a longer time period. An objective was also to provide safety data with longer use. Since neither of the interventions was superior to placebo, the continuation phase data were not needed, nor analysed in the original paper. If the interventions were more effective, less missing data should have resulted by design. However, if the interventions were more effective, it is possible that the placebo arm could have a lot of missing data, while the active arms have less missing data. This could still create problems with valid fit of models, at least in the placebo arm. The wisdom of a design such as this, from a statistical viewpoint, is questionable. From an ethical viewpoint, this design is very attractive, since participants are not kept on a regimen that is not working for a long time. A better compromise between the ethics and statistics is to allow the study clinicians to initiate participants on a rescue treatment at any time it is deemed clinically necessary. Participants are not discontinued from the study and data collection continues when rescue treatments are started.

The data collected allow one to get the conditional response over time, given that the participant entered the continuation phase and to have safety data with prolonged use; but it is not possible to calculate the unconditional effect of either of the drugs at 26 weeks. In this instance it did not matter because the drug was found to have no effect at 8 weeks, which implied that there was also no scientific interest in the effect at 26 weeks. Because of non-random exclusion of participants, the comparability of the three treatment arms after Week 8 is not equivalent to a randomised trial. At best, this provides an observational study about the longer term effects of the drugs. Any efficacy analysis in this continuation phase should be interpreted with caution.

A different study design, where all participants enter an open label safety extension, regardless of response in the active phase, could also be employed. This design was used in the studies of Prucalopride for the treatment of constipation (Camilleri et al., 2008; Tack et al., 2009). This design enable one to collect long term safety data, and maybe even long term efficacy data, without inducing missing data by design. This study has been analysed using proper missing data techniques (Janssens et al., 2012).

The conclusion drawn in the original paper was that there was no difference between either the hypericum arm and placebo or the sertraline arm and placebo. This was taken to mean that the study results were inconclusive. Although it showed that hypericum was no better than placebo, it

also did not find the expected difference between sertraline and placebo (Hypericum Depression Trial Study Group, 2002). The same was found when the continuation data were analysed, and the placebo effect was again noted and discussed (Sarris et al., 2012). We re-analysed the data using methods that are appropriate with missing data. We fitted various models using different assumptions about the missing data and various analysis methods. Under this extended sensitivity analysis we draw a different conclusion.

Some of our conclusions are similar to the original analysis, but both the MAR and some of the MNAR analyses lead to the conclusion that sertraline is significantly better than placebo in reducing depression symptoms over 8 weeks. No difference was found between hypericum and placebo in any of the analyses. This implies that the conclusion reached 10 years ago could be amended to state that hypericum does not seem to provide any benefit over placebo, in a trial where the active control did provide a benefit over placebo in some of the sensitivity analyses; thus concluding that this herbal remedy should not be recommended for the treatment of depression. Although this study gives only weak evidence of the efficacy of sertraline (in some of the models under some of the assumptions about missing data), there was absolutely no evidence that hypericum was superior to placebo under any reasonable assumption about missing data. This illustrates the point that not taking account of missing data in the analysis could introduce bias and lead to incorrect results.

This re-analysis of the data and sensitivity analysis adds to the field of applied statistics by showing how a proper analysis of data taking missing data into account can lead to different results than a more naïve analysis. Careful sensitivity analysis, or even re-analysis of data taking proper assumptions into account, brought novel insight; or even contradictory insight to the question regarding the effectiveness of St John's Wort in the treatment of depression.

Adjusting the analysis to take missing data into account does not imply changing the proposed estimate of effectiveness. The measure of effectiveness reported in the original paper was change from baseline to Week 8. We analysed the interaction between treatment arm and time under MAR and MNAR assumptions. In general multiple imputation allows any measure of effectiveness, since the analysis of choice is done in the second step.

The analysis was done with standard statistical software, using resources that should be available to most researchers. The availability of software is no longer a reason not to do the proper principled analyses when data are incomplete.

# Chapter 7
# Conclusions and discussions

## 7.1 Implication for practice

Missing data is an active area of research in theoretical statistics. Many methodological advances have been made in recent years. New text books on missing data are appearing regularly. Two text books on missing data were published in each of 2010 (Enders, 2010; Tsiatis, 2010), 2011 (Bethlehem, Cobben, & Schouten, 2011; Drukker, 2011) and 2012 (Graham, 2012; Tan, Tian, & Ng, 2012). In the first two months of 2013, two text books on missing data have appeared (Carpenter & Kenward, 2013; Mallinckrodt, 2013).

Although methods to handle missing data have been in development for the last 30 years, it is still not part of mainstream statistical methods training. It is unlikely that an undergraduate university course in statistics would include extensive training in the handling of missing data. Most non-statistical users of statistical methods and non-statistical medical researchers are still likely to ignore missing data or analyse participants with complete data only. Missing data are simply excluded from the analysis. This is compounded by the default setting in many statistical programs that exclude missing observations (SAS Institute Inc., 2004), therefore doing a complete case analysis unless the user overrides the defaults. An exception to this is most longitudinal procedures that do an analysis of the available cases by default. This enables direct likelihood and direct Bayesian analyses.

About ten years ago, it was common practice to analyse clinical trials using LOCF. This is changing in favour of more appropriate methods. Inappropriate methods, such as single imputations and LOCF are still being used; although the recommendations advice against these (Carpenter et al., 2004; Mallinckrodt et al., 2001; Molenberghs & Kenward, 2007; Molenberghs et al., 2004).

The ignorance about the correct methods to analyse missing data are shown in a recent paper where three methods for the analysis of missing data are tested using simulated data: the three methods were complete case analysis, single imputation and multiple imputation (Groenwold et al., 2012). Of these three methods only multiple imputation is regarded as an unbiased method to analyse MAR data (Carpenter & Kenward, 2007). The conclusion reached in that publication was that complete case analysis with covariate adjustment should be used more often. This is simply wrong, as pointed out by Liublinska and Rubin (2012) in a response to the paper.

Some of the methods of missing data analysis are hard to use and definitely not mainstream or likely to be. Some of the methods make lots of assumptions, require programming skills or knowledge of advanced statistical techniques. But some of the methods are easy to use, available in standard software and should be familiar to most statisticians; especially likelihood-based methods, such as those based on mixed models.

The major hurdle to the implementation of missing data methods is not the methodology or availability of software anymore. The major hurdle is rather a lack of awareness of the problems posed by missing data. Many researchers do not see excluding data as a problem. Many researchers exclude participants with missing data or missing outcomes from the data set prior to bringing the data set to a statistician. Unless the statistician is specifically vigilant and asks about missing data or excluded observations the statistician is unlikely to realise that bias might be present and that missing data issues should be addressed during the analysis of the data. Some researchers replace participants with missing data, thus creating a complete data set, which also poses a problem.

This lack of awareness of the importance of missing data is slowly changing. A large impetus in this direction might come from journals publishing medical research and the results of clinical trials. I am aware of at least two prominent journals that focus on the handling of missing data during the peer review of submitted papers. One is the Annals of Internal Medicine. The guidelines for authors has a section on missing data where they specifically require authors to report the frequency of missing variables and how the analysis handled missing data. They also

say that an LOCF analysis should not be used. To quote from the guidelines for authors: "*Appropriate methods for handling missing data include imputation, pattern-mixture (mixed) models, and selection models. Application of these methods requires consideration of the patterns and potential mechanisms behind the missing data*." (American College of Physicians, 2012). Our research group has recently submitted a paper to this journal and the majority of comments received focused on the handling of missing data and requested a sensitivity analysis. Of interest, this medical journal has a statistics deputy editor, six statistical associate editors and two statistical consultants; this could be the reason they are focusing on missing data.

The second is the New England Journal of Medicine, which published an update to their policies regarding missing data in response to a report by the National Research Council (Ware et al., 2012). The change in policy included a need to justify the use of complete case analyses and single imputation. Weighted estimating equations, multiple imputation or model based methods are preferred. Sensitivity analyses might be required if missing data are substantial. Authors are required to describe the rationale for the choice of models as well as details of the models fitted. The journal of Clinical Pharmacology and Therapeutics also covered the handling of missing data recently (O'Neill & Temple, 2012).

If prominent journals such as the Annals of Internal Medicine and the New England Journal of Medicine require submissions to handle missing data adequately, researchers and statisticians will be forced to become familiar with standard methodology for the analysis of missing data. This is a giant step in the right direction.

Authors, and especially expert panels (CHMP, 2010; Little et al., 2012; National Research Council, 2010), are quick to point out that the problem of missing data should not primarily be addressed at the analysis stage. They focus also on the design and conduct of clinical trials as important in reducing the amount of missing data. This is definitely an area to focus on in an attempt to improve clinical trials. More research into missing data methodology might not improve clinical trials as much as more guidelines to researchers on methods to design clinical trials in a way that will reduce missing data in the first place.

Sensitivity analyses are seen as an important element of any analysis when a large amount of data are missing. However, no clear guidelines are available for sensitivity analyses (Ware et al., 2012). The most useful sensitivity analyses would probably include modelling a MNAR model, which uses much more complicated methods than the likelihood-based methods that can be used when data are assumed to be MAR. An area where innovation is needed is the development of guidelines

for sensitivity analyses and the development of software that would be able to easily, without too much statistical programming by the user, do MNAR analyses.

What lessons could be learnt about the application of missing data techniques from the SAPiT study? Maybe the most discouraging lesson is that, although it seemed from the data that the missing data process was probably not ignorable, and several models under different assumptions were fitted, the conclusions were similar to the conclusions under a naïve model. In the end the more complicated analyses did not change the interpretation that one would have arrived at if a simple likelihood-based analysis was fitted. Thus, taking account of the effect of missing data did not change the conclusions in this instance.

The St John's Wort data were difficult to analyse because such a large proportion of the data was missing. In the original article published using this data, data were only analysed up to 8 weeks, before most of the data were missing (Hypericum Depression Trial Study Group, 2002). The conclusion drawn at 8 weeks was that hypericum was not efficacious, thus obviating the need to analyse the follow-up data. However, had it been necessary to analyse the follow-up data, it might not have been possible to do any useful analysis, other than a conditional analysis, namely looking at the outcomes given that a participant entered into the follow-up phase. The St John's Wort study highlighted the shortcomings of most of the missing data methods once a substantial proportion of the data is missing. This is a significant weakness, since missing data are more likely to lead to bias and becomes more important the more missing data there is. Even in instances where missing data are MNAR, the final results will probably be fairly unbiased as long as there is a small amount of missing data. This means that missing data techniques becomes more difficult to use, exactly when they are more needed. Statistical methodological research could therefore focus on cases where substantial amounts of data are missing and find optimum methods for these instances. It would also be helpful to know what the maximum amount of missing data is that a specific method can reliably handle.

Bayesian data analysis methods are not often used in the analysis of clinical trial data. This is also changing and Bayesian methods are being used more readily. In the SAPiT trial, the Bayesian models were easy to apply and gave results fairly similar to the likelihood-based methods. Bayesian analyses could provide a very usable framework in which to deal with missing data, since Bayesian methods can be applied when data are incomplete in much the same way as when data are complete. Some of the likelihood-based methods had computational problems with models that did not converge. Bayesian methods are much more robust and could be fitted in all instances in this thesis. This means that Bayesian models could be a viable alternative in certain settings where the computations became difficult.

In summary: The theoretical development of missing data techniques is light years ahead of the application of missing data techniques in everyday use. At a minimum researchers should be aware of the bias that could be introduced by missing data and should report on the missing data present in their data and what effect this could have on the results. It is unlikely that routine analyses will include more advanced missing data techniques such as pattern-mixture models and selection models soon. However, likelihood-based methods, and to a lesser extent Bayesian methods, are easily available and provide a valid, easy to use analysis if the missing data mechanism is MAR.

The latest buzzword is sensitivity analysis, but there is no clarity on the requirements for a sensitivity analysis. It is logical that a sensitivity analysis should include models that assume both MAR and MNAR missing data. However, MNAR models are much more complicated and make several assumptions and could probably not be fitted by the 'average' researcher without additional training in statistical programming or missing data methodology. It might also require advanced statistical knowledge to understand how to translate practical assumptions about the missing data into statistical elements of the models fitted. This makes the requirement for sensitivity analysis rather onerous to apply in most contexts. In practice a sensitivity analysis may be done, but may not really include all relevant alternative models. A sensitivity analysis could also never be exhaustive, since several alternative assumptions could be made and several models could be fitted under these assumptions. This means that even when a sensitivity analysis is presented, it might still not give reliable evidence about the robustness of the results.

# 8. References

Abdool Karim, S. S., Naidoo, K., Grobler, A., Padayatchi, N., Baxter, C., Gray, A., Gengiah, T., Nair, G., Bamber, S., Singh, A., Khan, M., Pienaar, J., El-Sadr, W., Friedland, G., & Abdool Karim, Q. (2010). Timing of initiation of antiretroviral drugs during tuberculosis therapy. *New England Journal of Medicine, 362*(8), 697-706. doi: 362/8/697 [pii]10.1056/NEJMoa0905848

Abdool Karim, S. S., Naidoo, K., Grobler, A., Padayatchi, N., Baxter, C., Gray, A. L., Gengiah, T., Gengiah, S., Naidoo, A., Jithoo, N., Nair, G., El-Sadr, W. M., Friedland, G., & Abdool Karim, Q. (2011). Integration of Antiretroviral Therapy with Tuberculosis Treatment. *New England Journal of Medicine, 365*(16), 1492-1501. doi: 10.1056/NEJMoa1014181

Albert, P. S., & Follmann, D. (2009). Shared-parameter models. In G. Fitzmaurice, M. Davidian, G. Verbeke & G. Molenberghs (Eds.), *Longitudinal data analysis*. Boca Raton: Chapman & Hall/CRC.

Albert, P. S., & Follmann, D. A. (2000). Modeling Repeated Count Data Subject to Informative Dropout. *Biometrics, 56*(3), 667-677. doi: 10.1111/j.0006-341X.2000.00667.x

Allison, P. D. (2009). Missing data. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The SAGE Handbook of Quantitative Methods in Psychology* (pp. 72-89). Thousand Oaks: Sage Publications Inc.

American College of Physicians. (2012). Annals of internal medicine information for authors. Retrieved 13 December 2012, 2012, from http://annals.org/public/authorsinfo.aspx

Azur, M., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International Journal Methods Psychiatric Research, 20*(1), 40-49.

Bang, H., & Robins, J. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics, 61*(4), 962-973.

Bethlehem, J., Cobben, F., & Schouten, B. (2011). *Handbook of non-response in household surveys*. Hoboken, NJ: Wiley.

Beunckens, C., Sotto, C., Molenberghs, G., & Verbeke, G. (2009). A multifaceted sensitivity analysis of the Slovenian public opinion survey data. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 58*(2), 171-196. doi: 10.1111/j.1467-9876.2009.00647.x

Bodner, T. E. (2008). What Improves with Increased Missing Data Imputations? *Structural Equation Modeling: A Multidisciplinary Journal, 15*(4), 651-675. doi: 10.1080/10705510802339072

Burton, A., & Altman, D. G. (2004). Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. *British Journal of Cancer, 91*(1), 4-8. doi: 10.1038/sj.bjc.66019076601907 [pii]

Camilleri, M., Kerstens, R., Rykx, A., & Vandeplassche, L. (2008). A Placebo-Controlled Trial of Prucalopride for Severe Chronic Constipation. *New England Journal of Medicine, 358*(22), 2344-2354. doi: 10.1056/NEJMoa0800670

Carpenter, J., & Kenward, M. G. (2007). *Missing data in randomised controlled trials - a practical guide*.

Carpenter, J., & Kenward, M. G. (2013). *Multiple imputation and its application* Chichester, United Kingdom: John Wiley and Sons, Ltd.

Carpenter, J., Kenward, M. G., Evans, S., & White, I. (2004). Last observation carried forward and last observation analysis. Letter to the editor. *Statistics in Medicine, 23*(3241-3244). doi: 10.1002/sim.1891

Carpenter, J., Pocock, S., & Lamm, C. J. (2002). Coping with missing data in clinical trials: A model-based approach applied to asthma trials. *Statistics in Medicine, 21*(8), 1043-1066. doi: 10.1002/sim.1065

Carpenter, J. R., Kenward, M. G., & Vansteelandt, S. (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 169*(3), 571-584. doi: 10.1111/j.1467-985X.2006.00407.x

Carpenter, J. R., Roger, J. H., & Kenward, M. G. (2013). Analysis of Longitudinal Trials with Protocol Deviation: A Framework for Relevant, Accessible Assumptions, and Inference via Multiple Imputation. *Journal of Biopharmaceutical Statistics, 23*(6), 1352-1371. doi: 10.1080/10543406.2013.834911

Carrigan, G., Barnett, A. G., Dobson, A. J., & Mishra, G. (2007). Compensating for missing data from longitudinal studies using WINBUGS. *Journal of statistical software, 19*(7).

Casella, G., & Berger, R. L. (2002). *Statistical Inference*. Pacific Grove, CA: Duxbury.

Chi, Y.-Y., & Ibrahim, J. G. (2006). Joint Models for Multivariate Longitudinal and Multivariate Survival Data. *Biometrics, 62*(2), 432-445. doi: 10.1111/j.1541-0420.2005.00448.x

CHMP. (2010). Guideline on missing data in confirmatory clinical trials. Retrieved 1 November 2013, from http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/09/WC500096793.pdf

Churchyard, G. J., Kleinschmidt, I., Corbett, E. L., Mulder, D., Smit, J., & de Kock, K. M. (2000). Factors associated with an increased case-fatality rate in HIV-infected and non-infected South African gold miners with pulmonary tuberculosis. *International Journal of Tuberculosis Lung Disease, 4*, 705-712.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6*(4), 330-351.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society Series B, 34*(2), 187-220.

Cox, D. R. (1975). Partial likelihood. *Biometrika, 72*(2), 269-276.

