Missing Data Methodology:- Sensitivity analysis after multiple imputation

Melanie Smuk

Thesis submitted in accordance with the requirements for the degree of Doctor of Philosophy of the University of London
May 2015

Department of Medical Statistics
Faculty of Epidemiology and Population Health
London School of Hygiene and Tropical Medicine
Funded by the Medical Research Council Clinical Trials Unit, London

In memory of Abhirami Wimalathasan & Sydney Burnett.

Declaration

Statement of Own Work

All students are required to complete the following declaration when submitting their thesis. A shortened version of the School's definition of Plagiarism and Cheating is as follows (the full definition is given in the Research Degrees Handbook):

The following definition of plagiarism will be used:

Plagiarism is the act of presenting the ideas or discoveries of another as ones own. To copy sentences, phrases or even striking expressions without acknowledgement in a manner which may deceive the reader as to the source is plagiarism. Where such copying or close paraphrase has occurred the mere mention of the source in a biography will not be deemed sufficient acknowledgement; in each instance, it must be referred specifically to its source. Verbatim quotations must be directly acknowledged, either in inverted commas or by indenting. (University

of Kent)

Declaration by candidate

I have read and understood the School's definition of plagiarism and cheating given in the Research Degrees Handbook. I declare that this thesis is my own work, and that I have acknowledged all results and quotations from the published or unpublished work of other people.

Signed:

Date: **May 2015**

Full name: **Melanie Smuk**

Abstract

Within epidemiological and clinical research, missing data are a common issue which are often inappropriately handled in practice. Multiple imputation (MI) is a popular tool used to 'fill in' partially observed data with plausible values drawn from an appropriate imputation distribution. Software generally implements MI under the assumption that data are 'missing at random' (MAR) i.e. that the missing mechanism is not dependent on the missing data conditional on the observed data. This is a strong inherently untestable assumption, and if incorrect can result in misleading inferences. The sensitivity of inferences to this assumption needs to be assessed by exploring the alternative assumption that missing data are 'missing not at random' (MNAR) i.e. even conditioned on the observed data, the probability of missing observations depends on their unseen, underlying values.

Broadly there are two ways to frame, and perform sensitivity analyses (SA) to accomplish this: using a pattern mixture model or a selection model. Motivated by a cancer dataset, we develop a novel pattern mixture approach to collecting and incorporating in the analysis prior

information elicited from experts. We demonstrated the inferential validity of our approach by simulation.

Our second example is an individual patient meta-analysis of sudden infant death syndrome studies. We extended existing multilevel MI software to perform SA for the risk of bed sharing in these complex data. Inferences were found to be robust.

Finally we considered a proposal of Carpenter *et al.* (2007) for SA after MI by reweighting. We developed a modification, which dramatically improves its performance in small data sets.

The routine use of SA in applied research is held back by the lack of practical methodology and examples. This thesis addresses these issues, and so lowers the barrier to the widespread adoption of SA.

Acknowledgements

My sincere gratitude goes to my supervisor, Professor James Carpenter, for his wisdom and guidance. His supervision opened up the fascinating world of statistics and supported my enthusiasm to learn.

I wish to thank the Medical Research Council Clinical Trials Unit London for funding my research, without your support this thesis would not be possible.

I would also like to thank my friends, family and loved ones, for their continuing support and encouragement throughout the last few years and my life.

Acronyms and Abbreviations

AECM Alternating Expectation Conditional Maximisation

AIPW Augmented Inverse Probability Weighting

AOR Adjusted Odds Ratios

CI Confidence Interval

CM Conditional Maximisation

CR Complete Record

DF Degrees of Freedom

ECM Expectation Conditional Maximisation

ECME Expectation Conditional Maximisation Either

EM Expectation Maximisation

GEE Generalised Estimating Equations

GEM Generalise Expectation Maximisation

HES Hospital Episode Statistics

IPW Inverse Probability Weighting

LOCF Last Observation Carried Forward

MAR Missing At Random

MCAR Missing Completely At Random

MCMC Markov Chain Monte Carlo

MI Multiple Imputation

MICE Multiple Imputation Chained Equations

MMI Multiple Multilevel Imputation

MNAR Missing Not Random

MOI Model Of Interest

NCDR National Cancer Data Repository

NHS National Health Service

OR Odds Ratio

PX-EM Parameter Expanded Expectation Maximisation

SA Sensitivity Analysis

SAGE Space Alternating Generalised Expectation Maximisation

SIDS Sudden Infant Death Syndrome

Table of contents

List of tables

List of figures

1

Introduction

1.1 Background

This thesis is motivated by the relatively poor handling of missing data, and particularly sensitivity analysis, in publications. It argues that sensitivity analysis after multiple imputation is important, and tackles applied and methodological barriers to its widespread adoption in practice. We define missing data as data which we sought to collect but were unable to do so for one reason or another. Therefore the data actually exists but we cannot observe it. We introduce the issues involved with missing data using Rubin's missing data mechanism framework (Rubin, 1976). This framework can be expanded to think about counterfactual data. An example is data from a clinical trial, where we may consider what might have happened if a placebo patient took the active treatment. This data is counterfactual as the patient did not take the active treatment. While this thesis does not explore counterfactual data, the framework is conceptually similar.

Within epidemiological and clinical research, missing data are a common issue and often over looked both in publications of trials, and in the analysis of observational data, e.g., among many, Mackinnon (2010). In an ideal setting missing data would be non existent, however due to cost, time and feasibility this would be very difficult to achieve. Multiple imputation (MI) is an increasingly popular tool — see Figure 1.1 — used to 'fill in' partially observed data with plausible values drawn from an appropriate imputation distribution to allow valid inference. Software generally implements MI under the assumption that missing data are 'Missing At Random' (MAR) i.e. that the missingness mechanism is not dependent on the missing observations conditional on the observed data. This is a strong inherently untestable assumption, and if incorrect can result in misleading inferences and conclusions. Thus the sensitivity of

the inferences and conclusions to this assumption needs to be assessed by exploring robustness of inferences to the alternative assumption that missing data are 'Missing Not At Random' (MNAR). The MNAR mechanism states that, even conditional on the observed data, the probability of data being unobserved (i.e. missing) depends on their unseen, underlying value. At present there is no general consensus about how sensitivity analysis should be framed and performed either in experimental or observational studies. The goal of this thesis is to develop, apply and advocate sensitivity analysis using multiple imputation. Our two key motivating examples are observational studies, with missing covariate data. However the pattern mixture approach may also be applied to missing outcome data, which often arises in clinical trails, although the details differ. Further, the method we develop in Chapter 5 can be applied to both missing outcome data (the setting we consider) and to missing covariate data (Bousquet *et al.*, 2012).

To further motivate the thesis, we performed a search of 16 online databases (including PubMed and MEDLINE) looking at all publications from January 2000 with "Multiple Imputation" (or synonyms) in either the title, abstract or keywords (see Appendix A for database list). A similar second search was completed, with "Missing Not At Random" (MNAR) in union with "Multiple Imputation" (or synonyms). Results were cross examined to remove duplicates of articles. There were 2194 publications found in the search.

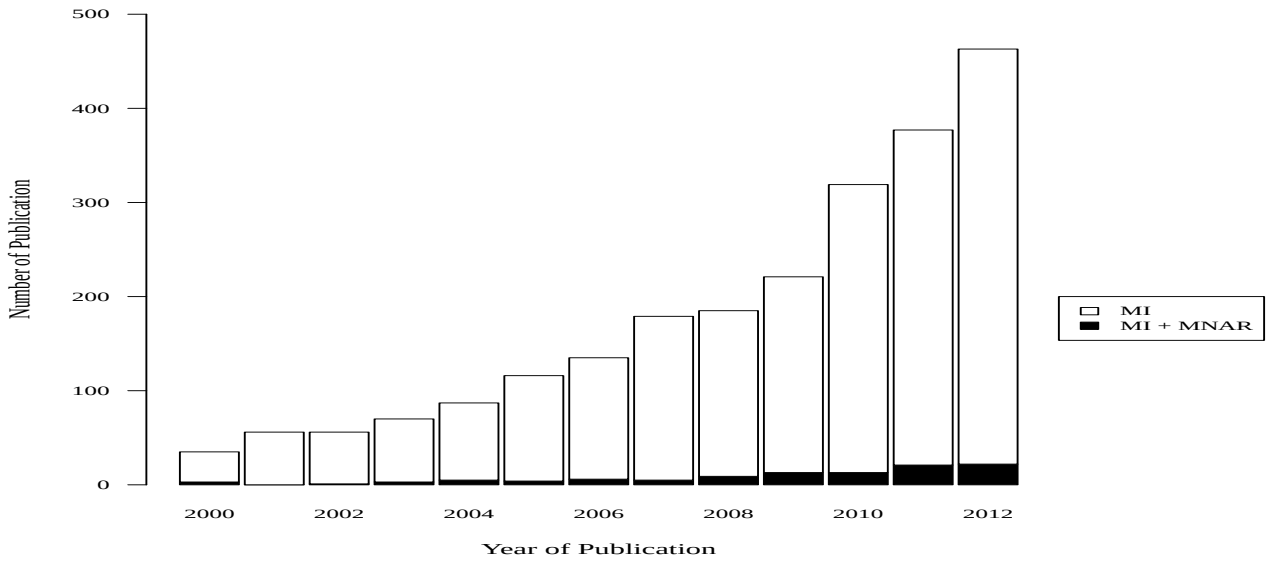

Figure 1.1: Number of publications between 1st January 2000 and 31st December 2012, shown for publications which contained the words "multiple imputation" and publications with "multiple imputation" and "missing not at random" within the title, abstract or keywords.

Figure 1.1 shows an approximately exponential increase in the number of publications which use or explore MI; however the number of publications which mention MNAR is only slowly increasing. We make the assumption that if sensitivity analysis of the MAR assumption was completed, then the article would include the phrase "Missing Not At Random". Even if it is only roughly an approximation to the truth, it is nevertheless clear that sensitivity of inferences obtained via MI to the assumption that data are MAR is typically not published and very likely not been undertaken (Carpenter *et al.*, 2012b). Since inferences may often not be robust to the MAR assumption, there is a need for valid, accessible and practical approaches to sensitivity analysis. Rubin (1976) pointed out that MI is well suited to sensitivity analysis and it seems clear that its full potential in this regard is as yet largely unrecognised and untapped. This thesis aims to develop and apply sensitivity analysis through MI, demonstrating

its advantages/importance, and overcoming barriers to its widespread use.