Curran, D., Molenberghs, G., Thijs, H., & Verbeke, G. (2004). Sensitivity Analysis for Pattern Mixture Models. *Journal of Biopharmaceutical Statistics, 14*(1), 125-143. doi: 10.1081/BIP-120028510

Daniels, M., & Hughes, M. D. (1997). Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine, 16*, 1965-1982.

Daniels, M. J., & Hogan, J. W. (2000). Reparameterizing the Pattern Mixture Model for Sensitivity Analyses Under Informative Dropout. *Biometrics, 56*(4), 1241-1248. doi: 10.1111/j.0006-341X.2000.01241.x

Daniels, M. J., & Hogan, J. W. (2008). *Missing data in longitudinal studies. Strategies for Bayesian modeling and sensitivity analysis* (1 ed.). Boca Raton, FL: Chapman & Hall/CRC.

Day, C., Gray, A., & Budgell, E. (2012). South African Health Review: Health and related indicators. Durban: Health Systems Trust.

Degruttola, V., & Tu, X. M. (1994). Modeling progression of CD-4 lymphocyte count and its relationship to survival time. *Biometrics, 50*, 1003-1014.

Demirtas, H., & Schafer, J. L. (2003). On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out. *Statistics in Medicine, 22*(16), 2553-2575. doi: 10.1002/sim.1475

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

Department of Health. (2004). The South African National Tuberculosis Control Programme: Practical Gudielines 2004. Pretoria: South African Department of Health.

Diggle, P. (2002). *Analysis of longitudinal data* (2nd ed.). Oxford ; New York: Oxford University Press.

Diggle, P., & Kenward, M. G. (1994). Informative Drop-Out in Longitudinal Data Analysis. *Applied Statistics, 43*(1), 49-93.

Drukker, D. M. (2011). *Missing-Data methods (box set) (Advances in Econometrics)*. Bingley, United Kingdom: Emerald Group Publishing Limited.

Dufouil, C., Brayne, C., & Clayton, D. (2004). Analysis of longitudinal studies with death and drop-out: a case study. *Statistics in Medicine, 23*(14), 2215-2226. doi: 10.1002/sim.1821

Enders, C. K. (2010). *Applied missing data analysis (Methodology in social sciences)*. New York, NY: The Guilford press.

Fegert, J., Kolch, M., Zito, J. M., Glaeske, G., & Janhsen, K. (2006). Antidepressant Use in Children and Adolescents in Germany. *Journal of Child and Adolescent Psychopharmacology, 16*(1-2), 197-206. doi: 10.1089/cap.2006.16.197

Fleming, T. (2011). Addressing missing data in clinical trials. *Annals of internal medicine, 154*(2), 113-117.

Follmann, D., & Wu, M. C. (1995). An approximate generalized linear model with random effects for informative missing data. *Biometrics, 51*, 151-168.

Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics, 58*, 21-29.

Friedman, L. M., Furberg, C. D., & DeMets, D. L. (1998). *Fundamentals of clinical trials* (Third ed.). New York: Springer.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian Data Analysis* (Second ed.). London: Chapman and Hall.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6*, 721-741. doi: 10.1109/TPAMI.1984.4767596

Gilks, W. R., Richardson, S., & Spiegelhalter, D. (1996). *Markov chain Monte Carlo in practice* (First ed.). London: Chapman and Hall.

Girardi, E., Palmieri, F., Cingolani, A., Ammassari, A., Petrosillo, N., Gillini, L., Zinzi, D., de Luca, A., Antinori, A., & Ippolito, G. (2001). Changing clinical presentation and survival in HIV-associated tuberculosis after highly active antiretroviral therapy. *Journal of acquired immune deficiency syndromes 26*(4), 326-331.

Gould, A. L. (1980). A new approach to the analysis of clinical drug trials with withdrawals. *Biometrics, 36*(4), 721-727.

Graham, J. W. (2012). *Missing data: Analysis and design (Statistics for behavioral sciences)*. New York, NC: Springer.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory. *Prevention Science, 8*(3), 206-213. doi: 10.1007/s11121-007-0070-9

Gray, C., Malatsi, N., Riou, C., & de Assis Rosa, D. (2010). Cellular immunity in HIV: a synthesis of responses to preserve self. In S. S. Abdool Karim & Q. Abdool Karim (Eds.), *HIV/AIDS in South Africa* (2 ed.). Cambridge: Cambridge University Press.

Groenwold, R. H. H., Donders, A. R. T., Roes, K. C. B., Harrell, F. E., & Moons, K. G. M. (2012). Dealing With Missing Outcome Data in Randomized Trials and Observational Studies. *American Journal of Epidemiology, 175*(3), 210-217. doi: 10.1093/aje/kwr302

Hamilton, M. (1960). A rating scale for depression. *Journal of neurology, neurosurgery and psychiatry 23*, 56-62. doi: 10.1136/jnnp.23.1.56

Health Systems Trust. (2008). District Health Barometer 2006/2007. Accessible at http: www.hst.org.za/publications/717 [Accessed: January 2009].

Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods, 2*(1), 64-78.

Hedeker, D., & Mermelstein, R. J. (2007). Mixed-effects regression models with heterogeneous variance: analyzing ecological momentary assessment (EMA) data of smoking. In T. D. Little, J. A. Bovaird & N. A. Card (Eds.), *Modeling contextual effects in longitudinal studies*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc., Publishers.

Henderson, R., Diggle, P., & Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics, 1*(4), 465-480. doi: 10.1093/biostatistics/1.4.465

Hogan, J. W., & Laird, N. M. (1997). Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine, 16*(3), 239-257. doi: 10.1002/(sici)1097-0258(19970215)16:3<239::aid-sim483>3.0.co;2-x

Hogan, J. W., Roy, J., & Korkontzelou, C. (2004). Handling drop-out in longitudinal studies. *Statistics in Medicine, 23*(9), 1455-1497. doi: 10.1002/sim.1728

Horton, N. J., & Kleinman, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *American Statistician, 61*(1), 79-90. doi: 10.1198/000313007X172556

Horton, N. J., & Switzer, S. S. (2005). Statistical methods in the journal. *New England Journal of Medicine, 353*(18), 1977-1979. doi: 353/18/1977 [pii] 10.1056/NEJM200511033531823

Hulley, S. B., Cummings, S. R., Browner, W. S., Grady, D. G., & Newman, T. B. (2007). *Designing clinical research*. Philadelphia: Lippicott Williams & Wilkins.

Hypericum Depression Trial Study Group. (2002). Effect of Hypericum perforatum (St John's Wort) in Major Depressive Disorder. *Journal of the American Medical Association, 287*(14), 1807-1814. doi: 10.1001/jama.287.14.1807

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., & Herring, A. H. (2005). Missing-Data Methods for Generalized Linear Models. *Journal of the American Statistical Association, 100*(469), 332-346. doi: 10.1198/016214504000001844

ICH E9 Expert Working Group. (1999). Statistical principles for clinical trials (ICH E9): an introductory note on an international guideline. *Statistics in Medicine, 18*(15), 1903-1942. doi: 10.1002/(SICI)1097-0258(19990815)18:15<1903::AID-SIM188>3.0.CO;2-F [pii]

Janssens, M., Molenberghs, G., & Kerstens, R. (2012). Handling of missing data in long-term clinical trials: a case study. *Pharmaceutical Statistics, 11*, 442-448.

Joint United Nations Programme on HIV/AIDS (UNAIDS). (2010). *Global report: UNAIDS report on the global AIDS epidemic, 2010*. Geneva: UNAIDS.

Joint United Nations Programme on HIV/AIDS UNAIDS. (2012). UNAIDS Report on the global AIDS epidemic. Geneva: UNAIDS.

Kalb, R., Trautmann-Sponsel, R. D., & Kieser, M. (2001). Efficacy and Tolerability of Hypericum Extract WS 5572 versus Placebo in Mildly to Moderately Depressed Patients. *Pharmacopsychiatry, 34*(03), 96,103. doi: 10.1055/s-2001-14280

Kalbfleisch, J. D., & Prentice, R. L. (2002). *The statistical analysis of failure time data* (Vol. Second edition). New York: John Wiley.

Keene, O. N. (2011). Intent-to-treat analysis in the presence of off-treatment or missing data. *Pharmaceutical Statistics, 10*(3), 191-195. doi: 10.1002/pst.421

Kenward, M. G. (2006). *Workshop on the handling of missing data in clinical trials*. 49th annual conference of the South African Statistical Association. Stellenbosch.

Kenward, M. G., & Molenberghs, G. (1998). Likelihood based frequentist inference when data are missing at random. *Statistical science, 13*(3), 236-247.

Kenward, M. G., & Molenberghs, G. (2009). Last Observation Carried Forward: A Crystal Ball? *Journal of Biopharmaceutical Statistics, 19*(5), 872-888. doi: 10.1080/10543400903105406

Kenward, M. G., Molenberghs, G., & Thijs, H. (2003). Pattern mixture models with proper time dependence. *Biometrika, 90*(1), 53-71. doi: 10.1093/biomet/90.1.53

Kenward, M. G., & Roger, J. H. (1997). Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood. *Biometrics, 53*(3), 983-997.

Kurland, B. F., Johnson, L. L., Egleston, B. L., & Diehr, P. H. (2009). Longitudinal data with follow-up truncated by death: match the analysis method to research aims. *Statistical Science, 24*(2), 211-222.

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

Li, X., Mehrotra, D. V., & Barnard, J. (2006). An analysis of incomplete longitudinal binary data using multiple imputation. *Statistics in Medicine, 25*, 2107-2124.

Linde, K., Berner, M. M., & Kriston, L. (2008). St John's wort for major depression. *Cochrane Database of Systematic Reviews, 4*(4). doi: 10.1002/14651858.CD000448.pub3

Linde, K., Ramirez, G., Mulrow, C. D., Pauls, A., Weidenhammer, W., & Melchart, D. (1996). St John's wort for depression - an overview and meta-analysis of randomised clinical trials. *British Medical Journal, 313*(7052), 253-258. doi: 10.1136/bmj.313.7052.253

Little, R. (1993). Pattern-Mixture Models for Multivariate Incomplete Data. *Journal of the American Statistical Association, 88*(421), 125-134.

Little, R. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika, 81*(3), 471-483.

Little, R., & Yau, L. (1996). Intent-to-Treat Analysis for Longitudinal Studies with Drop-Outs. *Biometrics, 52*(4), 1324-1333.

Little, R. J., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., Frangakis, C., Hogan, J. W., Molenberghs, G., Murphy, S. A., Neaton, J. D., Rotnitzky, A., Scharfstein, D., Shih, W. J., Siegel, J. P., & Stern, H. (2012). The Prevention and Treatment of Missing Data in Clinical Trials. *New England Journal of Medicine, 367*(14), 1355-1360. doi: 10.1056/NEJMsr1203730

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data* (2 ed.). New York: Wiley.

Little, R. J. A., & Wang, Y. (1996). Pattern-Mixture Models for Multivariate Incomplete Data with Covariates. *Biometrics, 52*(1), 98-111.

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

Liublinska, V., & Rubin, D. B. (2012). Re: “Dealing With Missing Outcome Data in Randomized Trials and Observational Studies”. *American Journal of Epidemiology, 176*(4), 357-358. doi: 10.1093/aje/kws215

Mallinckrodt, C. (2013). *Preventing and Treating Missing Data in Longitudinal Clinical Trials: A Practical Guide*. New York, NY: Cambridge university press.

Mallinckrodt, C. H., Clark, W. S., Carroll, R. J., & Molenberghs, G. (2003). Assessing response profiles from incomplete longitudinal clinical trial data under regulatory considerations. *Journal of Biopharmaceutical Statistics, 13*(2), 179-190.

Mallinckrodt, C. H., Clark, W. S., & David, S. R. (2001). Type I error rates from mixed-effects model repeated measures versus fixed effects analysis of variance with missing values imputed via last observation carried forward. *Drug Information Journal, 35*, 1215-1225. doi: 10.1177/009286150103500418

Mallinckrodt, C. H., Lin, Q., Lipkovich, I., & Molenberghs, G. (2012). A structured approach to choosing estimands and estimators in longitudinal clinical trials. *Pharmaceutical Statistics, 11*(6), 456-461. doi: 10.1002/pst.1536

Mallinckrodt, C. H., Sanger, T. M., Dube, S., DeBrota, D. J., Molenberghs, G., Carroll, R. J., Potter, W. Z., & Tollefson, G. D. (2003). Assessing and interpreting treatment effects in longitudinal clinical trials with missing data. *Biological Psychiatry, 53*(8), 754-760. doi: S000632230201867X [pii]

Martin, D. J., Sim, J. G. M., Sole, G. J., Rymer, L., Shalekoff, S., van Niekerk, A. B. N., Becker, P., Weilbach, C. N., Iwanik, J., Keddy, K., Miller, G. B., Ozbay, B., Ryan, A., Viscovic, T., & Woolf, M. (1995). CD4+ lymphocyte count in African patients co-infected with HIV and tuberculosis. *Journal of acquired immune deficiency syndromes and Human retrovirology, 8*, 386-391.

Mason, A., Richardson, S., Plewis, I., & Best, N. (2012). Strategy for modelling nonrandom missing data mechanisms in observational studies using Bayesian methods. *Journal of offician statistics, 28*(2), 279-302.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. Boca Raton: Chapman & Hall.

Michiels, B., Molenberghs, G., Bijnens, L., Vangeneugden, T., & Thijs, H. (2002). Selection models and pattern-mixture models to analyse longitudinal quality of life data subject to drop-out. *Statistics in Medicine, 21*(8), 1023-1041.

Molenberghs, G., Beunckens, C., Sotto, C., & Kenward, M. G. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70*(2), 371-388. doi: 10.1111/j.1467-9868.2007.00640.x

Molenberghs, G., & Kenward, M. G. (2007). *Missing data in Clinical Studies*. Chichester: Wiley.

Molenberghs, G., Kenward, M. G., & Goetghebeur, E. (2001). Sensitivity analysis for incomplete contingency tables: the Slovenian plebiscite case. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 50*(1), 15-29. doi: 10.1111/1467-9876.00217

Molenberghs, G., Michiels, B., Kenward, M. G., & Diggle, P. J. (1998). Monotone missing data and pattern-mixture models. *Statistica Neerlandica, 52*(2), 153-161. doi: 10.1111/1467-9574.00075

Molenberghs, G., Thijs, H., Jansen, I., Beunckens, C., Kenward, M. G., Mallinckrodt, C., & Carroll, R. J. (2004). Analyzing incomplete longitudinal clinical trial data. *Biostatistics, 5*(3), 445-464. doi: 10.1093/biostatistics/kxh001

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

Mukadi, Y. D., Maher, D., & Harries, A. (2001). Tuberculosis case fatality rates in high HIV prevalence populations in sub-Saharan Africa. *AIDS, 15*, 143-152.

National Department of Health. (2004). National Antiretroviral Treatment Guidelines. First Edition. Pretoria: South African Department of Health.

National Research Council. (2010). *The prevention and treatment of missing data in clinical trials*. Washington DC: The National Academic Press.

Nevalainen, J., Kenward, M. G., & Virtanen, S. M. (2009). Missing values in longitudinal dietary data: A multiple imputation approach based on a fully conditional specification. *Statistics in Medicine, 28*(29), 3657-3669. doi: 10.1002/sim.3731

O'Neill, R. T., & Temple, R. (2012). The Prevention and Treatment of Missing Data in Clinical Trials: An FDA Perspective on the Importance of Dealing With It. *Clinical Pharmacology and Therapeutics, 91*(3), 550-554.

Piscitelli, S. C., & Gallicano, K. D. (2001). Interactions among drugs for HIV and opportunistic infections. *New England Journal of Medicine, 344*, 984-996.

Rahimi, R., Nikfar, S., & Abdollahi, M. (2009). Efficacy and tolerability of Hypericum perforatum in major depressive disorder in comparison with selective serotonin reuptake inhibitors: A meta-analysis. *Progress in Neuro-Psychopharmacology and Biological Psychiatry, 33*(1), 118-127. doi: http://dx.doi.org/10.1016/j.pnpbp.2008.10.018

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data with applications in R*. Boca Raton: Chapman & Hall.

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

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

Robins, J. M., Greenland, S., & Hu, F. C. (1999). Reply to comments on Estimation of the Causal Effect of a Time-Varying Exposure on the Marginal Mean of a Repeated Binary Outcome. *Journal of the American Statistical Association, 94*(447), 708-712.

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

Rothman, K. J., Greenland, S., & Lash, T. L. (2008). *Modern epidemiology* (3 ed.). Phaladelphia: Lippicott Williams & Wilkins.

Rotnitzky, A., Robins, J. M., & Scharfstein, D. O. (1998). Semiparametric Regression for Repeated Outcomes with Nonignorable Nonresponse. *Journal of the American Statistical Association, 93*(444), 1321-1339.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581-592. doi: 10.1093/biomet/63.3.581

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons. Inc.

Rubin, D. B. (1996). Multiple Imputation After 18+ Years. *Journal of the American Statistical Association, 91*(434), 473-489.

Rubin, D. B., & Frangakis, C. E. (1999). Comment on Estimation of the Causal Effect of a Time-Varying Exposure on the Marginal Mean of a Repeated Binary Outcome. *Journal of the American Statistical Association, 94*(447), 702-704.

Sarris, J., Fava, M., Schweitzer, I., & Mischoulon, D. (2012). St John's Wort (Hypericum perforatum) versus Sertraline and Placebo in Major Depressive Disorder: Continuation Data from a 26-Week RCT. *Pharmacopsychiatry, 45*(07), 275-278. doi: 10.1055/s-0032-1306348

SAS Institute Inc. (2004). *SAS/STAT® 9.1 User's Guide*. Cary, NC: SAS Institute Inc.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data* United States of America: Chapman & Hall.

Scharfstein, D. O., Rotnitzky, A., & Robins, J. M. (1999). Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models. *Journal of the American Statistical Association, 94*(448), 1096-1120.

Schluger, N. W. (1999). Issues in the treatment of Active Tuberculosis in Human Immunodeficiency virus–infected participants. *Clinical Infectious Diseases, 28*, 130-135.

Schulz, K. F., Altman, D. G., Moher, D., & for the, C. G. (2010). CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *British Medical Journal, 340*, c332-. doi: 10.1136/bmj.c332

Shelton, R. C., Keller, M. B., Gelenberg, A., Dunner, D. L., Hirschfeld, R., Thase, M. E., Russell, J., Lydiard, B., Crits-Christoph, P., Gallop, R., Todd, L., Hellerstein, D., Goodnick, P., Keitner, G., Stahl, S. M., & Halbreich, U. (2001). Effectiveness of St John's wort in major depression: A randomized controlled trial. *JAMA, 285*(15), 1978-1986. doi: 10.1001/jama.285.15.1978

South African Department of Health. (2007). Tuberculosis Strategic Plan for South Africa, 2007-2011. Pretoria: Department of Health.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*(4), 583-639. doi: 10.1111/1467-9868.00353

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Gilks, W. R. (1995). *BUGS: Bayesian Inference Using Gibbs Sampling* (M. B. Unit Ed.). Cambridge: MRC Biostatistics Unit.

Tack, J., Van Outryve, M., Beyens, G., Kerstens, R., & Vandeplassche, L. (2009). Prucalopride (Resolor) in the treatment of severe chronic constipation in patients dissatisfied with laxatives. *Gut, 58*, 357-365.

Tan, M. T., Tian, G.-L., & Ng, K. W. (2012). *Bayesian missing data problems: EM, data augmentation and noniterative computation* London: Chapman and Hall/CRC.

Thijs, H., Molenberghs, G., Michiels, B., Verbeke, G., & Curran, D. (2002). Strategies to fit pattern-mixture models. *Biostatistics, 3*(2), 245-265. doi: 10.1093/biostatistics/3.2.245

Tsiatis, A. (2010). *Semiparametric theory and missing data*. New York, NY: Springer.

Tsiatis, A. A., & Davidian, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika, 88*(2), 447-458. doi: 10.1093/biomet/88.2.447

Tsiatis, A. A., & Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica, 14*, 809-834.

Tsiatis, A. A., Degruttola, V., & Wulfsohn, M. S. (1995). Modeling the Relationship of Survival to Longitudinal Data Measured with Error. Applications to Survival and CD4 Counts in Patients with AIDS. *Journal of the American Statistical Association, 90*(429), 27-37. doi: 10.1080/01621459.1995.10476485

Twisk, J. W. R. (2013). *Applied longitudinal data analysis for epidemiology*. Cambridge: Cambridge University Press.

Vansteelandt, S., Carpenter, J., & Kenward, M. G. (2010). Analysis of incomplete data using inverse probability weighting and doubly robust estimators. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 6*(1), 37-48.

Vansteelandt, S., Goetghebeur, E., Kenward, M. G., & Molenberghs, G. (2006). Ignorance and uncertainty regions as inferential tools in sensitivity analysis. *Statistica Sinica, 16*, 953-979.

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

Wang, Y., & Taylor, J. M. G. (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association, 96*(895-905).

Ware, J. H., Harrington, D., Hunter, D. J., & D'Agostino, R. B. (2012). Missing Data. *New England Journal of Medicine, 367*(14), 1353-1354. doi: 10.1056/NEJMsm1210043

White, I. R., Carpenter, J., Evans, S., & Schroter, S. (2007). Eliciting and using expert opinions about dropout bias in randomized controlled trials. *Clinical Trials, 4*(2), 125-139. doi: 10.1177/1740774507077849

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine, 30*(4), 377-399. doi: 10.1002/sim.4067

Wood, A. M., White, I. R., & Thompson, S. G. (2004). Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials, 1*(4), 368-376. doi: 10.1191/1740774504cn032oa

World Health Organisation. (2003). Treatment of Tuberculosis: guidelines for national programmes. 3rd Edition. Geneva, Switzerland: World Health Organisation.

World Health Organisation. (2011). Global TB Database (pp. Available from: http://www.who.int/tb/country/data/download/en/index.html.). Geneva, Switzerland.

World Health Organization. (2012). Global tuberculosis report 2012. Geneva: World Health Organization.

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

Wu, M. C., & Carroll, R. J. (1988). Estimation and Comparison of Changes in the Presence of Informative Right Censoring by Modeling the Censoring Process. *Biometrics, 44*(1), 175-188.

**Appendix 1**
**The extension of the delta method in Section 3.3.2.1 to four groups**

For four groups, say A, B, C and D, let the population weights be $\pi_a$, $\pi_b$, $\pi_c$ and $\pi_d$ and

$$\pi_a + \pi_b + \pi_c + \pi_d = 1,$$

$$\text{so that } \pi_d = 1 - (\pi_a + \pi_b + \pi_c).$$

The parameters for each of the patterns are $\beta_A$, $\beta_B$, $\beta_C$ and $\beta_D$. Let

$$\boldsymbol{\theta} = (\pi_a, \pi_b, \pi_c, \beta_A, \beta_B, \beta_C, \beta_D)$$

and as in the two group case, let

$$g(\boldsymbol{\theta}) = \pi_a\beta_A + \pi_b\beta_B + \pi_c\beta_C + \pi_d\beta_D$$

which can also be written as

$$g(\boldsymbol{\theta}) = \pi_a\beta_A + \pi_b\beta_B + \pi_C\beta_C + [1 - (\pi_a + \pi_b + \pi_c)]\beta_D \text{ and}$$

$$\mathrm{V}(\widehat{\boldsymbol{\theta}}) = \begin{bmatrix} var(\hat{\pi}_a) & cov_{AB} & cov_{AC} & 0 & 0 & 0 & 0 \\ cov_{AB} & var(\hat{\pi}_b) & cov_{BC} & 0 & 0 & 0 & 0 \\ cov_{AC} & cov_{BC} & var(\hat{\pi}_c) & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & var(\hat{\beta}_A) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & var(\hat{\beta}_B) & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & var(\hat{\beta}_C) & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & var(\hat{\beta}_D) \end{bmatrix}$$

where $cov_{AB}$= *cov*($\hat{\pi}_a$, $\hat{\pi}_b$),

$cov_{BC}$= *cov*($\hat{\pi}_b$, $\hat{\pi}_c$),

$cov_{AC}$= *cov*($\hat{\pi}_a$, $\hat{\pi}_c$).

The assumption is made that the four dropout status groups are independent, therefore the covariances between the *β*'s are *0*. However, the proportions $\hat{\pi}_a$, $\hat{\pi}_b$ and $\hat{\pi}_c$ are not independent and the covariances are not *0* in that case. The required derivatives are:

$$\frac{\partial g(\boldsymbol{\theta})}{\partial \pi_a} = \beta_A - \beta_D$$

$$\frac{\partial g(\boldsymbol{\theta})}{\partial \pi_b} = \beta_B - \beta_D$$

$$\frac{\partial g(\boldsymbol{\theta})}{\partial \pi_c} = \beta_C - \beta_D$$

$$\frac{\partial g(\boldsymbol{\theta})}{\partial \beta_A} = \pi_a$$

$$\frac{\partial g(\boldsymbol{\theta})}{\partial \beta_B} = \pi_b$$

$$\frac{\partial g(\boldsymbol{\theta})}{\partial \beta_C} = \pi_c$$

$$\frac{\partial g(\boldsymbol{\theta})}{\partial \beta_{\mathrm{D}}} = 1 - (\pi_a + \pi_b + \pi_c).$$

$$\mathrm{V}\left[\mathrm{g}(\widehat{\boldsymbol{\theta}})\right] = \left[\hat{\beta}_A - \hat{\beta}_D, \hat{\beta}_B - \hat{\beta}_D, \hat{\beta}_C - \hat{\beta}_D, \pi_{\mathrm{a}}, \pi_{\mathrm{b}}, \pi_{\mathrm{c}}, \{1 - (\pi_{\mathrm{a}} + \pi_{\mathrm{b}} + \pi_{\mathrm{c}})\}\right]\mathrm{V}(\hat{\theta})\left[\frac{\partial g(\boldsymbol{\theta})}{\partial \theta}\right]$$

$$= [(\hat{\beta}_A - \hat{\beta}_D)var(\hat{\pi}_a) + (\hat{\beta}_B - \hat{\beta}_D)cov_{AB} + (\hat{\beta}_C - \hat{\beta}_D)cov_{AC},$$
$$(\hat{\beta}_A - \hat{\beta}_D)cov_{AB} + (\hat{\beta}_B - \hat{\beta}_D)var(\hat{\pi}_b) + (\hat{\beta}_C - \hat{\beta}_D)cov_{BC},$$
$$(\hat{\beta}_A - \hat{\beta}_D)cov_{AC} + (\hat{\beta}_B - \hat{\beta}_D)cov_{BC} + (\hat{\beta}_C - \hat{\beta}_D)var(\hat{\pi}_c),$$
$$\pi_a var(\hat{\beta}_A), \pi_b var(\hat{\beta}_B), \pi_c var(\hat{\beta}_C), \quad var(\hat{\beta}_D)\{1 - (\pi_a + \pi_b + \pi_c)\}]\left[\frac{\partial g(\boldsymbol{\theta})}{\partial \theta}\right]$$

$$= (\hat{\beta}_A - \hat{\beta}_D)^2 var(\hat{\pi}_a) + (\hat{\beta}_A - \hat{\beta}_D)(\hat{\beta}_B - \hat{\beta}_D)cov_{AB} + (\hat{\beta}_A - \hat{\beta}_D)(\hat{\beta}_C - \hat{\beta}_D)cov_{AC}$$
$$+ (\hat{\beta}_B - \hat{\beta}_D)(\hat{\beta}_A - \hat{\beta}_D)cov_{AB} + (\hat{\beta}_B - \hat{\beta}_D)^2 var(\hat{\pi}_b)$$
$$+ (\hat{\beta}_C - \hat{\beta}_D)(\hat{\beta}_B - \hat{\beta}_D)cov_{BC} + (\hat{\beta}_A - \hat{\beta}_D)(\hat{\beta}_C - \hat{\beta}_D)cov_{AC}$$
$$+ (\hat{\beta}_B - \hat{\beta}_D)(\hat{\beta}_C - \hat{\beta}_D)cov_{BC} + (\hat{\beta}_C - \hat{\beta}_D)^2 var(\hat{\pi}_c) + \pi_a{}^2 var(\hat{\beta}_A)$$
$$+ \pi_b{}^2 var(\hat{\beta}_B) + \pi_c{}^2 var(\hat{\beta}_C) + \{1 - (\pi_a + \pi_b + \pi_c)\}^2 var(\hat{\beta}_D)$$

$$= (\hat{\beta}_A - \hat{\beta}_D)^2 var(\hat{\pi}_a) + 2(\hat{\beta}_A - \hat{\beta}_D)(\hat{\beta}_B - \hat{\beta}_D)cov_{AB} + 2(\hat{\beta}_A - \hat{\beta}_D)(\hat{\beta}_C - \hat{\beta}_D)cov_{AC}$$
$$+ (\hat{\beta}_B - \hat{\beta}_D)^2 var(\hat{\pi}_b) + 2(\hat{\beta}_C - \hat{\beta}_D)(\hat{\beta}_B - \hat{\beta}_D)cov_{BC} + (\hat{\beta}_C - \hat{\beta}_D)^2 var(\hat{\pi}_c)$$
$$+ \pi_a{}^2 var(\hat{\beta}_A) + \pi_b{}^2 var(\hat{\beta}_B) + \pi_c{}^2 var(\hat{\beta}_C) + \{1 - (\pi_a + \pi_b + \pi_c)\}^2 var(\hat{\beta}_D)$$

$$cov_{AB} = cov_{BA} = cov(\hat{\pi}_a, \hat{\pi}_b) = \frac{-\pi_a \pi_b}{N}$$
$$cov_{AC} = cov_{CA} = cov(\hat{\pi}_a, \hat{\pi}_c) = \frac{-\pi_a \pi_c}{N}$$
$$cov_{CB} = cov_{BC} = cov(\hat{\pi}_b, \hat{\pi}_c) = \frac{-\pi_c \pi_b}{N}$$
$$var(\hat{\pi}_a) = \frac{\pi_a(1 - \pi_a)}{N}$$
$$var(\hat{\pi}_b) = \frac{\pi_b(1 - \pi_b)}{N}$$
$$var(\hat{\pi}_c) = \frac{\pi_c(1 - \pi_c)}{N}$$

If $cov_{AB}$, $cov_{AC}$, $cov_{CB}$, var($\hat{\pi}_a$) var($\hat{\pi}_b$) and var($\hat{\pi}_c$) are replaced in $\mathrm{V}[\mathrm{g}(\hat{\boldsymbol{\theta}})]$ above the variance is estimated by the following:

$$\widehat{var}[g(\hat{\boldsymbol{\theta}})] = (\hat{\beta}_A - \hat{\beta}_D)^2 \frac{\hat{\pi}_a(1 - \hat{\pi}_a)}{N} - 2\frac{\hat{\pi}_a \hat{\pi}_b}{N}(\hat{\beta}_A - \hat{\beta}_D)(\hat{\beta}_B - \hat{\beta}_D) - 2\frac{\hat{\pi}_a \hat{\pi}_c}{N}(\hat{\beta}_A - \hat{\beta}_D)(\hat{\beta}_C - \hat{\beta}_D)$$
$$+ (\hat{\beta}_B - \hat{\beta}_D)^2 \frac{\hat{\pi}_b(1 - \hat{\pi}_b)}{N} - 2\frac{\hat{\pi}_c \hat{\pi}_b}{N}(\hat{\beta}_C - \hat{\beta}_D)(\hat{\beta}_B - \hat{\beta}_D) + (\hat{\beta}_C - \hat{\beta}_D)^2 \frac{\hat{\pi}_c(1 - \hat{\pi}_c)}{N}$$
$$+ \hat{\pi}_a^2 var(\hat{\beta}_A) + \hat{\pi}_b^2 var(\hat{\beta}_B) + \hat{\pi}_c^2 var(\hat{\beta}_C) + [1 - (\hat{\pi}_a + \hat{\pi}_b + \hat{\pi}_c)]^2 var(\hat{\beta}_D)$$

This formulation of the equation is not in the most natural form to fit the model. The following changes can be made to write the formula in a form that is easier to use:

As above:

$$\pi_a + \pi_b + \pi_c + \pi_d = 1,$$

$$\text{so that } \pi_d = 1 - (\pi_a + \pi_b + \pi_c)$$

$$\text{and } \overline{\boldsymbol{\beta}} = g(\boldsymbol{\theta}) = \pi_a\beta_A + \pi_b\beta_B + \pi_c\beta_C + \pi_d\beta_D$$

which can also be written as

$$g(\boldsymbol{\theta}) = \pi_a\beta_A + \pi_b\beta_B + \pi_c\beta_C + [1 - (\pi_a + \pi_b + \pi_c)]\beta_D.$$

A change is made by defining

$$\beta_B = \beta_A + \beta_{\Delta 1}$$

$$\beta_C = \beta_A + \beta_{\Delta 2} \text{ and } \beta_D = \beta_A + \beta_{\Delta 3}$$

with $\beta_{\Delta 1}$ indicating how group B differs from group A, $\beta_{\Delta 2}$ indicating how group C differs from group A and $\beta_{\Delta 3}$ indicating how group D differs from group A then

$$g(\boldsymbol{\theta}) = \overline{\boldsymbol{\beta}} = \pi_a\beta_A + \pi_b\beta_B + \pi_c\beta_C + \pi_d\beta_D$$

$$= \pi_a\beta_A + \pi_b(\beta_A + \beta_{\Delta 1}) + \pi_c(\beta_A + \beta_{\Delta 2}) + [1 - (\pi_a + \pi_b + \pi_c)](\beta_A + \beta_{\Delta 3})$$

$$= \pi_a\beta_A + \pi_b\beta_A + \pi_c\beta_A + \beta_A - (\pi_a + \pi_b + \pi_c)(\beta_A + \beta_{\Delta 3}) + \pi_b\beta_{\Delta 1} + \pi_c\beta_{\Delta 2} + \beta_{\Delta 3}$$

$$= \beta_A + \pi_b\beta_{\Delta 1} + \pi_c\beta_{\Delta 2} + \pi_d\beta_{\Delta 3}$$