1.2 Outline of this thesis

This thesis is divided into six chapters. Chapter 2 explores the various methods for analysing data with missing observations. We explain how each method works mathematically and discuss its limitations and advantages. Chapter 3 applies model based sensitivity analysis using a pattern mixture approach to a motivating colorectal cancer data set. Chapter 4 first applies multilevel multiple imputation for a multi-centre case-control study of sudden infant death. After imputing the data we apply pattern mixture sensitivity analysis to investigate the robustness of inferences to departure from the MAR assumption. In Chapter 5 we introduce a method proposed by Carpenter *et al.* (2007) for re-weighting data sets imputed under MAR as an approximation to selection modelling. The method is then developed and evaluated for small data sets where the initial proposal performed poorly. The final chapter summarises the thesis conclusions and discusses areas of further research.

2

Literature Review

2.1 Introduction

In this chapter we discuss what missing data are, missing data concepts, and review important missing data patterns and mechanisms. We then use a simple linear regression example to introduce and discuss the various methods that have been proposed to handle missing data in the literature.

2.2 What Are Missing Data

Missing data are observations which exist but were not recorded or recorded and then lost. In clinical studies missing data often result from withdrawal, attrition and loss to follow up. In other settings the missing data could be generated through a coarsening design. An example of a coarsening design is when continuous data are deliberately collected in intervals to maximise the chance of response. A classic example is collecting income in ranges, see Heitjan and Rubin (1991). Often there is some information about why data are missing, however the information is generally not definitive. The best way to avoid missing data issues is to have a good study design which reduces the chance of missing data occurring. Often the simplest method for handling missing data in the analysis is simply to use only those records which have been fully observed (complete record analysis). We therefore begin with this approach, after introducing Rubin's terminology using an illustrative example for missing data mechanisms.

2.3 Missing Data Methodology Illustrative Example

We introduce here, a simple linear regression model as an illustrative example to discuss Rubin's terminology and methods to handle missing data. The simple linear regression model has n pairs (Y_i, X_i), where observations Y_i are conditionally independent given the observations X_i. We write this as

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + \varepsilon_i \qquad i = 1, \ldots, n, \tag{2.3.1}$$

where $\varepsilon_i \sim N(0, \sigma^2)$ i.i.d and $\boldsymbol{X}_i$ is the design matrix. Let R_i be a binary missing value indicator with $R_i = 0$ implying that either Y_i or X_i is missing depending on the context of the example. Let $(\boldsymbol{Y}, \boldsymbol{X}) = \boldsymbol{Z} = (\boldsymbol{Z}^m, \boldsymbol{Z}^o)$ where $\boldsymbol{Z}^o$ is defined as observed data and $\boldsymbol{Z}^m$ as missing data.

2.4 Missing Data Patterns and Monotone Missing Data

A missing data pattern describes the pattern of missing and observed data. We begin by discussing an important missing data pattern called monotone. A monotone missing data pattern is when the variables can be arranged so that $X_{j+1}, \ldots, X_M$ are all missing if X_j is missing, where $j = 1, \ldots, M$ represents variables in the data set or more realistically repeat measurements. For example if the data set contained 10 records and 5 variables ($M = 5$), the data would have a monotone missing data pattern if the variables could be arranged so the frequency of missing observations by record increases, see Figure 2.1.

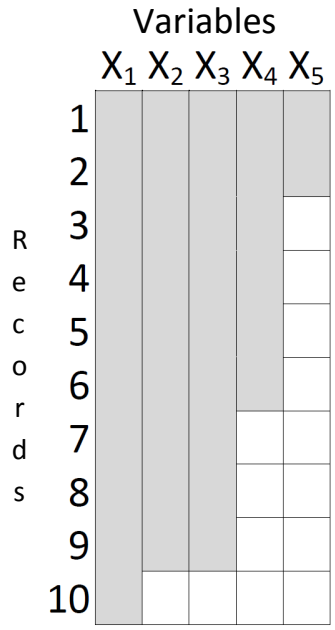

Figure 2.1: A visual illustration of a monotone missing data pattern, with 5 variables and 10 records. The unshaded areas represent missing observations

The monotone missing data pattern is often (at least approximately) observed in clinical longitudinal studies with repeat measurement. These studies often suffer from attrition where patients drop out before the end of the study. The presence of a monotone missing pattern in other settings is infrequent. When data are not monotone the reason for data being missing is generally more difficult to model. Exploring the missing data pattern gives a better insight into the missing data mechanism.

2.5 Missing Data Mechanisms

A missing data mechanism specifies how the underlying value of the missing observation is associated with the reason for being missing. Understanding the missing data mechanism is a key stage in comprehending the impact of the missing data on a specific analysis, or missing data methods. Rubin (1976) classified missing data mechanisms as Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR). These are now universally used terminology.

Using the illustrative example (2.3.1) (with Y_i partially observed), Y_i is said to be MCAR if the distribution of R_i is independent of the possibly unobserved value of Y_i and the value of X_i. This means that $P(R_i|Y_i, X_i) = P(R_i)$ which implies $P(Y_i|X_i, R_i) = P(Y_i|X_i)$ and $P(Y_i|R_i) = P(Y_i)$. This means the conditional and marginal distribution can always be validly estimated from the observed data. The advantage of MCAR is it allows the data to be analysed as if there had been no missing observations. MCAR can be seen as a special case of MAR, defined below.