Define $\boldsymbol{\theta} = (\beta_A, \beta_{\Delta 1}, \beta_{\Delta 2}, \beta_{\Delta 3}, \pi_b, \pi_c, \pi_d)$. It then follows that

$$\frac{\partial g(\boldsymbol{\theta})}{\partial \beta_{\mathrm{A}}} = 1$$

$$\frac{\partial g(\boldsymbol{\theta})}{\partial \beta_{\Delta 1}} = \pi_b$$

$$\frac{\partial g(\boldsymbol{\theta})}{\partial \beta_{\Delta 2}} = \pi_c$$

$$\frac{\partial g(\boldsymbol{\theta})}{\partial \beta_{\Delta 3}} = \pi_d$$

$$\frac{\partial g(\boldsymbol{\theta})}{\partial \pi_{\mathrm{B}}} = \beta_{\Delta 1}$$

$$\frac{\partial g(\boldsymbol{\theta})}{\partial \pi_{\mathrm{C}}} = \beta_{\Delta 2}$$

$$\frac{\partial g(\boldsymbol{\theta})}{\partial \pi_{\mathrm{D}}} = \beta_{\Delta 3}.$$

$$\mathrm{V}(\widehat{\boldsymbol{\theta}}) = \begin{bmatrix} \mathrm{var}(\hat{\beta}_A) & \mathrm{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 1}) & \mathrm{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 2}) & \mathrm{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 3}) & 0 & 0 & 0 \\ \mathrm{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 1}) & \mathrm{var}(\hat{\beta}_{\Delta 1}) & 0 & 0 & 0 & 0 & 0 \\ \mathrm{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 2}) & 0 & \mathrm{var}(\hat{\beta}_{\Delta 2}) & 0 & 0 & 0 & 0 \\ \mathrm{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 3}) & 0 & 0 & \mathrm{var}(\hat{\beta}_{\Delta 3}) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathrm{var}(\hat{\pi}_b) & \mathrm{cov}(\hat{\pi}_b, \hat{\pi}_c) & \mathrm{cov}(\hat{\pi}_b, \hat{\pi}_d) \\ 0 & 0 & 0 & 0 & \mathrm{cov}(\hat{\pi}_b, \hat{\pi}_c) & \mathrm{var}(\hat{\pi}_c) & \mathrm{cov}(\hat{\pi}_c, \hat{\pi}_d) \\ 0 & 0 & 0 & 0 & \mathrm{cov}(\hat{\pi}_b, \hat{\pi}_d) & \mathrm{cov}(\hat{\pi}_c, \hat{\pi}_d) & \mathrm{var}(\hat{\pi}_d) \end{bmatrix}$$

$$\left(\frac{\partial g(\boldsymbol{\theta})}{\partial \theta}\right) \mathbf{V}(\widehat{\boldsymbol{\theta}}) \left(\frac{\partial g(\boldsymbol{\theta})}{\partial \theta}\right)' = [1, \pi_b, \pi_c, \pi_d, \beta_{\Delta 1}, \beta_{\Delta 2}, \beta_{\Delta 3}] \mathbf{V}(\widehat{\boldsymbol{\theta}}) \begin{bmatrix} 1 \\ \pi_b \\ \pi_c \\ \pi_d \\ \beta_{\Delta 1} \\ \beta_{\Delta 2} \\ \beta_{\Delta 3} \end{bmatrix}$$

$$= \left[\mathrm{var}(\hat{\beta}_A) + \pi_b \,\mathrm{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 1}) + \pi_c \,\mathrm{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 2}) + \pi_d \,\mathrm{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 3}),\ \mathrm{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 1}) + \pi_b \,\mathrm{var}(\hat{\beta}_{\Delta 1}),\ \mathrm{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 2}) + \pi_c \,\mathrm{var}(\hat{\beta}_{\Delta 2}),\ \mathrm{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 3}) + \pi_d \,\mathrm{var}(\hat{\beta}_{\Delta 3}),\right.$$
$$\beta_{\Delta 1} \,\mathrm{var}(\hat{\pi}_b) + \beta_{\Delta 2} \,\mathrm{cov}(\hat{\pi}_b, \hat{\pi}_c) + \beta_{\Delta 3} \,\mathrm{cov}(\hat{\pi}_b, \hat{\pi}_d),\ \beta_{\Delta 1} \mathrm{cov}(\hat{\pi}_b, \hat{\pi}_c) + \beta_{\Delta 2} \,\mathrm{var}(\hat{\pi}_c) + \beta_{\Delta 3}\, cov(\hat{\pi}_c, \hat{\pi}_d),\ \beta_{\Delta 1} \mathrm{cov}(\hat{\pi}_b, \hat{\pi}_d) + \beta_{\Delta 2}\, cov(\hat{\pi}_c, \hat{\pi}_d) +$$
$$\left.\beta_{\Delta 3}\, var(\hat{\pi}_d)\right]\ [1, \pi_b, \pi_c, \pi_d, \beta_{\Delta 1}, \beta_{\Delta 2}, \beta_{\Delta 3}]'$$

$$= \text{var}(\hat{\beta}_A) + \pi_b \text{ cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 1}) + \pi_c \text{ cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 2}) + \pi_d \text{ cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 3}) + \pi_b \text{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 1}) + \pi_{\text{b}}^2 \text{var}(\hat{\beta}_{\Delta 1}) +$$
$$\pi_c \text{ cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 2}) + \pi_{\text{c}}^2 \text{var}(\hat{\beta}_{\Delta 2}) + \pi_{\text{d}}^2 \text{var}(\hat{\beta}_{\Delta 3}) + {\beta_{\Delta 1}}^2 \text{var}(\hat{\pi}_b) + \pi_d \text{ cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 3}) +$$
$$2\beta_{\Delta 1}\beta_{\Delta 2} \text{ cov}(\hat{\pi}_b, \hat{\pi}_c) + 2 \beta_{\Delta 1}\beta_{\Delta 3} \text{ cov}(\hat{\pi}_b, \hat{\pi}_d) + {\beta_{\Delta 2}}^2 \text{var}(\hat{\pi}_c) +$$
$$+ 2\beta_{\Delta 2}\beta_{\Delta 3} \text{ cov}(\hat{\pi}_c, \hat{\pi}_d) + {\beta_{\Delta 3}}^2 \text{var}(\hat{\pi}_d)$$

$$\text{var}(\hat{\beta}) = \text{var}(\hat{\beta}_A + \pi_b \hat{\beta}_{\Delta 1} + \pi_c \hat{\beta}_{\Delta 2} + \pi_d \hat{\beta}_{\Delta 3})$$
$$= \text{var}(\hat{\beta}_A) + \text{var}(\pi_{\text{b}} \hat{\beta}_{\Delta 1}) + \text{var}(\pi_c \hat{\beta}_{\Delta 2}) + \text{var}(\pi_d \hat{\beta}_{\Delta 3}) + 2\text{cov}(\hat{\beta}_A, \pi_b \hat{\beta}_{\Delta 1}) + 2\text{cov}(\hat{\beta}_A, \pi_c \hat{\beta}_{\Delta 2})$$
$$+ 2\text{cov}(\hat{\beta}_A, \pi_d \hat{\beta}_{\Delta 3}) + 2\text{cov}(\pi_b \hat{\beta}_{\Delta 1}, \pi_c \hat{\beta}_{\Delta 2}) + 2\text{cov}(\pi_b \hat{\beta}_{\Delta 1}, \pi_d \hat{\beta}_{\Delta 3})$$
$$+ 2\text{cov}(\pi_c \hat{\beta}_{\Delta 2}, \pi_d \hat{\beta}_{\Delta 3})$$

Since the covariance between $\hat{\beta}_{\Delta 1}$, $\hat{\beta}_{\Delta 2}$ and $\hat{\beta}_{\Delta 3}$ is *0*:

$$= \text{var}(\hat{\beta}_A) + \pi_{\text{b}}^2 \text{var}(\hat{\beta}_{\Delta 1}) + \pi_{\text{c}}^2 \text{var}(\hat{\beta}_{\Delta 2}) + \pi_{\text{d}}^2 \text{var}(\hat{\beta}_{\Delta 3}) + 2\pi_b \text{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 1})$$
$$+ 2\pi_c \text{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 2}) + 2\pi_d \text{cov}(\hat{\beta}_A, \hat{\beta}_{\Delta 3})$$

Combining $\text{var}(\hat{\beta})$ and $\text{V}(\widehat{\boldsymbol{\theta}})$ gives the following:

$$\left(\frac{\partial g(\boldsymbol{\theta})}{\partial \theta}\right) V(\widehat{\boldsymbol{\theta}}) \left(\frac{\partial g(\boldsymbol{\theta})}{\partial \theta}\right)' = \text{var}(\hat{\beta}) + {\beta_{\Delta 1}}^2 \text{var}(\hat{\pi}_b) + 2\beta_{\Delta 1}\beta_{\Delta 2}\text{cov}(\hat{\pi}_b, \hat{\pi}_c) + 2\beta_{\Delta 1}\beta_{\Delta 3}\text{cov}(\hat{\pi}_b, \hat{\pi}_d) +$$
$${\beta_{\Delta 2}}^2 \text{var}(\hat{\pi}_c) + 2\beta_{\Delta 2}\beta_{\Delta 3}\text{cov}(\hat{\pi}_c, \hat{\pi}_d) + {\beta_{\Delta 3}}^2 \text{var}(\hat{\pi}_d)$$
$$= \text{var}(\hat{\beta}) + {\beta_{\Delta 1}}^2 \frac{\pi_{\text{b}}(1 - \pi_{\text{b}})}{\text{N}} - 2\beta_{\Delta 1}\beta_{\Delta 2} \frac{\pi_{\text{b}}\pi_{\text{c}}}{\text{N}} - 2\beta_{\Delta 1}\beta_{\Delta 3} \frac{\pi_{\text{b}}\pi_{\text{d}}}{\text{N}} + {\beta_{\Delta 2}}^2 \frac{\pi_{\text{c}}(1 - \pi_{\text{c}})}{\text{N}}$$
$$- 2\beta_{\Delta 2}\beta_{\Delta 3} \frac{\pi_{\text{c}}\pi_{\text{d}}}{\text{N}} + {\beta_{\Delta 3}}^2 \frac{\pi_{\text{d}}(1 - \pi_{\text{d}})}{\text{N}}$$

This means that the variance can be calculated in standard statistical software by calculating the variance for $\hat{\beta}$ using the estimate statement and then adding the contribution to the variance made by the other terms; $\hat{\beta}_{\Delta 1}$ , $\hat{\beta}_{\Delta 2}$ and $\hat{\beta}_{\Delta 3}$ are available as standard output and $\pi_{\text{b}}$, $\pi_{\text{c}}$ and $\pi_{\text{d}}$ can be calculated easily from the sample.

**Appendix 2**
**Definition of variables used in SAS code**

| Variable | Description | Coding |
|---|---|---|
| Pid | Unique participant identifier | |
| Point<br>Or t<br>Or pointx | Time point in study | 1 = baseline<br>2 = 6 months<br>3 = 12 months<br>4 = 18 months<br>5 = 24 months |
| Pointsquared | $Point^2$ (quadratic time effect) | |
| Treatment | Arm randomised to | 1 = early integrated treatment arm<br>2 = late integrated treatment arm<br>3 = sequential treatment arm |
| Dummyearly | Dummy variable indicating belonging to early treatment arm | 0 = Does not belong to early treatment arm<br>1 = Belongs to early treatment arm |
| Dummylate | Dummy variable indicating belonging to late treatment arm | 0 = Does not belong to late treatment arm<br>1 = Belongs to late treatment arm |
| Dropout | Variable indicating dropout category | Dropout is coded differently in different models. In the case of a binary dropout variable, it would be<br>0 = Completed the study<br>1 = Dropped out of the study |
| Cd4count | CD4+ count | |
| CD4Square | Square root of CD4+ count | |
| Cd4at6 | CD4+ count at 6 months | |
| Cd4at12 | CD4+ count at 12 months | |
| Cd4at18 | CD4+ count at 18 months | |
| Cd4at24 | CD4+ count at 24 months | |
| Sqr_cd4atx | Square root of CD4+ count at x months | |
| Logvl | Log10 of viral load at baseline | |
| Who | WHO status at baseline | 0 = WHO stage 1,2 or 3<br>1 = WHO stage 4 |
| Baselinecd4 | CD4+ count at baseline | CD4+ count divided by 50 |
| Age | Age at baseline in years | |
| Gender | Gender | 0 = male<br>1 = female |
| Historytb | History of tuberculosis at baseline | 0 = no history of tuberculosis<br>1 = history of tuberculosis |
| Extrapul | Presence of extra-pulmonary tuberculosis | 1 = extra pulmonary tuberculosis present<br>0 = extra pulmonary tuberculosis not present |
| MDR | Presence of multidrug resistant tuberculosis | 1 = multidrug resistant tuberculosis detected<br>0 = multidrug resistant tuberculosis not detected |

# Appendix 3
# SAS code

## Code for the identifying restriction from Section 5.5.1.3

```
data growth4;
 set dataset;
 array trt_t[*] trt1t1 trt1t2 trt1t3 trt1t4 trt1t5 trt2t1 trt2t2 trt2t3 trt2t4 trt2t5 trt3t1 trt3t2 trt3t3 trt3t4 trt3t5  ;
 do k=1 to 15;
  trt_t[k]=0;
 end;

    if treatment = 1 and t = 1 then trt1t1  = 1;
    if treatment = 1 and t = 2 then trt1t2  = 1;
    if treatment = 1 and t = 3 then trt1t3  = 1;
    if treatment = 1 and t = 4 then trt1t4  = 1;
    if treatment = 1 and t = 5 then trt1t5  = 1;

    if treatment = 2 and t = 1 then trt2t1  = 1;
    if treatment = 2 and t = 2 then trt2t2  = 1;
    if treatment = 2 and t = 3 then trt2t3  = 1;
    if treatment = 2 and t = 4 then trt2t4  = 1;
    if treatment = 2 and t = 5 then trt2t5  = 1;

    if treatment = 3 and t = 1 then trt3t1 = 1;
    if treatment = 3 and t = 2 then trt3t2  = 1;
    if treatment = 3 and t = 3 then trt3t3  = 1;
    if treatment = 3 and t = 4 then trt3t4  = 1;
    if treatment = 3 and t = 5 then trt3t5  = 1;
run;

step 1: estimate model parameters given monotone missing data, by pattern;
%macro step1;
%do t=1 %to 5;
proc mixed data=growth4 asycov;
 where last_t=&t and t<=&t;   * last_t was created earlier as the last visit where data were observed ;
 class pid t;
 %if &t=1 %then model cd4square = trt1t1 trt2t1 trt3t1/ noint solution covb ddfm=kr;;
 %if &t=2 %then model cd4square = trt1t1 trt1t2 trt2t1 trt2t2 trt3t1 trt3t2/ noint solution covb ddfm=kr;;
 %if &t=3 %then model cd4square = trt1t1 trt1t2 trt1t3 trt2t1 trt2t2 trt2t3 trt3t1 trt3t2 trt3t3/ noint solution
covb ddfm=kr;;
 %if &t=4 %then model cd4square = trt1t1 trt1t2 trt1t3 trt1t4 trt2t1 trt2t2 trt2t3 trt2t4 trt3t1 trt3t2 trt3t3
trt3t4/
      noint solution covb ddfm=kr;;
 %if &t=5 %then model cd4square = trt1t1 trt1t2 trt1t3 trt1t4 trt1t5 trt2t1 trt2t2 trt2t3 trt2t4 trt2t5 trt3t1 trt3t2
trt3t3 trt3t4 trt3t5/
      noint solution covb ddfm=kr;;
 repeated t / subject=pid type=un;

 ods output solutionf=beta_point_&t;  * estimates;
 ods output covb=beta_covar_&t; * precision;
ods output covparms=alpha_point_&t; * estimates...;
 ods output asycov=alpha_covar_&t; * precision;
run;
ods output close;
%end;
%mend;

%step1;
```

```
* step 2: impute data under ACMV;
%macro step2( data=,        /* data set used in 'step 1'                    */
 subject=,      /* subject indicator                            */
 time=,         /* time indicator                               */
 last_time=,    /* indicator of last observed y within each subject */
 y=,            /* target variable                              */
 x=,            /* predictor variable(s)                        */
 nimputation=,  /* number of imputations                        */
 seed=,         /* seed used for multiple imputations           */
 restrictions=  /* identifying restrictions: CCMV, NCMV, or ACMV    */);

proc sql noprint;
 select count(distinct &last_time) into :ntp from &data; %let ntp=%left(&ntp); * number of all patterns or
timepoints;
 select count(distinct &last_time)-1 into :nt from &data; %let nt=%left(&nt); * number of incomplete
patterns;
 select count(distinct &subject) into :n from &data;
 %do t=1 %to &ntp;
 select count(distinct &subject) into :n&t from &data where &last_time=&t;
 %end;
 %do t=1 %to &ntp;
 select count(distinct &subject)/&n into :p&t from &data where &last_time=&t;
 %end;
quit;

* 1. read data;
data _orig;
 set &data;
 _sorting=_n_;
run;

proc iml;

%do t=1 %to &ntp; * loop over all patterns;

use _orig(where=(&last_time=&t)); * only select subjects with incomplete data;
read all var{&subject &last_time &time &y &x
} into data&t;
%end;

* 2. read models: beta and alpha data sets;
%do t=1 %to &ntp; * loop over all patterns;

use beta_point_&t(drop=effect stderr df tvalue probt); * only select the estimates;
read all into beta_point_&t;
use beta_covar_&t(drop=row effect);
read all into beta_covar_&t;
use alpha_point_&t(drop=covparm subject);
read all into alpha_point_&t;
use alpha_covar_&t(drop=row covparm);
read all into alpha_covar_&t;
%end;

* 3. construct the conditional distributions and draw from them;
create _done var{
_imputation &subject &last_time &time &y};

%do imputation=1 %to &nimputation; * start of multiple imputation loop;

 call randseed(&seed+&imputation,1);
```

```
 * sample beta and alpha from their estimated sampling distribution, as is customarily done in multiple
imputation;
 %do j=2 %to &ntp;
  beta_&j=t(randnormal(1,t(beta_point_&j),beta_covar_&j));
  do until (ok); * ensure that each alpha is positive definite;
   temp=t(randnormal(1,t(alpha_point_&j),alpha_covar_&j));
   alpha_&j=j(&j,&j);
   k=0;
   do j=1 to &j;
    do i=1 to j;
     k=k+1;
     alpha_&j[i,j]=temp[k];
     alpha_&j[j,i]=alpha_&j[i,j];
    end;
   end;
   ok=1;
   do r=1 to &j;
    ok=ok & det(alpha_&j[1:r,1:r])>0; * Sylvesters criterion;
   end;
  end;
 %end;

 %do t=1 %to &nt; * loop over incomplete patterns;

  call symput('ni', char(nrow(data&t)));

  %do i=1 %to &ni %by &ntp; * loop through subject * timepoint blocks;

   y=data&t[&i:%eval(&i+&ntp-1),4]; * these are the outcomes for 1 subject;
   x=data&t[&i:%eval(&i+&ntp-1),5:ncol(data&t)]; * these are the covariates for 1 subject;

   %do s=%eval(&t+1) %to &ntp; * loop through missing timepoints s for subject i;

    start=&s;
    stop=&ntp;
    if "&restrictions"="CCMV" then start=&ntp; * for CCMV, start=stop=T, index j=T will be fixed (see
below);
    if "&restrictions"="NCMV" then stop=&s; * for NCMV, start=stop=s, index j=s will be fixed (see
below);;
    call symput('start', char(start));
    call symput('stop', char(stop));

    y1=y[1:%eval(&s-1)]; * y1 only depends on s;
    y2given1=j(%eval(&stop-&start+1),1); * initiate y2given1;
    w=j(nrow(y2given1),1); * initiate w;
    %do j=&start %to &stop; * j>=s -- for CCMV j=T, for NCMV j=s;
     * the choice of beta and alpha depends on j -- for CCMV/NCMV index j is fixed, for ACMV j is running
from s to T;
     beta=beta_&j;
     alpha=alpha_&j;
     * the below only depends on j through beta and alpha;
     mu=x[,1:nrow(beta)]*beta;
     mu1=mu[1:%eval(&s-1)];
     mu2=mu[&s];
     alpha11=alpha[1:%eval(&s-1),1:%eval(&s-1)];
     alpha12=alpha[1:%eval(&s-1),&s];
     alpha21=alpha[&s,1:%eval(&s-1)];
     alpha22=alpha[&s,&s];
     if det(alpha11)>0 then do;
      y2given1[%eval(&j-&start+1),1]=mu2+alpha21*inv(alpha11)*(y1-mu1);
```

```
     w[%eval(&j-&start+1),1]=&&p&j*(det(alpha11)**-0.5)*exp(-0.5*(t(y1-mu1)*inv(alpha11)*(y1-mu1)));
    end;
    else do; * catch exception: if the conditional variance is not positive then give a low weight to y2given1;
     y2given1[%eval(&j-&start+1),1]=mu2;
     w[%eval(&j-&start+1),1]=0.0001;
    end;
   %end;
   if nrow(w)=1 then y[&s]=y2given1[1,1];
   else do;
    abs=w#(1/sum(w)); * ensure that weights sum up to 1;
    cum=j(nrow(abs),1);
    do u=1 to nrow(w);
     cum[u]=sum(abs[1:u]); * ensure that weights become cumulative;
    end;
    ranu=ranuni(&seed+&imputation);
    select=1;
    do u=2 to nrow(w);
     if ranu>cum[u-1] & ranu<=cum[u] then select=u;
    end;
    y[&s]=y2given1[select,1];
   end;

  %end; * end of loop through missing timepoints s for subject i;

  data&t[&i:%eval(&i+&ntp-1),4]=y;

 %end; * end of loop through subject * timepoint blocks;

 impdata=impdata // data&t[,1:4];

%end; * end of loop over incomplete patterns;

alldata=impdata // data&ntp[,1:4]; * add completers;
alldata=j(nrow(alldata),1,&imputation) || alldata; * add imputation index;
setout _done;
append from alldata;
free impdata alldata; * clear these matrices;

%end; * end of imputation loop;
close _done;
quit; * end of proc iml;

* 4. write data;
proc sql;
 create table &data._&restrictions as
 select b._imputation, a.*, b.&y as _&y._&restrictions
 from _orig a, _done b
 where a.&subject=b.&subject and a.&time=b.&time
 order by _imputation, _sorting;
quit;
data &data._&restrictions;
 set &data._&restrictions;
 if &y=. then &y=_&y._&restrictions;
 drop _sorting _&y._&restrictions;
run;

%mend;

%step2(data=growth4, subject=pid, time=t, last_time=last_t,
 y=cd4square, x= trt1t1 trt2t1 trt3t1 trt1t2 trt2t2 trt3t2 trt1t3 trt2t3 trt3t3 trt1t4 trt2t4 trt3t4 trt1t5 trt2t5 trt3t5 ,
```

```
 nimputation=70, seed=999, restrictions=ACMV);

* truncate imputed data;
 data growth4_acmv;
   set growth4_acmv;
      if y gt 35 then y = 35;
              if y lt 0 and y gt . then y = 3;
              t2 = t;
              tsquare = t*t;
run;

* step 3: estimate model parameters given imputed data;

proc mixed data=growth4_acmv asycov;
 by _imputation;
 class pid t treatment;
 model y = treatment t2 treatment*t2 tsquare/ solution covb ddfm=kr;
 repeated t / subject=pid type=un;

 ods output solutionf = mix_acmv covb = mixcovb;
run;

ods output close;

data mixparms2;
  set mix_acmv;
  _imputation_ = _imputation;

  if effect="treatment" and treatment=1 then effect="TRT1";
  if effect="treatment" and treatment=2 then effect="TRT2";
  if effect="t2*treatment" and treatment=1 then effect="TRT1_WEEK";
  if effect="t2*treatment" and treatment=2 then effect="TRT2_WEEK";
  if effect="treatment" and treatment=3 then effect="TRT3";
run;

  proc mianalyze parms=mixparms2 ;
     modeleffects Intercept t2 TRT1 TRT2 TRT1_WEEK TRT2_WEEK tsquare;
   run;
```

The same macro is then used to do imputation under ACMV, NCMV and CCMV. The example is for ACMV.

## Code for pattern-mixture model using multiple imputation in Section 5.5.1.2

```
/*  Macro name: Modify
Parameters
Data= Name of original data set
Imp= Name of imputed data set
Out= Name of Output data set
Var= List of variables as used in Var statement for MI
Deltadrop and deltadead= The amount to decrease imputed value by
S= SD for Normal distribution from which Change is
sampled with mean Delta for each imputation
Trx= Name of variable holding treatment classification
n = number of variables used to impute*/

%macro modify(data= ,imp= ,out= , var =, deltadrop=, deltadead= ,s= ,trx=, n = );
%local i n;

* Get number of elements in the Var List as macro variable n;
%let i=1;
%let txt=%scan(&var, &i, %str( ));
%do %while(%length(&txt)) ;
%let i=%eval(&i +1);
%let txt=%scan(&var, &i, %str( ));
%end;
%let n=%eval(&i -1);

data Temp1;
set &data;
array My_Var[1:&n] &Var;
array My_Ind[1:%eval(&n+1)] My_Ind1-My_Ind&n No;
keep My_row My_ind1-My_Ind&n;
My_Row=_N_;

No=0;

do i= &n to 1 by -1;
My_ind[i]= ( ((My_Var[i] > .z) + (My_ind[i+1])) > 0 );
end;
run;

data Temp2;
set &imp;
by _imputation_;
retain My_Row 0;
if first._imputation_ then My_Row=0;
My_Row=My_Row+1;
run;

proc sql;
create table Temp3 as
select A.*, B.*
from Temp2 A left join Temp1 B
on A.My_Row = B.My_Row
order by _imputation_, &Trx;
quit;

* My_Ind=1 if data is Real;
* My_Ind=0 if data is Imputed;

data &out;
```

```
set Temp3;
by _imputation_ &trx;

array My_Var[1:&n] &Var;
array My_Ind[1:&n] My_Ind1-My_Ind&n;
drop My_Row Change i My_ind1-My_Ind&n;
retain deltadrop deltadead;

* Change allows us to build up delta within the subject;

if first.&trx then do;
seednum = _imputation_*&trx;
Deltadead=&deltadead+&s*rannor(seednum);
Deltadrop=&deltadrop+&s*rannor(seednum);
end;
Change=0;
do i=1 to &n;

* If it is imputed then increase Change by delta;
if My_ind[i]=0 then do;
   if dropoutdead = 1 then change = change + deltadrop;
          if dropoutdead = 0 then change = change + deltadead;
end;

My_Var[i]=My_Var[i]-(i)*Change;
end;
run;
%mend modify;

/*  Multiple imputation  */
proc mi data = dataset out = full nimpute = 50 seed = 1 round = 1 1 1 1 0.001 . maximum = 1000 1000 1000
1000 8
  minimum = 1 1 1 1 0;
  em initial = cc;
  MCMC nbiter=5000 niter=5000;
  mcmc impute = full chain = single timeplot (mean(sqr_cd4at6 sqr_cd4at12 sqr_cd4at18 sqr_cd4at24))
       acfplot (mean (sqr_cd4at6 sqr_cd4at12 sqr_cd4at18 sqr_cd4at24));
  var  sqr_cd4at6 sqr_cd4at12 sqr_cd4at18 sqr_cd4at24 logvl who baselinecd4 age gender historytb extrapul
mdr ;
  by treatment;
run;

proc sort data = full; by _imputation_;

%macro analiseer;

data full2;
  set bbb;
    cd4at6kwa = sqr_cd4at6*sqr_cd4at6;
    cd4at12kwa = sqr_cd4at12*sqr_cd4at12;
    cd4at18kwa = sqr_cd4at18*sqr_cd4at18;
    cd4at24kwa = sqr_cd4at24*sqr_cd4at24;
run;

proc sort; by _imputation_ treatment; run;

/*  Estimate treatment effect at final time point using each imputed data set   */

proc univariate data=full2 NOPrint;
   var cd4at6kwa cd4at12kwa cd4at18kwa cd4at24kwa;
```

```
  output out=outuni mean= cd4at6 cd4at12 cd4at18 cd4at24 stderr= Scd4at6 Scd4at12 Scd4at18 Scd4at24
     n = ncd4at6 ncd4at12 ncd4at18 ncd4at24;
  by _Imputation_ treatment;
run;

PROC SORT; by treatment;

proc glm data=full2 ;
  model cd4at6kwa= treatment/inverse;
  by _Imputation_;
  ods output ParameterEstimates=glmparmssix
   InvXPX=glmxpxisix;
quit;

*  etc for each time point ;

/*  Some data manipulation steps to create dataset in correct format to use with proc MIXED  */

proc mixed data = fulla method = reml;
 class pid treatment point ;
 model col1 = point2 treatment point2*treatment pointsquared/ solution covb ddfm = kr ;
 repeated point / subject = pid type = un group = treatment;
 by _imputation_;
 ods output solutionf = mixparms covb = mixcovb;
run;
ods output close;

data mixparms2; set mixparms;
if effect="treatment" and treatment=1 then effect="TRT1";
if effect="treatment" and treatment=2 then effect="TRT2";
if effect="point2*treatment" and treatment=1 then effect="TRT1_WEEK";
if effect="point2*treatment" and treatment=2 then effect="TRT2_WEEK"; run;
run;

  data mixcovb2; set mixcovb;
if effect="treatment" and treatment=1 then effect="TRT1";
if effect="treatment" and treatment=2 then effect="TRT2";
if effect="point2*treatment" and treatment=1 then effect="TRT1_WEEK";
if effect="point2*treatment" and treatment=2 then effect="TRT2_WEEK";
rename col3=TRT1;
rename col4=TRT2;
rename col6=TRT1_WEEK;
rename col7=TRT2_WEEK;
rename col1= intercept;
rename col2 = point2;
rename col9 = POINTSQUARED;
run;

proc mianalyze parms=mixparms2 covb=mixcovb2;
modeleffects intercept point2 TRT1 TRT2 TRT1_WEEK TRT2_WEEK POINTSQUARED ;
test1: test trt1+TRT1_WEEK= trt2+ TRT2_WEEK/mult; run;

proc mianalyze data=outuni ;
  modeleffects cd4at6 cd4at12 cd4at18 cd4at24;
  stderr Scd4at6 Scd4at12 Scd4at18 Scd4at24;
  BY treatment;
run;
%mend;

%macro pvals;
```

```
proc mianalyze parms=glmparmssix xpxi=glmxpxisix;
modeleffects Intercept treatment;
run;

*  etc and repeat for all time points ;
%mend;

title1 'dropout = 0.5 and dead = 1';
%modify(data=dataset, imp=full, out=bbb, var= sqr_cd4at6 sqr_cd4at12 sqr_cd4at18 sqr_cd4at24,
deltadrop= 0.5, deltadead=1, trx=treatment, s=1);

%analiseer;

%pvals;

* These steps are then repeated for each of the combinations of delta and sigma;
```

**Appendix 4**

**Grobler, A.C, Matthews, G, Molenberghs, G. The impact of missing data on clinical trials: a re-analysis of a placebo controlled trial of Hypericum perforatum (St Johns wort) and sertraline in major depressive disorder. Psychopharmacology, 2013, DOI 10.1007/s00213-013-3344-x**

# The impact of missing data on clinical trials: a re-analysis of a placebo controlled trial of *Hypericum perforatum* (St Johns wort) and sertraline in major depressive disorder

**Anneke C. Grobler · Glenda Matthews · Geert Molenberghs**

Received: 23 May 2013 / Accepted: 18 October 2013


**Abstract**

*Rationale and objective* *Hypericum perforatum* (St John's wort) is used to treat depression, but the effectiveness has not been established. Recent guidelines described the analysis of clinical trials with missing data, inspiring the reanalysis of this trial using proper missing data methods. The objective was to determine whether hypericum was superior to placebo in treating major depression.

*Methods* A placebo-controlled, randomized clinical trial was conducted for 8 weeks to determine the effectiveness of hypericum or sertraline in reducing depression, measured using the Hamilton depression scale. We performed sensitivity analyses under different assumptions about the missing data process.

*Results* Three hundred forty participants were randomized, with 28 % lost to follow-up. The missing data mechanism was not missing completely at random. Under missing at random assumptions, some sensitivity analyses found no difference between either treatment arm and placebo, while some sensitivity analyses found a significant difference from baseline to week 8 between sertraline and placebo (−1.28, 95 % credible interval [−2.48; −0.08]), but not between hypericum and placebo (0.56, [−0.64;1.76]). The results were similar when the missing data process was assumed to be missing not at random.

*Conclusions* There is no difference between hypericum and placebo, regardless of the assumption about the missing data process. There is a significant difference between sertraline and placebo with some statistical methods used. It is important to conduct an analysis that takes account of missing data using valid statistically principled methods. The assumptions about the missing data process could influence the results.

**Keywords** St John's wort · *Hypericum perforatum* · Herbal medicine · Antidepressant · Sertraline · Hamilton depression scale · Bayesian · Multiple imputation · Missing at random · Missing not at random

Trial Registration: Clintrials.gov, NCT00005013, http://www.clinicaltrials.gov/ct2/show/NCT00005013?term=Hypericum+perforatum+major+depression

A. C. Grobler (✉)
Centre for the AIDS Programme of Research in South Africa (CAPRISA), Nelson R Mandela School of Medicine, University of KwaZulu-Natal, Private Bag X7, Durban 4013, South Africa
e-mail: grobler@ukzn.ac.za

G. Matthews
School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban, South Africa

G. Molenberghs
I-BioStat, Universiteit Hasselt and KU Leuven, Leuven, Belgium

## Introduction

### *Hypericum perforatum*

*H. perforatum* (St John's wort), is a herbal remedy used in the treatment of depression, especially in European countries (Fegert et al. 2006). It was shown to be more effective than placebo (Kalb et al. 2001) in treating depression and is believed to have fewer side effects than standard antidepressive therapies (Kasper et al. 2010; Linde et al. 1996). Some studies and meta-analyses have found hypericum to be as effective as standard antidepressive therapies (Linde et al. 2008; Rahimi et al. 2009), while other studies found no difference between hypericum and placebo (Shelton et al. 2001). Because different studies had contradictory results about the effectiveness of hypericum compared to placebo and standard antidepressive drugs, a trial was designed to compare both a standard antidepressive therapy (sertraline) and hypericum to placebo (Hypericum Depression Trial Study Group 2002). Sertraline

(trademark Zoloft) is an antidepressant of the selective serotonin reuptake inhibitor (SSRI) class. The practicing clinician will be interested in these results, since herbal preparations should be used only when evidence exist that it is efficacious.

### Missing data

Missing data are common in longitudinal clinical trials. Over the last 30 years, methods were designed for the proper analysis of clinical trial data when missing data are present. This is summarized in two guidance documents by the European Medicines agency, "Guideline on Missing Data in Confirmatory Clinical Trials"(CHMP 2010), and the National Research Council in the United States of America, "The Prevention and Treatment of Missing Data in Clinical Trials"(National Research Council 2010). The second report was recently summarized in the New England Journal of Medicine (NEJM) (Little et al. 2012). In the same issue, the journal announced new review policies regarding missing data (Ware et al. 2012). Reviewers will in future look at aspects of trial design that reduce the impact of missing data. Weighting or model-based methods will be preferred over complete case analysis or single imputation methods. The NEJM will in future also require sensitivity analysis when missing data are extensive.

Rubin (1976) described three different missing data mechanisms based on the level of dependence between the missing data process and the measurement process. These are missing completely at random, missing at random, and missing not at random. Data are missing completely at random when the probability of dropout is independent of both observed and unobserved data, for example, when a sample was lost in the laboratory or a patient did not attend a visit due to transport problems. Data are missing at random when the reason for dropout is known and associated with trial-related events (Carpenter et al. 2002). The reason for dropout can depend on observed data, but not on unobserved data, for example, when a participant who is doing poorly is subsequently discontinued from the trial by the clinician or as per participant's choice and a poor efficacy outcome is recorded in the study database. When neither missing completely at random nor missing at random is valid, data are missing not at random. In this instance, the missingness can be explained by unobserved outcomes, for example, when a participant whose condition worsens stops coming to the clinic, and this worsened condition is not observed. The missing data mechanism cannot be determined using the data observed, except possibly to confirm that the missing data mechanism is not missing completely at random.

Ten years ago, it was standard practice to analyze clinical trials using complete case analysis or single imputation techniques including last observation carried forward (LOCF). This is changing in favor of more appropriate methods and recommendations now advise against these (Carpenter et al. 2004; Carpenter et al. 2013; CHMP 2010; Mallinckrodt et al. 2001; Molenberghs and Kenward 2007; Molenberghs et al. 2004; National Research Council 2010; Ratitch et al. 2013). In 2002, the hypericum clinical trial was published using complete case analysis and last observation carried forward (Hypericum Depression Trial Study Group 2002). In this paper, we reanalyzed the data up to week 8, using principled methods suggested by current guidelines, which did not exist in 2002. We use this trial as an example of the changes in analyses that are required if the new guidelines are to be adopted.