The variable Y_i is said to be MAR if the distribution of R_i is conditionally independent of the possibly unobserved Y_i given X_i. This means $P(R_i|Y_i, X_i) = P(R_i|X_i)$ which implies $P(Y_i|X_i, R_i) = P(Y_i|X_i)$; however now $P(Y_i|R_i) \neq P(Y_i)$. This means that given the observed data, the missing indicator does not depend on the unobserved data. The terminology MAR does not mean that the missing data values are a simple random sample of the data values. MAR is often seen as less restrictive than MCAR "as it requires only that the missing values behave like a random sample of all values within subclasses defined by observed data" (Schafer,

1999).

When MCAR and MAR are not valid we say the data is MNAR. The variable Y_i is said to be MNAR if the distribution of R_i is dependent on the unobserved Y_i value even when conditioned on X_i, thus $P(R_i|Y_i, X_i)$ does not simplify. This implies $P(Y_i|X_i, R_i) \neq P(Y_i|X_i)$ and $P(Y_i|X_i, R_i = 1) \neq P(Y_i|X_i, R_i = 0)$. To create unbiased inferences, a joint model of both $\boldsymbol{Y}$ and $\boldsymbol{R}$ is required, however there will typically be a very wide range of models for $\boldsymbol{R}$ compatible with the observed data, which may result in very different substantive inferences.

There is no test which can definitively show if the missing mechanism is MCAR, MNAR or MAR, thus when applying missing data methods we are required to make an assumption. There are many different methods to handling missing data, choosing which one to apply should depend on both the missing pattern and the missing data mechanism assumption. We next discuss how these mechanisms affect inferences in a complete records approach.

2.6 Complete Record (CR) Analysis

A Complete Record (CR) (also known as a complete case) analysis is based on the subset of complete records; i.e. records with missing values are not included. For example if any data for record i is missing then all data for record i is removed from the dataset. CR is valid in a regression setting if the probability of being a CR is independent of the outcome when conditioned on the explanatory variables, regardless of whether they contain missing observations or not (White and Carlin, 2010). This definition includes the case of MCAR (Myers, 2000).

Suppose we have four variables $\boldsymbol{W}$, $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$, with $\boldsymbol{Y}$ being partially observed and the other variables being complete. Let our model of interest be a regression of $\boldsymbol{Y}$ on $\boldsymbol{X}$ and $\boldsymbol{Z}$. If $\boldsymbol{Y}$ is MAR given $\boldsymbol{X}$ and $\boldsymbol{Z}$, then CR analysis will be unbiased as the contributions to the likelihood from records with missing data are integrated out. If $\boldsymbol{Y}$ is MAR given $\boldsymbol{W}$, but $\boldsymbol{W}$ is independent of $\boldsymbol{Y}$, then $\boldsymbol{W}$ contains no information about $\boldsymbol{Y}$ so complete record analysis will be unbiased. In the setting where $\boldsymbol{W}$ and $\boldsymbol{Y}$ are associated and $\boldsymbol{W}$ is predictive of $\boldsymbol{Y}$ being missing ($\boldsymbol{Y}$ is thus MAR given $\boldsymbol{W}$), the inferences will generally be biased if $\boldsymbol{W}$ is omitted from the analysis, because missingness will then depend on $\boldsymbol{Y}$ and thus the mechanism will be MNAR. If $\boldsymbol{Y}$ is MNAR, then the inferences will be biased unless a valid assumption about the difference between the observed data distribution and missing data distribution is made.

Suppose again we have the same four variables and model of interest, but now the covariate $\boldsymbol{X}$ is partially observed. If $\boldsymbol{X}$ is MCAR when complete records will produce unbiased inferences however in this setting we can recover information on $\boldsymbol{X}$ from the observed data by modelling $\boldsymbol{X}$ in terms of $\boldsymbol{Y}$ and $\boldsymbol{Z}$, for example through multiple imputation. If $\boldsymbol{X}$ is MAR given $\boldsymbol{Z}$, with the missing mechanism independent of $\boldsymbol{Y}$ given $\boldsymbol{Z}$, then inferences from complete records will again be unbiased. However if $\boldsymbol{X}$ is MAR and the missing mechanism is dependent on $\boldsymbol{Y}$ and $\boldsymbol{Z}$ then complete records will be biased. Similarly if $\boldsymbol{X}$ is MAR with the missing mechanism dependent on $\boldsymbol{W}$ (not in the model of interest), with $\boldsymbol{W}$ and $\boldsymbol{Y}$ associated, then complete record analysis will be biased. If $\boldsymbol{X}$ is MNAR depending on $\boldsymbol{X}$ or $\boldsymbol{Z}$ but independent of $\boldsymbol{Y}$, then complete records analysis is unbiased. Finally if $\boldsymbol{X}$ is MNAR depending on $\boldsymbol{X}$, $\boldsymbol{Y}$ and maybe $\boldsymbol{Z}$, then again the complete records analysis will be biased.

We demonstrate the CR behaviour under various missing mechanisms with partially observed

$\boldsymbol{X}$ through a simulation. We created 5 million observations from the model:

$$\boldsymbol{Y} = \beta_0 + \beta_1 \boldsymbol{X} + \beta_2 \boldsymbol{Z} + \boldsymbol{e} \tag{2.6.1}$$

where $\boldsymbol{X} \sim N(0,1)$ similarly for $\boldsymbol{Z}$ and $\boldsymbol{e}$. We made approximately half the observations missing using different mechanisms and set $\beta_0 = 10$, $\beta_1 = 5$ and $\beta_2 = 3$. The results can be seen in Table 2.1.