## Materials and methods

### Trial design

The trial was a randomized, double-blind, parallel-arm, 8-week outpatient trial of hypericum, sertraline, or placebo treatment for major depressive disorder, followed by 18 weeks of double-blind continuation treatment in participants meeting response criteria at week 8. The study consisted of an acute phase (the first 8 weeks) and an optional continuation phase (from weeks 8 to 26). The focus of this paper is on the acute phase only. The specific inclusion and exclusion criteria are given in the trial publication (Hypericum Depression Trial Study Group 2002). The eligibility of participants was assessed, after which they gave written informed consent and participated in a 1-week placebo run-in. Participants meeting eligibility criteria after the run-in were randomized to one of the three treatment arms in a 1:1:1 ratio. Participants were assessed weekly from week 1 to week 8. The Hamilton depression scale (HAM-D) (Hamilton 1960), global assessment of functioning (GAF) scale, Clinical Global impressions scale for severity (CGI-S) and improvement (CGI-I), and the Beck Depression Inventory (BDI) were assessed at all visits. Other safety-related information was also collected, such as vital signs, adverse events, and blood chemistry and hematology (Hypericum Depression Trial Study Group 2002). We analyze the response over time on HAM-D, which is a measure of depression, with a higher score indicative of more severe depression.

The primary hypothesis was whether hypericum is superior to placebo after 8 weeks. The endpoint was defined as the change in the HAM-D score from baseline to week 8. The principal comparison was between the hypericum and placebo arms. The sertraline arm was included as an active control arm to validate the study, but no comparison between hypericum and sertraline was intended and the trial was not powered for such a comparison or for multiple comparisons with placebo (Hypericum Depression Trial Study Group 2002). Details about the proportion of patients discontinued, timing of discontinuation, and reason for discontinuation are provided.

## Statistical methods

In the original analysis, treatment differences in the change in HAM-D total score from baseline to week 8 were evaluated through a random coefficient regression model. Available longitudinal scores through week 8 were modelled as a linear function of fixed effects for treatment, site, sex, week, and treatment by week, with random intercept and slope for each patient. A secondary analysis was restricted to participants who completed the acute phase (completer analysis), and analysis of covariance models used last observation carried forward; although these results were not given in the paper (Hypericum Depression Trial Study Group 2002). An analysis of the data in the continuation phase was also done in a separate paper. In this analysis, the HAM-D scores for completers at the final time point were compared and last observation carried forward was applied (Sarris et al. 2012). While the random coefficient regression model could be appropriate if the missing data mechanism is missing completely at random or missing at random, last observation carried forward is not an appropriate way of handling missing data, because it might inflate the Type I error and create bias in the estimation of mean change from baseline while producing standard errors that are too small (Mallinckrodt et al. 2001). The random coefficient regression model takes into account the expectation of the missing measurements, given the observed measurements and is valid and unbiased under missing at random (Molenberghs and Kenward 2007). It can be thought of as aiming to estimate the treatment effect that would have been observed if all participants had continued on treatment for the full study duration (CHMP 2010).

We reanalyzed the data using principled missing data techniques. Assumptions were made about the missing data mechanism and a series of sensitivity analyses were done under these assumptions. If the missing data mechanism is missing completely at random, an available case or complete case analysis would be valid, although this analysis would have reduced power because of the exclusion of some participants. If one assumes the missing data mechanism is missing at random, several methods would be valid, including maximum likelihood models, multiple imputation, and Bayesian analysis.

Likelihood-based approaches use a parametric model to formulate a statistical model for the missing data and base inferences on the likelihood function of the incomplete data. The objective is to draw inference about a parameter, $\theta$, in a model $f(y|\theta)$ for the response data that is not fully observed. Under the missing at random assumption, $\theta$ and the missing data model are functionally independent and missing data can be treated as ignorable. In this case, inference is drawn about $\theta$ without having to specify a model that relates the missing data process to the observed data (National Research Council 2010).

In the absence of missing data, likelihood-based methods entail the maximization of the full data likelihood. With incompleteness, this likelihood is replaced by the observed data likelihood, where the individual likelihoods are integrated over the missing values, $\prod_{i=1}^{N} \int f(y_i^{obs}, y_i^{miss}, r_i|\theta,\psi)\mathrm{d}y_i^{miss}$; where $y$ indicate the outcome variable, $r$ is the missing data indicator, $\theta$ and $\psi$ are parameter vectors describing the measurement and missingness processes, respectively, and $N$ is the number of participants. Under ignorability and missing completely at random or missing at random missingness, the integral can be rewritten as an integral over the missing values and the distribution of the missing data mechanism (under a selection model). Under missing completely at random, this becomes $\prod_{i=1}^{N} f(y_i^{obs}|\theta)f(r_i|\psi)$ and under missing at random this becomes $\prod_{i=1}^{N} f(y_i^{obs}|\theta)f(r_i|y_i^{obs},\psi)$. We fitted a model with fixed and random effects, including treatment, week, and the interaction between treatment and week, adjusted for repeated measures.

Single imputation has several limitations. When analyzing observed data, it is assumed that measurements are made with error. To assume that if data are missing, we can impute the missing value without error (a single value) is unrealistic. With conditional mean imputation, the imputed data are much less variable than the observed data would have been. Thus, analyzing imputed data as observed data leads to an underestimation of standard errors, $p$ values, and confidence intervals (Carpenter and Kenward 2007; CHMP 2010).

Multiple imputation, first suggested by Rubin (1976), overcomes the limitations of single imputation. Multiple imputation is done in three steps. In step 1, plausible values for missing observations are imputed that reflect uncertainty about the missing data models, generally assuming the missing data process is missing at random. These values are used to fill in or impute the missing values. This process is repeated, resulting in the creation of several complete data sets, taking into account the uncertainty in estimating both the relationship between the variables and the residual variability. These provide a representation of the distribution of the missing data given the observed. In step 2, each of these data sets is analyzed using complete data methods that would have been appropriate had there been no missing data. In Step 3 the results are combined, taking the uncertainty regarding the imputations into account (Rubin 1976). This additional step is needed to correctly estimate the variability of quantities estimated from a completed data set. These results are unbiased and have approximately the correct standard error (Horton and Kleinman 2007; Molenberghs and Kenward 2007).

Multiple imputation is said to be proper if it leads to consistent asymptotically normal estimators, correct variance estimators, and valid tests. Generally, the imputation will be

proper if all sources of variability and uncertainty are included in the imputation model, including prediction errors of the individual values and errors of estimation in the fitted coefficients of the imputation model (White et al. 2011). Multiple imputation is done using Bayesian predictive distribution and Monte Carlo Markov Chain sampling to generate the imputation, assuming that the data follows a multivariate normal distribution. The model used for imputation should include all the variables included in the analysis model (to ensure proper imputation), and all variables that could improve the prediction.

Multiple imputation with 100 imputations was used to analyze the change from baseline to week 8. The imputation model included all observed values of HAM-D, age, sex, race, duration of depression, BDI, CGI-S, CGI-I, and GAF scale at baseline. The variables were selected because they were believed to be factors that could predict HAM-D scores or were included in the analysis model. The imputed datasets were used to get the estimates of the change from baseline to week 8, and the appropriate $p$ values and summary statistics were calculated using Rubin's rule (Rubin 1976). The imputed datasets were also analyzed with a mixed model as described previously.

At its core, multiple imputation uses Bayesian techniques, since the imputations are sampled from a Bayesian posterior distribution. Fully Bayesian approaches, where multiple datasets are not imputed, are appropriate for the analysis of missing data by specifying priors on all the parameters and specifying distributions for the missing covariates (Daniels and Hogan 2008; Horton and Kleinman 2007). The missing data are then sampled from their conditional distribution via the Gibbs sampler, an algorithm that samples a Markov chain where the kernel is the product of the sequentially updated full conditional distributions of the parameters and the stationary distribution is the posterior distribution (Geman and Geman 1984).

In Bayesian analyses, parameters are treated as random variables. Probability statements are made about the model parameters and not about the data. Bayesian analysis has three components. The prior distribution, $p(\theta)$, reflects the distribution of the parameters before the data are seen. The likelihood, $L(\theta|D)$, gives the distribution of the observed data. The posterior distribution uses Bayes' theorem to combine information from the prior distribution and the likelihood and expresses uncertainty about the unknown parameters after seeing the data. In ignorable methods, the posterior distribution is $p(\theta|D) \propto p(\theta)L(\theta|D)$. The Bayesian inference is done by specifying a model, specifying prior distributions for the parameters of the model, and then updating the prior information on the parameters using the model specified and the data observed to obtain the posterior distribution of the parameters. In addition to specifying a missingness model, assumptions about the missing data and the uncertainty around the missing data can be made explicit through the prior distributions (Daniels and Hogan 2008). The priors can be used to encode information about the missing data process. Bayesian methods are a natural way of handling missing data because a probability distribution is estimated for each missing value, allowing for uncertainty to be captured. Missing data are treated as additional unknown quantities, thus, no distinction is made between missing data and unknown parameters. After specifying an appropriate joint model for the observed and missing data and the model parameters, posterior samples of the model parameters and missing values will be generated using Markov Chain Monte Carlo methods (Mason et al. 2012).

Letting $Y_{ij}$ be the HAM-D score for participant $i$ at occasion $j$, where $Y_{ij} \sim \text{Normal}(\mu_{ij}, \sigma^2)$, the following Bayesian model was fitted:

$$\mu_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 S_i + \beta_3 H_i + (\beta_4 S_i + \beta_5 H_i) t_{ij}$$

where, $t_{ij}$ indicates occasion $j$ (week$_j$) for participant $i$. $S_i$ is an indicator variable for the sertraline arm and equals 1 if participant $i$ belongs to the sertraline arm and is 0 otherwise. Similarly, $H_i$ is an indicator variable for the hypericum arm and equals 1 if participant $i$ belongs to the hypericum arm and is 0 otherwise. A participant $i$ belonging to the placebo arm will thus have $H_i = S_i = 0$. The placebo arm is therefore the reference group. We assume $N$ independent participants.

Vague priors were specified for the unknown parameters and are given in Table 4. Sensitivity analyses were performed with various prior distributions, to assess the sensitivity of the Bayesian models to the choice of prior distribution. Priors could also have been used to include data about the missingness process. One could potentially use different prior distributions for each of the treatment arms, if the prior beliefs about the missingness mechanism warranted this.

Under missing not at random, an assumption we cannot test using the observed data, Bayesian analyses, pattern mixture models, selection models, and shared parameter models can be used, at the expense of increased complexity.

Under a missing not at random assumption, we fitted three Bayesian models under several different assumptions about the missing data mechanism. Bayesian analysis provides a flexible way to model missing not at random data, using a selection model factorization of a joint model, consisting of a model of interest and a model of missingness. The same model for the observed data was fitted as under missing at random and a model of missingness of the form $m_{ij} \sim \text{Bernoulli}(p_{ij})$, $\text{logit}(p_{ij}) = \theta_0 + \Delta y_{ij}$ was added, where $i = 1, \ldots, 340$ indicates the participant and $j = 1, \ldots 8$ indicates the visit, $m_{ij}$ is a binary missing value indicator for $y_{ij}$. This model allows the missingness to depend on the value that would have been observed.

A second missing not at random model was fitted where the model of missingness had the following form: logit $p_{i,w}=\theta_0+\theta_1 y_{i,w-1}+\theta_2(y_{i,w}-y_{i,w-1})$. This model allows the missingness to depend on the value that would have been observed at the current occasion, where the value is possibly missing, as well as on the previous observed value. In so doing, in line with what is oftentimes done in a selection model specification, missingness can be seen to depend on both the level of the outcome (represented by the average of the current and previous values) and the increment between the previous and current values. It should be noted that the data do not carry information on $\theta_2$ in the usual sense. While under a likelihood and Bayesian paradigm parameters may be identified, the usual asymptotics may not hold, in the sense that information accrual may be minimal with increasing sample sizes. This subtle issue is discussed and illustrated with data analyses and simulations in Jansen et al. (2006) and the references therein. Practically, it means that parameters distinguishing missing not at random from missing at random under the posited model may be identifiable from the observed data, but only barely so.

A third missing not at random model was fitted where the model of missingness has the following form: logit $p_{ij}=\theta_0+\theta_1 y_{ij}+\theta_2 y_{ij}S_i+\theta_3 y_{ij}H_i$. This model allows the missingness to depend on the unobserved HAM-D value, while allowing for a different mechanism in each treatment arm by including the HAM-D score by treatment interaction. The priors are given in Table 4. The priors were chosen to be flat and therefore uninformative. The parameters were varied from extremely flat priors, to less flat priors in order to investigate whether the models were sensitive to the choice of prior. Many other models for nonrandom missingness could be fitted, depending on the assumptions made regarding the missing data mechanism.

The assumptions made by the models fitted were tested. The assumption of linear regression was tested by plotting the studentized residuals against the predicted means. This plot showed no deviations from the assumption of linear regression. The normality assumption was tested by looking at the normal probability plot of the residuals. No deviation from normality was present. Different variance covariance structures were fitted and the most appropriate was chosen using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

All analyses were performed on the data collected from baseline to week 8, using SAS version 9.3 software (SAS Institute Inc, Cary, NC), with the exception of the Bayesian analyses, which were performed using OPENBUGS version 3.2.2. All assumptions of linear regression were satisfied.

## Results

The trial enrolled 428 participants in the run-in phase and 340 were randomized to the tree treatment arms. The demographics of the sample and the CONSORT diagram is described elsewhere (Hypericum Depression Trial Study Group 2002). The mean pooled baseline HAM-D score was 22.8 (SD=2.7). A large number of participants was lost to follow-up before week 8; 28.8 % in the sertraline arm, 27.4 % in the hypericum arm, and 27.6 % in the placebo arm. At the end of week 8, the percentage participants who dropped out were similar across the three arms; however, participants discontinued sooner in the active arms, especially in the sertraline arm. The extent of and reason for missing data at each visit is given in Table 1.

A central question is whether the missing data are missing completely at random, missing at random, or missing not at random. Although the data cannot exclude missing not at random, it can hold some evidence of informative missingness. Figures 1 and 2 compare the HAM-D score of participants who dropped out at the next visit to those of participants who did not drop out at the next visit. Participants who dropped out are different from those who did not drop out. Prior to week 4, dropouts had lower HAM-D scores than those who did not drop out. From week 4 onwards, dropouts had higher HAM-D scores than participants who did not drop out; thus, dropout depends at least on observed HAM-D score. Among demographic and baseline variables, duration of depression and age were significantly associated with dropout in a logistic regression.

The fact that some withdrawals were due to insufficient response suggested that HAM-D score should not be analyzed without taking missing data into account. In addition, the fact that participants started dropping out sooner in the sertraline arm might reflect that tolerability issues were related to dropout, casting further doubt on the suitability of a missing at random model. Ignoring participants who drop out due to insufficient response or tolerability issues will introduce an important bias in the complete case analysis. In this context, the missing at random based model expresses the assumption that a participant's observed history is deemed adequate to derive his or her probability of dropping out. Here, history is to be understood as the combination of the patient's outcomes from the beginning of the study up to but not including the current one, and the covariate information, collected at baseline and during follow-up, up to the current time. Up to week 8, a missing completely at random analysis is not appropriate; a missing at random analysis might be appropriate, but missing not at random cannot be ruled out. Assumptions about the missing data, other than missing at random should be considered.

In the original paper, the mean HAM-D scores using available case analysis were given in a figure. According to this figure, the sertraline arm showed the largest improvement over time. It also included a random coefficient regression analysis on the longitudinal HAM-D total scores. This analysis detected a downward linear trend with week ($p$ value$<0.001$). Linear trends with week did not

**Table 1** Number of participants attending each visit

| | Sertraline $N=111$ | | Hypericum $N=113$ | | Placebo $N=116$ | |
|---|---|---|---|---|---|---|
| Number with Hamilton depression score at | $N$ | Number missing (%) | $N$ | Number missing (%) | $N$ | Number missing (%) |
| Baseline | 111 | – | 113 | – | 116 | – |
| Week 1 | 101 | 10 (9.0) | 101 | 12 (10.6) | 111 | 5 (4.3) |
| Week 2 | 90 | 21 (18.9) | 102 | 11 (9.7) | 107 | 9 (7.8) |
| Week 3 | 90 | 21 (18.9) | 100 | 13 (11.5) | 94 | 22 (19.0) |
| Week 4 | 89 | 22 (19.8) | 97 | 16 (14.2) | 99 | 17 (14.7) |
| Week 6 | 82 | 29 (25.1) | 91 | 22 (19.5) | 93 | 23 (19.8) |
| Week 7 | 79 | 32 (28.8) | 82 | 31 (27.4) | 84 | 32 (27.6) |
| Week 8 | 79 | 32 (28.8) | 82 | 31 (27.4) | 84 | 32 (27.6) |
| Enter continuation phase | 49 | | 38 | | 42 | |
| Reasons not completing acute phase (week 8) | | $N=32$ | | $N=31$ | | $N=32$ |
| Loss to follow-up | 10 | 31.3 % | 8 | 25.8 % | 7 | 21.9 % |
| Insufficient response | 7 | 21.9 % | 6 | 19.4 % | 11 | 34.4 % |
| Withdrew consent | 8 | 25.0 % | 7 | 22.6 % | 8 | 25.0 % |
| Adverse event | 5 | 15.6 % | 2 | 6.5 % | 3 | 9.4 % |
| Protocol violation | 2 | 6.3 % | 8 | 25.8 % | 3 | 9.4 % |

differ significantly by treatment (hypericum versus placebo, $p$ value=0.59; sertraline versus placebo, $p$ value=0.18). If this analysis was done using all data points while fitting a mixed model using maximum likelihood estimates, this analysis is consistent with missing at random assumptions (Hypericum Depression Trial Study Group 2002). From the original paper, model estimates for the mean change from baseline to week 8 in HAM-D score were calculated for each of the treatment arms (Table 2). The conclusion was that neither hypericum nor sertraline was superior to placebo. The authors highlighted the high level of improvement in the placebo arms often seen in depression trials (Hypericum Depression Trial Study Group 2002).