Mechanism depends on:	$\boldsymbol{X}$ is missing when	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
-	-	10.00	5.00	3.00
$\boldsymbol{Y}$	$\boldsymbol{Y} > E[\boldsymbol{Y}]$	9.65	4.76	2.86
$\boldsymbol{X}$	$\boldsymbol{X} > E[\boldsymbol{X}]$	10.00	5.00	3.00
$\boldsymbol{Z}$	$\boldsymbol{Z} > E[\boldsymbol{Z}]$	10.00	5.00	3.00
$\boldsymbol{X}$, $\boldsymbol{Z}$	$E[\boldsymbol{X} + \boldsymbol{Z}] > E[E[\boldsymbol{X}] + E[\boldsymbol{Z}]]$	10.00	5.00	3.00
$\boldsymbol{Y}$, $\boldsymbol{X}$	$E[\boldsymbol{Y} + \boldsymbol{X}] > E[E[\boldsymbol{Y}] + E[\boldsymbol{X}]]$	9.69	4.78	2.89
$\boldsymbol{Y}$, $\boldsymbol{Z}$	$E[\boldsymbol{Y} + \boldsymbol{Z}] > E[E[\boldsymbol{Y}] + E[\boldsymbol{Z}]]$	9.67	4.80	2.84
$\boldsymbol{Y}$, $\boldsymbol{X}$, $\boldsymbol{Z}$	$E[\boldsymbol{X} + \boldsymbol{Y} + \boldsymbol{Z}] > E[E[\boldsymbol{X}] + E[\boldsymbol{Y}] + E[\boldsymbol{Z}]]$	9.71	4.81	2.87

Table 2.1: Estimated coefficients from complete records when $\boldsymbol{X}$ is partially observed, with various missing data mechanisms.

We see from Table 2.1 that biased coefficients are created when the missing data depends on the outcome $\boldsymbol{Y}$. In our simulation example we have used the missing mechanism to create missing observations in $\boldsymbol{X}$, however if we retained the missing mechanism but instead created missing observations in $\boldsymbol{Y}$ or $\boldsymbol{Z}$ the same bias will occur. It is thus important to note that it does not matter where the missing observations occur; rather it is the mechanism which creates bias in inferences when we restrict to the complete records. Further complications arise when estimators possess a symmetry (i.e. odds ratios); this has been explored in Harel and Carpenter (2014). An empirical illustration similar to Table 2.1 with logistic regression is given by Carpenter and Kenward (2013, p.33).

In conclusion the CR method is often chosen due to its ease of application. However the CR subset can be very small when the number of missing observations is high, affecting modelling and power (Nakai and Ke, 2011). When fitting several regression models with varying sets of explanatory variables, the subset size is different each time, affecting the validity of model comparisons. We discussed that bias in inferences restricted to the complete records is dependent on the predictors of complete records regardless of where the missing observations are. However, to include information in the partially observed records, we need to explicitly consider where the missing data occur, and the corresponding form of the missing data mechanism. In the next section we discuss methods which are often applied under the missing data mechanism assumption of MAR, we begin with model based methods.

2.7 Expectation Maximisation (EM) Algorithm

The Expectation Maximisation (EM) algorithm is a relatively old idea for handling missing data. It was first noted as a missing data method in a medical application by Mckendrick (1926). It was developed by Orchard and Woodbury (1972), but was constrained by the assumption that the full data had a multivariate normal distribution. The method was later extended by Dempster *et al.* (1977) to a wider range of full data models, and can be applied to many different problems (Meng and Pedlow, 1992). It is typically used for point estimation in missing data problems.

The EM algorithm is an iterative process for maximising log-likelihoods involving missing observations. It is useful when the full data likelihood cannot be readily integrated to yield

the observed data likelihood for direct maximisation.

Using the illustrative example (2.3.1), the EM algorithm consists of two steps; the first is called the E step (expectation) and the second the M step (maximisation). The E step finds the expected value for the log-likelihood (ℓ) with respect to the conditional expectation of $\boldsymbol{Z}^m$ given $\boldsymbol{Z}^o$ and current estimate of $\boldsymbol{\beta}$, say $\boldsymbol{\beta}^{(t)}$:

$$Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(t)}) = E_{\boldsymbol{Z}^m|\boldsymbol{Z}^o,\boldsymbol{\beta}^{(t)}}[\ell(\boldsymbol{\beta}|\boldsymbol{Z}^o, \boldsymbol{Z}^m)]. \tag{2.7.1}$$

Since $E[g(x)|y] = \int_{-\infty}^{\infty} g(x)P(x|y)dx$, (2.7.1) becomes for the t^{th} iteration:

$$\int \ell(\boldsymbol{\beta}|\boldsymbol{Z})P(\boldsymbol{Z}^m|\boldsymbol{Z}^o, \boldsymbol{\beta} = \boldsymbol{\beta}^{(t)})d\boldsymbol{Z}^m \tag{2.7.2}$$

The M step maximises the expected complete data log-likelihood (2.7.2):

$$\boldsymbol{\beta}^{(t+1)} = arg\ max_{\boldsymbol{\beta}}\ Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(t)})$$

The steps are generally straight forward to derive, and can be approximated computationally (Meng and Rubin, 1993; Dempster *et al.*, 1977); however the EM algorithm may converge slowly if $\boldsymbol{Z}^M$ is a large subset. Another issue is that the EM algorithm does not automatically provide standard errors for the parameter estimates (Pawitan, 2001); for this the Louis formulation (Louis, 1982) is needed, which typically entails considerable model specific calculations.