We compared the change from baseline to week 8 using multiple imputation. The change was slightly larger in all arms using multiple imputation, but the $p$ values were similar to the previous analysis. We conclude that there is no difference between either of the treatment arms and the placebo arm (Table 2).

The results for the change from baseline to week 8 using multiple imputation and a mixed model are presented in Tables 2 and 3, respectively. Multiple imputation and likelihood-based methods make similar assumptions about the missing data, namely that it is missing at random. The conclusions are expected to be similar. Neither analysis found either of the treatments to be different from placebo. Analyzing the data

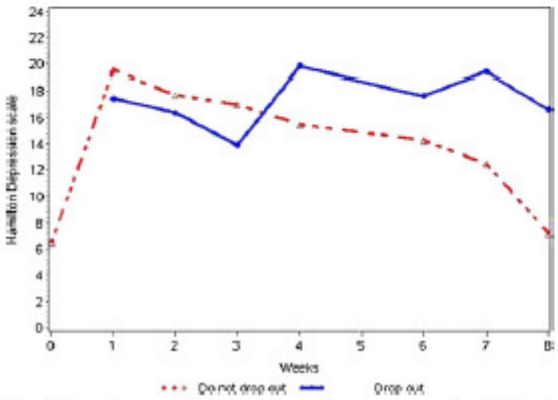



**Fig. 1** Sample mean of Hamilton depression score at each week for all treatment groups combined. The *triangles* are the means for participants who did not drop out before the subsequent measurement. The *circles* are the means for participants who dropped out before the subsequent measurement

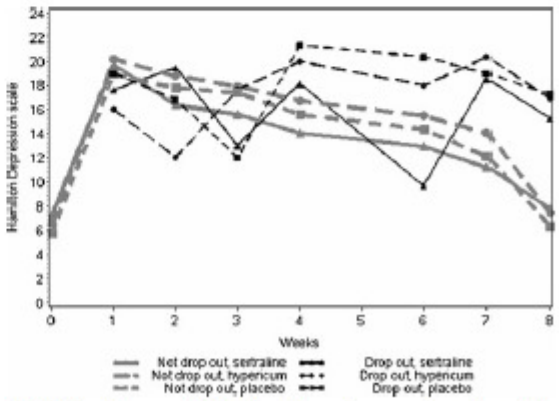



**Fig. 2** Sample mean of Hamilton depression score at each week by treatment group. Not drop out gives the means for participants who did not drop out before the subsequent measurement. Drop out gives the means for participants who dropped out before the subsequent measurement

**Table 2** Change in Hamilton depression score (week 8 to baseline)

| | Mean change from baseline (standard error of the mean change from baseline) [95 % confidence interval] | | | *p* value | |
|---|---|---|---|---|---|
| | Hypericum ($n=113$) | Placebo ($n=116$) | Sertraline ($n=109$) | Hypericum vs placebo | Sertraline vs placebo |
| Last observation carried forward, change from baseline to week 8 | −8.42 (0.68) [−9.76; −7.09] | −8.89 (0.73) [−10.33; −7.45] | −9.68 (0.64) [−10.94; −8.42] | 0.64 | 0.42 |
| From 2002 paper[a] | −8.68 (0.68) [−10.01; −7.35] | −9.20 (0.67) [−10.51; −7.89] | −10.53 (0.72) [−11.94; −9.12] | 0.59 | 0.18 |
| Multiple imputation: change calculated on the imputed values | −9.99 (0.74) [−11.45; −8.54] | −10.01 (0.80) [−11.58; −8.43] | −11.47 (0.67) [−12.78; −10.17] | 0.99 | 0.16 |
| Maximum likelihood models under missing at random assumptions, change based on model results[b] | −9.98 (0.76) [−11.47; −8.49] | −10.12 (0.74) [−11.58; −8.66] | −11.41 (0.78) [−12.94; −9.87] | 0.19 | 0.23 |
| | | | | Significant according to posterior 95 % credible interval | |
| Missing at random Bayes model (Prior 1) | −8.80 (0.44) [−9.65; −7.93] | −9.36 (0.43) [−10.2; −8.52] | −10.64 (0.45) [−11.53; −9.76] | Not significant | Significant |
| Missing not at random Bayes model 1 (Prior 1) | −8.79 (0.44) [−9.65; −7.94] | −9.40 (0.43) [−10.20; −8.53] | −10.64 (0.45) [−11.52; −9.77] | Not significant | Significant |
| Missing not at random Bayes model 2 (Prior 1) | −8.79 (0.44) [−9.65; −7.94] | −9.35 (0.43) [−10.20; −8.51] | −10.65 (0.45) [−11.53; −9.77] | Not significant | Significant |
| Missing not at random Bayes model 3 (Prior 1) | −8.79 (0.44) [−9.65; −7.93] | −9.37 (0.43) [−10.21; −8.53] | −10.65 (0.45) [−11.53; −9.77] | Not significant | Significant |

A lower score means a greater improvement

*Vs* versus

[a] Values are linear model estimates, and *p* values adjusted for site and sex based on modelling the longitudinal measures at week 0 and weeks 1 through 8 as a linear function of site, sex, treatment, week, and treatment by week, with a random intercept and slope over time for each participant including all available data. Results taken from the 2002 paper (Hypericum Depression Trial Study Group 2002).

[b] All participants included and all values modelled under missing at random assumptions. The model included site, treatment, week, treatment by week and a random intercept and slope over time for each participant

**Table 3** Results fitting a mixed model

| | With multiple imputation | | | | Without multiple imputation | | | |
|---|---|---|---|---|---|---|---|---|
| | Estimate | Standard error | 95 % CI | *p* value | Estimate | Standard error | 95 % CI | *p* value |
| Intercept | 22.70 | 0.25 | 22.22; 23.19 | | 23.17 | 0.25 | 22.68; 23.66 | |
| Week | −1.15 | 0.09 | −1.33; −0.96 | <0.001 | −1.23 | 0.08 | −1.40;−1.07 | <0.001 |
| Sertraline versus placebo | 0.09 | 0.35 | −0.60; 0.78 | 0.79 | −0.08 | 0.36 | −0.75; 0.66 | 0.81 |
| Hypericum versus placebo | 0.42 | 0.35 | −0.26; 1.10 | 0.23 | 0.27 | 0.35 | −0.43; 0.97 | 0.45 |
| Week by sertraline | −0.23 | 0.13 | −0.48; 0.02 | 0.08 | −0.19 | 0.12 | −0.43; 0.05 | 0.12 |
| Week by hypericum | 0.01 | 0.13 | −0.24; 0.25 | 0.95 | 0.07 | 0.12 | −0.16; 0.31 | 0.55 |

*CI* Confidence interval

using a likelihood-based model found the only significant effect to be week, indicating that depression decreased over time (Table 3).

The choice of a vague prior did not change the results appreciably in the Bayesian missing at random analysis (Table 4). The results were similar to the maximum likelihood model, with one notable exception. Under the maximum likelihood model, none of the treatment effects was significant; however, under this model, the interaction between the sertraline arm and week was significant, indicating that the decrease in HAM-D score over time was larger in the sertraline arm than in the placebo arm, regardless of the choice of prior.

Under the missing not at random assumption, we did several Bayesian analyses as a sensitivity analysis. The results under the first missing not at random Bayes model did not differ appreciably from the results under the missing at random Bayes model. Under prior sets 1 and 2, the results were almost similar. The posterior means were slightly different under the less flat and therefore more nonrandom priors used in prior set 3. Prior set 3 was the most informative prior used in that provided the strongest prior belief against the missing not at random assumption. However, the conclusions were the same regardless of prior set used. There was a significant interaction between sertraline and week, meaning that the sertraline arm had a larger decrease in HAM-D score than the placebo arm over time. Under missing not at random model 2, the posterior means for the β-coefficients were similar to the posterior means under missing not at random model 1, with the exception of sertraline and hypericum under prior sets 2 and 3. This model was more sensitive to the choice of prior than model 1. The conclusion drawn is the same, except under prior set 2, where the interaction between sertraline and week just did not meet statistical significance.

These results need to be interpreted with caution. First, as was discussed in Jansen et al. (2006), a test for missing not at random versus missing at random is valid only under the untestable assumption that the missing not at random alternative is correctly specified. Secondly, even then, this test has been shown not to have the usual power behavior, simply because there is information missing. Evidently, this problem cannot be avoided, hence the need for sensitivity analysis.

Missing not at random model 1 is sometimes called "protective" (Michiels and Molenberghs 1997) and is specific in the sense that dropout can depend on the current, possibly unobserved, measurement, but not on the previous one. Intuitively, it is a mirror image of a commonly used missing at random model, where missingness depends on the previous but not current value. Assuming that previous and current values are often relatively similar, these models are often not too different from each other, establishing that they retain some of the stability of missing at random models. Model 2, on the other hand, allows missingness to depend on previous and current measurements, and therefore also on the increment between them. This is a profound departure from missing at random, and about the increment there is often not a lot of information in the data, because it is by definition unknown for someone dropping out, at the time of dropout (Jansen et al. 2006). As a result, it is expected that the model is more sensitive to unverifiable assumptions or choices made, such as prior specification.

## Discussion

The conclusion drawn in the original paper was that there was no difference between either the hypericum arm and placebo or the sertraline arm and placebo. This was taken to mean that the study results were inconclusive. Although it showed that hypericum was no better than placebo, it also did not find the expected difference between sertraline and placebo (Hypericum Depression Trial Study Group 2002). The same was found when the continuation data was analyzed, and the placebo effect was again noted and discussed (Sarris et al. 2012). The last observation carried forward analysis used previously was not a plausible assumption in this instance because of the week effect found. It penalized the arm with higher or earlier dropout; in this instance, the sertraline arm. This explains why the original analyses had inconclusive results. We reanalyzed the

**Table 4** Hamilton depression score posterior means, standard deviations, and credible intervals according to Bayesian analysis

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Missing at random model | | | | | | | | | |
| | Prior set 1 | | | Prior set 2 | | | Prior set 3 | | |
| | β ~ Normal (0,10 000), precision~Gamma (0.001, 0.001). | | | β ~ Normal (0,1000), precision~Gamma (0.01, 0.01). | | | β ~ Normal (0,10), precision~Gamma (0.1,0.1). | | |
| Parameter | Mean | SD | 95 % credible interval | Mean | SD | 95 % credible interval | Mean | SD | 95 % credible interval |
| $\beta_0$ – Intercept | 21.01 | 0.48 | 20.08; 21.95 | 21.0 | 0.47 | 20.06; 21.94 | 20.58 | 0.46 | 19.68; 21.49 |
| $\beta_1$ – Week | −1.17* | 0.05* | −1.28; −1.07* | −1.17* | 0.05* | −1.28; −1.07* | −1.15* | 0.05* | −1.26; −1.05* |
| $\beta_2$ – Sertraline | −0.20 | 0.67 | −1.58; 1.05 | −0.17 | 0.68 | −1.54; 1.1 | 0.24 | 0.65 | −1.07; 1.47 |
| $\beta_3$ – Hypericum | 0.74 | 0.66 | −0.59; 1.99 | 0.76 | 0.66 | −0.57; 2.01 | 1.15 | 0.66 | −0.15; 2.40 |
| $\beta_4$ – Interaction week and sertraline | −0.16* | 0.08* | −0.31;−0.01* | −0.16* | 0.08* | −0.31; −0.01* | −0.18* | 0.08* | −0.33; −0.03* |
| $\beta_5$ – Interaction week and hypericum | 0.07 | 0.08 | −0.08; 0.22 | 0.07 | 0.08 | −0.08; 0.22 | 0.05 | 0.08 | −0.09; 0.21 |
| Standard deviation from τ | 3.94 | 0.06 | 3.82; 4.07 | 3.94 | 0.06 | 3.82; 4.07 | 3.94 | 0.06 | 3.82; 4.07 |
| Standard deviation from τ2 | 4.36 | 0.19 | 4.01; 4.76 | 4.36 | 0.19 | 4.36; 4.76 | 4.37 | 0.19 | 4.01; 4.76 |
| Missing not at random (MNAR) model 1 | | | | | | | | | |
| | 1st MNAR model: prior set 1 | | | 1st MNAR model: prior set 2 | | | 1st MNAR model: prior set 3 | | |
| | β~ Normal(0,10 000), precision~Gamma (0.001, 0.001), $\theta_0$ ~logistic (0,1), Δ~Normal (0, 10 000) | | | β~Normal (0,1000), precision~Gamma (0.01, 0.01), $\theta_0$ ~ logistic (0,1), Δ~Normal (0, 100) | | | β ~ Normal (0,10), precision~Gamma (0.1,0.1), $\theta_0$ ~ logistic (0,1), Δ~Normal (0, 10). | | |
| Parameter | Mean | SD | 95 % credible interval | Mean | SD | 95 % credible interval | Mean | SD | 95 % credible interval |
| $\beta_0$ – Intercept | 21.00 | 0.45 | 20.12; 21.93 | 21.0 | 0.48 | 20.07; 21.94 | 20.53 | 0.47 | 19.59; 21.46 |
| $\beta_1$ – Week | −1.17* | 0.05* | −1.28; −1.06* | −1.17* | 0.06* | −1.28; −1.06* | −1.15* | 0.05* | −1.26; −1.04* |
| $\beta_2$ – Sertraline | −0.21 | 0.66 | −1.49; 1.08 | −0.20 | 0.67 | −1.49; 1.14 | 0.31 | 0.68 | −1.02; 1.65 |
| $\beta_3$ – Hypericum | 0.76 | 0.65 | −0.52; 2.04 | 0.74 | 0.68 | −0.60; 2.09 | 1.19 | 0.65 | −0.09; 2.48 |
| $\beta_4$ – Interaction week and sertraline | −0.16* | 0.08* | −0.31; −0.01* | −0.16* | 0.08* | −0.32; −0.01* | −0.19* | 0.08* | −0.34; −0.03* |
| $\beta_5$ – Interaction week and hypericum | 0.07 | 0.08 | −0.08; 0.22 | 0.07 | 0.08 | −0.08; 0.23 | 0.05 | 0.08 | −0.10; 0.20 |
| Δ | 0.01 | 0.05 | −0.18; 0.07 | 0.02 | 0.03 | −0.02; 0.07 | 0.02 | 0.02 | −0.02; 0.07 |
| $\theta_0$ | −4.12* | 0.84* | −5.18; −0.89* | −4.28* | 0.49* | −5.27; −3.37* | −4.30* | 0.47* | −5.29; −3.43* |
| Standard deviation from τ | 3.94 | 0.06 | 3.82; 4.07 | 3.94 | 0.06 | 3.82; 4.07 | 3.94 | 0.06 | 03.82; 4.07 |
| Standard deviation from τ2 | 4.36 | 0.19 | 4.00; 4.75 | 4.36 | 0.19 | 4.00; 4.75 | 4.37 | 0.19 | 4.01; 4.76 |
| Missing not at random model 2 | | | | | | | | | |
| | 2nd MNAR model: prior set1 | | | 2nd MNAR model: prior set 2 | | | 2nd MNAR model: prior set 3 | | |
| | β ~ Normal (0,10 000), precision~Gamma (0.001, 0.001), $\theta_0$,$\theta_1$ and $\theta_2$ ~ logistic (0,1) | | | β ~Normal (0,1000), precision~Gamma (0.01, 0.01), $\theta_0$,$\theta_1$ and $\theta_2$ ~ logistic (0,1), | | | β ~ Normal (0,10), precision~Gamma (0.1,0.1), $\theta_0$,$\theta_1$ and $\theta_2$ ~ logistic (0,1) | | |