2.8 Variants of the EM Algorithm

Sometimes the M step in the EM algorithm does not have a simple computational form. To avoid this we can increase the Q function instead of maximising it, resulting in a generalised expectation maximisation (GEM) algorithm (Dempster *et al.*, 1977). "GEM algorithms increase the log-likelihood at each interaction, but appropriate convergence is not guaranteed without further specification of the process of increasing the Q function" (Little and Rubin, 2002). Another variant of the EM algorithm is the Expectation Conditional Maximisation (ECM) (Meng and Rubin, 1993) algorithm, which is a subset of the GEM algorithm. The ECM replaces the complicated M step of the EM algorithm with multiple Conditional Maximisation (CM) steps. These steps are simpler than the original M step as the estimated Q function found in the preceding E step is maximised with constraints on each β conditional on the other parameters remaining fixed. The constraints together must result in the maximisation over the parameter space of β.

Other extensions of the EM algorithm are the Expectation Conditional Maximisation Either (ECME)(Liu and Rubin, 1994; Liu *et al.*, 1998), the Space Alternating Generalised EM (SAGE)(Fessler and Hero, 1995), the Alternating ECM (AECM)(Meng and van Dyk, 1997), and the Parameter Expanded EM (PX-EM).

2.9 Direct Maximum Likelihood Estimation

As we noted, a disadvantage of the EM algorithm is that is does not provide standard error estimates. The direct maximum likelihood method (also known as a full information maximum likelihood) solves this problem at the cost of computational simplicity. However in cases where the longitudinal outcome measures are partially observed, this can be done relatively simply (Beale and Little, 1975). The method works by estimating the observed data likelihood for each record based on the variables that were observed for that particular record. For example let continuous outcome Y_i be regressed on X_i (which is partially observed) and $\boldsymbol{V}_i$ which are full observed covariates for unit i. The observed log-likelihood function for each record for which a binary variable X_i is missing, would be:

$$\ell_i = \ell_i(\boldsymbol{\beta}, \gamma; Y_i, \boldsymbol{V}_i) = \sum_{x=0,1} f\left(Y_i | X_i = x, \boldsymbol{V}_i, \boldsymbol{\beta}\right) \times p(X_i = x; \gamma). \tag{2.9.1}$$

Here $f(Y_i|X_i = x, \boldsymbol{V}_i, \boldsymbol{\beta})$ is the usual log-likelihood for the model, and $p(X_i = x; \gamma)$ is the marginal model for the partially observed binary covariate X_i, parameterised by γ. Otherwise if $\boldsymbol{X}$ is continuous,

$$\ell_i = \int f\left(Y_i | X_i, \boldsymbol{V}_i, \boldsymbol{\beta}\right) f(X_i, \gamma) dX, \tag{2.9.2}$$

where $p(X; \gamma)$ is the marginal model for X_i. The log-likelihood function are then accumulated for all records,

$$\ell(\boldsymbol{\beta}|\boldsymbol{Z}^o) = \sum_{i=1}^{N} \ell_i \tag{2.9.3}$$

where i denotes the record ($i = 1, \ldots, N$), $\boldsymbol{Z}^o$ defines the observed data and ℓ is the log-likelihood. This can then be maximised to find $\hat{\boldsymbol{\beta}}$. The standard errors can be calculated by conventional maximum likelihood methods for example taking the negative inverse of the

information matrix see (Arbuckle, 1996). A disadvantage of the direct maximum likelihood estimation method is that not all likelihoods can be factored easily and the method entails awkward numerical integration over the distribution of the covariates.

We now move on to describe weighting approaches to missing data. These methods have the advantage that the probability distribution of the missing data does need to be specified (removing bias due to possible misspecification) and thus can reduce complexity in some settings.

2.10 Inverse Probability Weighting (IPW)

In a simple setting Inverse Probability Weighting (IPW) has a long history as a method for correcting for biases caused by missing data going back beyond the seminal paper by Horvitz and Thompson (1952). The approach weights complete records by the inverse probability of observing the data. This is achieved generally by finding the probability of records being complete as a function of the observed data through a logistic regression. Observations which were observed but had a high probability of being missing will be up-weighted relative to observations which had a low probability of being missing and were observed (Carpenter *et al.*, 2011b).

Using the illustrative example (2.3.1) we take scalar X_i as MAR given fully observed Y_i. Assume we are interested in solving the normal equation for $\boldsymbol{\beta}$ (assuming no missing data):

$$\sum_{i=1}^{n} X_i(Y_i - X_i\boldsymbol{\beta}) = 0. \tag{2.10.1}$$

If the expectation of (2.10.1) over $P(Y_i|X_i)$ at the true value of $\boldsymbol{\beta}$ is zero, then the estimate obtained by solving (2.10.1) will be consistent (Cox and Hinkley, 1974). In a setting with X_i missing for some individuals, (2.10.1) can be written:

$$\sum_{X_i Observed} X_i(Y_i - X_i\boldsymbol{\beta}) + \sum_{X_i Missing} X_i(Y_i - X_i\boldsymbol{\beta}) = 0. \tag{2.10.2}$$

Now let $R_i = 1$ if X_i is observed and 0 otherwise. Assume X_i is MAR, so that $P(R_i = 1|Y_i, X_i) = P(R_i = 1|Y_i) = \pi_i(Y_i)$ say. Then (2.10.2) can be re-written as

$$\sum_{i=1}^{n} R_i X_i(Y_i - X_i\boldsymbol{\beta}) = 0. \tag{2.10.3}$$

If we take the expectation of the sum at the true $\boldsymbol{\beta}$ this is no longer zero, as R_i is only 1 for a non random selection of individuals, i.e. $[Y_i|R_i = 1] \neq [Y_i|R_i = 0]$, and $E[R_i] = \pi_i(Y_i)$ is a function of Y_i. However, if we include the inverse probability weights we have

$$\sum_{i=1}^{n} \frac{R_i}{\pi_i} X_i(Y_i - X_i\boldsymbol{\beta}) = 0. \tag{2.10.4}$$

Now we can first take expectation over $[R_i|Y_i, X_i]$ to obtain $\sum_{i=1}^{n} X_i(Y_i - X_i\boldsymbol{\beta})$ and then take expectation over $[Y_i|X_i]$ at the true value of $\boldsymbol{\beta}$. In this way we see that the expectation of (2.10.4) is zero when the inverse probability weights are included, giving a consistent estimate of $\hat{\boldsymbol{\beta}}$.

Under MAR, IPW gives consistent parameter estimates, however they are inefficient since all information from partially observed records is discarded (Carpenter *et al.*, 2006; Clayton *et al.*, 1998). IPW may also be unstable if the weight model is imprecise, and is difficult to apply when the pattern of missing is non-monotone (Rubin, 1987; Kang and Schafer, 2007).

A more efficient IPW method (augmented inverse probability weighting) has been proposed by Robins *et al.* (1995) which makes IPW both more efficient and less sensitive to imprecise weight models. In the next section we show a setting where a doubly robust IPW (Carpenter *et al.*, 2006; Vansteelandt *et al.*, 2010), can give consistent point estimates when the weight model or distribution of missing data given observed is correctly specified.

2.11 Augmented Inverse Probability Weighting (AIPW)

The augmented inverse probability weighting (AIPW) estimator (Robins and Rotnitzky, 1995; Robins *et al.*, 1994; Tsiatis, 2006) is a more efficient than the Inverse Probability Weighting (IPW) estimator. The IPW estimator (2.10.4), uses only the CR data and thus efficiency is lost due to information from partially observed data being discarded. The AIPW adds a term with expectation equal to zero to the IPW estimator, which contributes the partially observed data information. As the expectation of the term is zero the conditional expectation argument is unaffected, resulting in the estimate being still consistent.

Recalling from the previous section that $R_i = 1$ if X_i is observed, and 0 otherwise, and that X_i is MAR with $P(R_i = 1|X_i, Y_i) = \pi_i(Y_i)$, a plausible augmented IPW (2.10.4) is

$$\sum_{i=1}^{n} \left[\frac{R_i}{P(R_i = 1)} X_i(Y_i - X_i\boldsymbol{\beta}) - \left(\frac{R_i}{P(R_i = 1)} - 1 \right) \phi(Y_i) \right]. \tag{2.11.1}$$

The augmented parameter $\phi(.)$ is a function of Y_i, and hence when X_i is missing the record still contributes to the estimator. The calculations to find $\phi(X_i)$ are explored in Robins and

Rotnitzky (1995). The calculation requires the mapping of the residual of the orthogonal projection of the function

$$\frac{R_i}{P(R_i = 1)} X_i(Y_i - X_i\boldsymbol{\beta}) \tag{2.11.2}$$

onto the space of mean zero function

$$\Delta = \left[\left(\frac{R_i}{P(R_i = 1)} - 1 \right) \phi(Y_i) : \phi \text{ arbitrary} \right]. \tag{2.11.3}$$

The calculations to find ϕ involves Hilbert space theory and are beyond the scope of this thesis. More details can be found in Robins *et al.* (1995). In our setting the projection gives

$$\phi(Y_i) = E_{X_i|Y_i}[X_i(Y_i - X_i\boldsymbol{\beta})|Y_i]. \tag{2.11.4}$$

This may be estimated from the data since X_i is MAR given Y_i. Then the efficient AIPW estimate of $\boldsymbol{\beta}$ is found by solving

$$\sum_{i=1}^{n} \left[\frac{R_i}{P(R_i = 1)} X_i(Y_i - X_i\boldsymbol{\beta}) - \left(\frac{R_i}{P(R_i = 1)} - 1 \right) E_{X_i|Y_i}[X_i(Y_i - X_i\boldsymbol{\beta})|Y_i] \right] = 0. \tag{2.11.5}$$

If we look at (2.11.5) we see that it has expectation zero if either the weight model is correctly specified, or $E_{X_i|Y_i}[X_i(Y_i - X_i\boldsymbol{\beta})|Y_i]$ is correctly specified. To see this, suppose the weight model is correctly specified, so that $E[R_i] = \pi_i(Y_i)$. Take the expectation of the left hand side (LHS) of (2.11.5) over $R_i|Y_i, X_i$. This gives $\sum_{i=1}^{n} X_i(Y_i - X_i\boldsymbol{\beta})$. Taking expectation of this over $Y_i|X_i$ gives zero, so the estimator is consistent for $\boldsymbol{\beta}$. In particular it is consistent regardless of whether the conditional expectation $E_{X_i|Y_i}[X_i(Y_i - X_i\boldsymbol{\beta})|Y_i]$ is correctly specified.