**Table 4** (continued)

| Parameter | Mean | SD | 95% credible interval | Mean | SD | 95% credible interval | Mean | SD | 95% credible interval |
|---|---|---|---|---|---|---|---|---|---|
| $\beta_0$ – Intercept | 21.02 | 0.50 | 20.03; 21.98 | 21.12 | 0.48 | 20.18; 22.03 | 20.7 | 0.46 | 19.77; 21.61 |
| $\beta_1$ – Week | −1.18* | 0.05* | −1.28; −1.07* | −1.18* | 0.06* | −1.3; −1.08* | −1.16* | 0.05* | −1.27; −1.06* |
| $\beta_2$ – Sertraline | −0.10 | 0.73 | −1.49; 1.35 | −0.28 | 0.67 | −1.56; 1.04 | 0.17 | 0.66 | −1.10; 1.48 |
| $\beta_3$ – Hypericum | 0.79 | 0.71 | −0.58; 2.21 | 0.67 | 0.69 | −0.69; 2.00 | 1.06 | 0.66 | −0.23; 2.34 |
| $\beta_4$ – Interaction week and sertraline | −0.17* | 0.08* | −0.33; −0.01* | −0.16 | 0.08 | −0.31; <0.001 | −0.18* | 0.08* | −0.33; −0.03* |
| $\beta_5$ – Interaction week and hypericum | 0.07 | 0.08 | −008; 0.22 | 0.08 | 0.08 | −0.07; 0.24 | 0.06 | 0.08 | −0.10; 0.21 |
| Delta | −0.22 | 99.4 | −195; 191 | 0.11 | 10.05 | −19.74; 19.85 | −0.04 | 3.15 | −6.19; 6.21 |
| $\theta_0$ | −4.35* | 0.59* | −5.59; −3.26* | −4.35* | 0.55* | −5.51; −3.3* | −4.31* | 0.55* | −5.43; −3.30* |
| $\theta_1$ | 0.03 | 0.03 | −0.02; 0.09 | 0.03 | 0.03 | −0.02; 0.09 | 0.03 | 0.03 | −0.02; 0.08 |
| $\theta_2$ | 0.20* | 0.07* | 0.04; 0.34* | 0.20* | 0.07* | 0.05; 0.33* | 0.19* | 0.07* | 0.03; 0.32* |
| Standard deviation from τ | 3.97 | 0.06 | 3.84; 4.10 | 3.97 | 0.07 | 3.84; 4.10 | 3.96 | 0.07 | 3.84; 4.09 |
| Standard deviation from τ2 | 0.05 | 0.19 | 0.04; 0.06 | 4.37 | 0.19 | 4.02; 4.76 | 4.37 | 0.19 | 4.01; 4.75 |
| Missing not at random model 3 | | | | | | | | | |
| | 3rd MNAR model: prior set1 β ~ Normal(0, 10000), precision~Gamma (0.001, 0.001), $\theta_0$ ~ logistic (0,1), $\theta_1, \theta_2$ and $\theta_3$ ~ Normal (0, 10 000) | | | 3rd MNAR model: prior set 2 β ~ Normal(0,1000), precision~Gamma (0.01, 0.01), $\theta_0$ ~ logistic (0,1), $\theta_1, \theta_2, \theta_3$ ~ Normal (0, 100) | | | 3rd MNAR model: prior set 3 β ~ Normal(0,10), precision~Gamma (0.1,0.1), $\theta_0$ ~ logistic (0,1), $\theta_1, \theta_2, \theta_3$ ~ Normal (0, 10) | | |
| $\beta_0$ Intercept | 21.03 | 0.47 | 20.09; 21.97 | 20.98 | 0.48 | 20.03; 21.91 | 21.0 | 0.45 | 20.13; 21.89 |
| $\beta_1$ Week | −1.17* | 0.05* | −1.28; −1.07* | −1.17* | 0.05* | −1.28; −1.06* | −1.17* | 0.05* | −1.28; −1.06* |
| $\beta_2$ Sertraline | −0.23 | 0.68 | −1.56; 1.11 | −0.17 | 0.68 | −1.51; 1.17 | −0.19 | 0.64 | −1.45; 1.04 |
| $\beta_3$ Hypericum | 0.72 | 0.67 | −0.60; 2.04 | 0.78 | 0.67 | −0.55; 2.08 | 0.76 | 0.65 | −0.53; 2.04 |
| $\beta_4$ Interaction week and sertraline | −0.16* | 0.08* | −0.31; −0.01* | −0.16* | 0.08* | −0.31; −0.01* | −0.16* | 0.08* | −0.31; −0.01* |
| $\beta_5$ Interaction week and hypericum | 0.07 | 0.08 | −0.08; 0.22 | 0.07 | 0.08 | −0.08; 0.22 | 0.07 | 0.08 | −0.08; 0.22 |
| $\theta_0$ | −4.29* | 0.47* | −5.23; −3.42* | −4.31* | 0.50* | −5.36; −3.37* | −4.29* | 0.48* | −5.27; −3.36* |
| $\theta_1$ | 0.04 | 0.03 | −0.02; 0.09 | 0.04 | 0.03 | −0.02; 0.09 | 0.04 | 0.03 | −0.02; 0.09 |
| $\theta_2$ | −0.01 | 0.02 | −0.05; 0.03 | −0.01 | 0.02 | −0.04; 0.03 | −0.01 | 0.02 | −0.04; 0.03 |
| $\theta_3$ | −0.04 | 0.02 | −0.09; 0.002 | −0.04 | 0.02 | −0.09; 0.00 | −0.04* | 0.02* | −0.09; −0.001* |
| σ | 3.94 | 0.06 | 3.82; 4.07 | 3.94 | 0.06 | 3.82; 4.07 | 3.94 | 0.06 | 3.82; 4.07 |
| σ2 | 4.36 | 0.19 | 4.00;4.75 | 4.36 | 0.19 | 4.01; 4.75 | 4.36 | 0.19 | 4.01; 4.75 |

Letting $Y_{ij}$ be the HAM-D score for participant $i$ at occasion $j$, where $Y_{ij}$~Normal($\mu_{ij}$,$\sigma^2$), the following Bayesian model was fitted: $\mu_{ij}=\beta_0+\beta_1 t_{ij}+\beta_2 S_i+\beta_3 H_i+(\beta_4 S_i+\beta_5 H_i)t_{ij}$ where, $t_{ij}$ indicates occasion $j$ (week$_j$) for participant $i$. $S_i$ is an indicator variable for the sertraline arm and equals 1 if participant $i$ belongs to the sertraline arm and is 0 otherwise. Similarly, $H_i$ is an indicator variable for the hypericum arm and equals 1 if participant $i$ belongs to the hypericum arm and is 0 otherwise. A participant $i$ belonging to the placebo arm will thus have $H_i=S_i=0$. The placebo group is therefore the reference group. We assume $N$ independent participants

*SD* Standard deviation

*Statistically significant at 0.05 level

data using methods that are appropriate with missing data. We fitted various models using different assumptions about the missing data and various analysis methods. Under this extended sensitivity analysis, we draw a different conclusion.

The missing data in this study was not missing completely at random. We did a range of sensitivity analyses under missing at random and missing not at random assumptions. Some of our conclusions are similar to the original analysis, but both the missing at random and missing not at random analyses using Bayesian methods lead to the conclusion that sertraline is significantly better than placebo in reducing depression symptoms over 8 weeks. No difference was found between hypericum and placebo in any of the analyses. Our sensitivity analysis penalized the sertraline group less than the previous analysis and therefore showed that there was a difference between sertraline and placebo. The change from baseline to week 8 for sertraline was −10.64 (95 % CI −11.52, −9.77) and placebo was −9.36 (95 % CI −10.2, −8.52) according to the missing at random Bayesian model with prior set 1.

While there are strong similarities between likelihood and Bayesian missing at random analyses, a key difference is the absence versus presence of a prior specification. The impact of

**Table 5** Summary of all the statistical methods used; their key features and main assumptions

| Method | Key features | Assumptions |
|---|---|---|
| Missing at random | | |
| Likelihood-based approaches | Parametric model; Draw inference about a parameter, $\theta$, in a model $f(y\|\theta)$ for the response data that is not fully observed. Model with fixed and random effects, including treatment, week and the interaction between treatment and week, adjusted for repeated measures | Missing at random Missing data mechanism is ignorable. No need to specify a model that relates the missing data process to the observed data. |
| Multiple imputation | Produce several different imputed data sets. The imputed values are random draws from the posterior predictive distribution of the missing data, given the observed data. Apply likelihood-based estimation methods to each data set. Parameter estimates are averaged across the several analyses. Standard errors are calculated using Rubin's (1987) formula that combines variability within and between data sets. The imputation model included: HAM-D, age, sex, race, duration of depression, BDI, CGI-S, CGI-I, and GAF scale at baseline. | Missing at random Data are from a multivariate normal distribution. Missing values can occur on any of the variables. Missing data mechanism is ignorable. |
| Bayesian missing at random model | Bayesian model: $Y_{ij}\sim\text{Normal}(\mu_{ij},\sigma^2)$ $\mu_{ij}=\beta_0+\beta_1 t_{ij}+\beta_2 S_i+\beta_3 H_i+(\beta_4 S_i+\beta_5 H_i)t_{ij}$ The Bayesian inference is done by specifying a model and prior distributions for the parameters of the model, and then updating the prior information on the parameters using the model specified and the data observed to obtain the posterior distribution of the parameters. Vague priors were specified for the unknown parameters and are given in Table 4. | Missing at random Parameters are treated as random variables. Probability statements are made about the model parameters and not about the data |
| Missing not at random | | |
| Bayesian missing not at random model 1 | Bayesian model: $Y_{ij}\sim\text{Normal}(\mu_{ij},\sigma^2)$ $\mu_{ij}=\beta_0+\beta_1 t_{ij}+\beta_2 S_i+\beta_3 H_i+(\beta_4 S_i+\beta_5 H_i)t_{ij}$ Plus a model of missingness: $\text{logit}\, p_{i,w}=\theta_0+\theta_1 y_{i,w-1}+\theta_2(y_{i,w}-y_{i,w-1})$. Key features as described for previous missing at random model. Vague priors were specified for the unknown parameters and are given in Table 4. | Missing not at random Parameters are treated as random variables. Probability statements are made about the model parameters and not about the data |
| Bayesian missing not at random model 2 | Bayesian model: $Y_{ij}\sim\text{Normal}(\mu_{ij},\sigma^2)$ $\mu_{ij}=\beta_0+\beta_1 t_{ij}+\beta_2 S_i+\beta_3 H_i+(\beta_4 S_i+\beta_5 H_i)t_{ij}$ Plus a model of missingness: $\text{logit}\, p_{i,w}=\theta_0+\theta_1 y_{i,w-1}+\theta_2(y_{i,w}-y_{i,w-1})$. Key features as described for previous missing not at random model. Vague priors were specified for the unknown parameters and are given in Table 4. | Missing not at random Parasmeters are treated as random variables. Probability statements are made about the model parameters and not about the data |
| Bayesian missing not at random model 3 | Bayesian model: $Y_{ij}\sim\text{Normal}(\mu_{ij},\sigma^2)$ $\mu_{ij}=\beta_0+\beta_1 t_{ij}+\beta_2 S_i+\beta_3 H_i+(\beta_4 S_i+\beta_5 H_i)t_{ij}$ Plus a model of missingness: $\textit{logit}\, p_{ij}=\theta_0+\theta_1 y_{ij}+\theta_2 y_{ij} S_i+\theta_3 y_{ij} H_i$. Key features as described for previous missing not at random model. Vague priors were specified for the unknown parameters and are given in Table 4. | Missing not at random. This model allows the missingness to depend on the unobserved HAM-D value, while allowing for a different mechanism in each treatment arm by including the HAM-D score by treatment interaction. |

the prior is perhaps one of the most studied topics in Bayesian analyses, already in the context of no missing data. It suggests that the missing at random based Bayesian analysis can be more sensitive to assumptions made than the missing at random based likelihood or multiple imputation analyses, simply because more assumptions have to be made.

This implies that the conclusion reached 10 years ago could be amended to state that hypericum does not seem to provide any benefit over placebo, in a trial where it could not be ruled out that the active comparator could provide a slight benefit over placebo. This illustrates the point that not taking account of missing data in the analysis could introduce bias and lead to incorrect results.

Adjusting the analysis to take missing data into account does not imply changing the proposed estimate of effectiveness. The measure of effectiveness reported in the original paper was change from baseline to week 8. We analyzed the same estimate under missing at random assumptions using either multiple imputation or likelihood-based methods (Table 5). In general, multiple imputation allows any measure of effectiveness, since the analysis of choice is done in the second step.

The trial design only continued participants with a full response at week 8 to the continuation phase. Thus only a fraction of the participants (37.9 %) will have data in the continuation phase. Because of nonrandom exclusion of participants, the comparability of the three treatment arms after week 8 is not equivalent to a randomized trial. At best, this provides an observational study about the longer term effects of the drugs. Any efficacy analysis in this continuation phase should be interpreted with caution. This design should be discouraged, unless the objective is to estimate sustained response in those who responded initially. Because of the small number of participants in the continuation phase, the correct handling of missing data in this phase is important. The missing data mechanism is probably missing at random by design, since missingness can be predicted by the response to treatment at week 8. Missing not at random missing data cannot be excluded either, since additional mechanisms could also contribute to the missing data in this phase. It becomes even more important to analyze the data from week 8 onwards using appropriate methods for missing data.

The analysis was done with standard statistical software, using resources that should be available to most researchers. The unavailability of software should no longer be a reason not to do the proper principled analyses in the presence of missing data.

## Conclusion

There is no difference between hypericum and placebo, regardless of the assumption about the missing data process, but there is a significant difference between sertraline and placebo with some of the analyses assuming a missing at random missing data process and when a missing not at random missing data process is assumed. The assumptions about the missing data process could influence the results, as is shown by this example. This reanalysis of the original data, using proper missing data processes, changes the original conclusion of the trial. The original conclusion was that the trial was inconclusive, since the active control arm was not superior to placebo. The findings using these methods conclude that the sertraline arm could be superior to placebo under certain assumptions about the missing data process. This means that the original trial was not inconclusive, but found that hypericum was not superior to placebo. It is important to conduct an analysis that takes account of missing data using valid statistically principled methods.

**Acknowledgments** Data used in the preparation of this article were obtained from the limited access datasets (version 4.1) distributed from the NIH-supported "A Placebo-Controlled Clinical Trial of a Standardized Extract of H. perforatum in Major Depressive Disorder" (Hypericum). This is a multisite, clinical trial of persons with depression comparing the effectiveness of randomly assigned medication treatment. The purpose of this trial is to study the acute efficacy and safety of a standardized extract of the herb *H. perforatum* (St. John's wort) in the treatment of patients with major depression. The study was supported by NIMH contract # N01MH70007 to the Duke University Medical Center. The ClinicalTrials.gov identifier is NCT00005013. This manuscript reflects the views of the authors and may not reflect the opinions or views of the Hypericum Study Investigators or the NIH. Anneke Grobler had full access to the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Geert Molenberghs gratefully acknowledges financial support from the IAP research Network P7/06 of the Belgian Government (Belgian Science Policy).

**Conflict of interest** The authors declare that they have no conflicts of interest.

## References

Carpenter J, Kenward MG (2007) Missing data in randomised controlled trials—a practical guide

Carpenter J, Kenward MG, Evans S, White I (2004) Last observation carried forward and last observation analysis. Letter to the editor. Statistics in medicine 23

Carpenter J, Pocock S, Lamm CJ (2002) Coping with missing data in clinical trials: a model-based approach applied to asthma trials. Statistics in medicine 21:1043–1066

Carpenter J, Roger J, Kenward M (2013) Analysis of longitudinal trials with protocol deviation: a framework for relevant, accessible assumptions and inference via multiple imputation. J Biopharm Stat 23:1352–1371

CHMP (2010) Guideline on missing data in confirmatory clinical trials. European Medicines Agency, London

Daniels MJ, Hogan JW (2008) Missing data in longitudinal studies. Strategies for Bayesian modeling and sensitivity analysis, 1st edn. Chapman & Hall/CRC, Boca Raton, FL

Fegert J, Kolch M, Zito JM, Glaeske G, Janhsen K (2006) Antidepressant use in children and adolescents in Germany. J Child Adolesc Psychopharmacol 16:197–206

Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesion restoration of images. IEEE Trans on Pattern Anal and Mach Intell 6:721–741

Hamilton M (1960) A rating scale for depression. J Neurol, Neurosurg Psychiatry 23:56–62

Horton NJ, Kleinman KP (2007) Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. Am Stat 61:79–90

Hypericum Depression Trial Study Group (2002) Effect of *Hypericum perforatum* (St John's wort) in major depressive disorder. J Am Med Assoc 287:1807–1814

Jansen I, Hens N, Molenberghs G, Aerts M, Verbeke G, Kenward MG (2006) The nature of sensitivity in missing not at random models. Comput Stat and Data Anal 50:830–858

Kalb R, Trautmann-Sponsel RD, Kieser M (2001) Efficacy and tolerability of Hypericum extract WS 5572 versus placebo in mildly to moderately depressed patients. Pharmacopsychiatry 34:96,103

Kasper S, Gastpar M, Moller HJ, Muller WE, Volz HP, Dienel A, Kieser M (2010) Better tolerability of St. John's wort extract WS 5570 compared to treatment with SSRIs: a reanalysis of data from controlled clinical trials in acute major depression. Int Clin Psychopharmacol 25:204–213

Linde K, Berner MM, Kriston L (2008) St John's wort for major depression. Cochrane Database of Systematic Reviews 4.

Linde K, Ramirez G, Mulrow CD, Pauls A, Weidenhammer W, Melchart D (1996) St John's wort for depression—an overview and meta-analysis of randomised clinical trials. BMJ 313:253–258

Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, Frangakis C, Hogan JW, Molenberghs G, Murphy SA, Neaton JD, Rotnitzky A, Scharfstein D, Shih WJ, Siegel JP, Stern H (2012) The prevention and treatment of missing data in clinical trials. N Engl J Med 367:1355–1360

Mallinckrodt CH, Clark WS, David SR (2001) Type I error rates from mixed-effects model repeated measures versus fixed effects analysis of variance with missing values imputed via last observation carried forward. Drug Inf J 35:1215–1225

Mason A, Richardson S, Plewis I, Best N (2012) Strategy for modelling nonrandom missing data mechanisms in observational studies using Bayesian methods. J Offician Stat 28:279–302

Michiels B, Molenberghs G (1997) Protective estimation of longitudinal categorical data with nonrandom dropout. Commun in Stat, Theory and Methods 26:65–94

Molenberghs G, Kenward MG (2007) Missing data in clinical studies. Wiley, Chichester

Molenberghs G, Thijs H, Jansen I, Beunckens C, Kenward MG, Mallinckrodt C, Carroll RJ (2004) Analyzing incomplete longitudinal clinical trial data. Biostatistics 5:445–464

National Research Council (2010) The prevention and treatment of missing data in clinical trials. The National Academic Press, Washington DC

Rahimi R, Nikfar S, Abdollahi M (2009) Efficacy and tolerability of *Hypericum perforatum* in major depressive disorder in comparison with selective serotonin reuptake inhibitors: a meta-analysis. Prog Neuro-Psychopharmacol Biol Psychiatry 33:118–127

Ratitch B, O'Kelly M, Tosiello R (2013) Missing data in clinical trials: from clinical assumptions to statistical analysis using pattern mixture models. Pharmaceutical Statistics: n/a-n/a.

Rubin DB (1976) Inference and missing data. Biometrika 63:581–592

Sarris J, Fava M, Schweitzer I, Mischoulon D (2012) St John's Wort (*Hypericum perforatum*) versus sertraline and placebo in major depressive disorder: continuation data from a 26-week RCT. Pharmacopsychiatry 45:275–278

Shelton RC, Keller MB, Gelenberg A, Dunner DL, Hirschfeld R, Thase ME, Russell J, Lydiard B, Crits-Christoph P, Gallop R, Todd L, Hellerstein D, Goodnick P, Keitner G, Stahl SM, Halbreich U (2001) Effectiveness of St John's wort in major depression: a randomized controlled trial. JAMA 285:1978–1986

Ware JH, Harrington D, Hunter DJ, D'Agostino RB (2012) Missing Data. N Engl J Med 367:1353–1354

White IR, Royston P, Wood AM (2011) Multiple imputation using chained equations: issues and guidance for practice. Stat Med 30: 377–399