Now suppose the conditional expectation $E_{X_i|Y_i}[X_i(Y_i - X_i\boldsymbol{\beta})|Y_i]$ is correctly specified. Taking

expectation of the LHS of (2.11.5) over $X_i|Y_i, R_i$, gives

$$\sum_{i=1}^{n} E_{X_i|Y_i}[X_i(Y_i - X_i\boldsymbol{\beta})|Y_i]. \tag{2.11.6}$$

Now we take expectation of this over Y_i, i.e.

$$\sum_{i=1}^{n} E_{Y_i}\left[E_{X_i|Y_i}[X_i(Y_i - X_i\boldsymbol{\beta})|Y_i]\right]. \tag{2.11.7}$$

Now reverse the order of expectation, to see this is again zero because $E[Y_i] = X_i\boldsymbol{\beta}$. Thus the estimator is consistent for $\boldsymbol{\beta}$. In particular it is consistent whether or not the probability model, $\pi_i(Y_i)$ is correctly specified. This is known as the doubly robustness property (Vansteelandt *et al.*, 2010). Besides being doubly robust, estimate (2.11.5) is also fully efficient if both the weight model, $\pi_i(Y_i)$ and the conditional expectation model are correct.

This approach further extends to the logit setting though weighting general estimating equations. Generalised Estimating Equations (GEEs) are often used with longitudinal data where the correlation between variables/outcomes is unknown. Suppose that $y_i = (y_{i1}, y_{i2}, \ldots y_{ik})$ which denotes outcomes for subject i. Let y_i be fully observed for $i = 1, \ldots, f$ and partially observed/missing for $i = f + 1, \ldots, n$. We define $R_i = 1$ if y_i is fully observed and $R_i = 0$ if partially observed. Suppose we have c fully observed covariates x_i so $x_i = (x_{i1}, x_{i2}, \ldots, x_{ic})^T$. Suppose we wish to find the mean of the distribution of y_i given x_i with the assumed form $f(x_i, \beta)$. The function f is a regression with unknown parameter $\boldsymbol{\beta}$ which has r elements. If all data was fully observed (i.e. $R_i = 1$ for all i) then the GEE solution is,

$$\sum_{i=1}^{n} D_i(x_i, \boldsymbol{\beta})[y_i - f(x_i, \boldsymbol{\beta})] = 0, \tag{2.11.8}$$

where $D_i(x_i, \boldsymbol{\beta})$ is a matrix (r x k) of a known function of x_i which gives consistent estimates of $\boldsymbol{\beta}$. When data is missing the GEE process only uses fully observed records, thus equation (2.11.8) is changed to:

$$\sum_{i=1}^{c} D_i(x_i, \boldsymbol{\beta})[y_i - f(x_i, \boldsymbol{\beta})] = 0. \tag{2.11.9}$$

This gives consistent marginal regression parameter estimates providing the missing data mechanism is MCAR (Laird, 1988). If the missing data mechanism is MAR then GEEs generally yield biased regression parameter estimates (Lipsitz *et al.*, 2009). One method to eliminate this bias under MAR is inverse weighting, creating weighted GEEs (Robins *et al.*, 1995). The weighted GEE equation changes equation (2.11.9) to,

$$\sum_{i=1}^{c} w_i(\hat{\alpha}) D_i(x_i, \boldsymbol{\beta})[y_i - f(x_i, \boldsymbol{\beta})] = 0, \tag{2.11.10}$$

where $w_i(\hat{\alpha})$ is an estimated inverse probability of being a complete record. This is often found by logistic regression of R_i on x_i and any other auxiliary covariates believed to be predictors of R_i. If the weights are correctly specified then equation (2.11.10) will give a consistent estimate for β as long as the missing mechanism is MAR. Further, the approach can be extended similar to the AIPW to obtain doubly robust estimators. Extensions to the method to include information from partially observed records and data with an MNAR mechanism have been explored see Scharfstein *et al.* (1999).

We have outlined why that both complete records and inverse probability weighting lose efficiency by reducing the sample size to just the observed data with complete records. Further, weighting methods are awkward with non-monotone missingness pattern such as typically occur with observational data. We have also discussed the limitations of the EM algorithm when we have non or slow convergence of the iterative process. We thus next introduce an array of

methods which impute the data, overcoming some of the disadvantages of other methods.

2.12 Single Imputation

Single imputation imputes missing data values and then analyses the augmented data as if it were the original fully-observed data set. It thus retains all records, so has the potential to gain efficiency over a complete records analysis. The most common forms of single imputation are, imputation of the simple mean, imputation of the regression mean, stochastic regression imputation, hot deck imputation and last observation carried forward (LOCF).

Imputation of the simple mean is where a missing value of Y_i is replaced by the mean of $\boldsymbol{Y}$ from all observed records ignoring other variables. It should never be applied to categorical data as a mean category makes no sense. The method has many disadvantages. First, it ignores all other variables when creating an imputation so the association of the imputed variable with other variables is reduced. Second, by imputing all missing values to a single value, it will reduce the variability in the data. For example suppose we had a linear regression of $\boldsymbol{Y}$ on $\boldsymbol{X}$, where $\boldsymbol{Y}$ was partially observed with a MCAR missing mechanism. If we impute using the mean of $\boldsymbol{Y}$, we shrink the estimate of the slope towards 0.

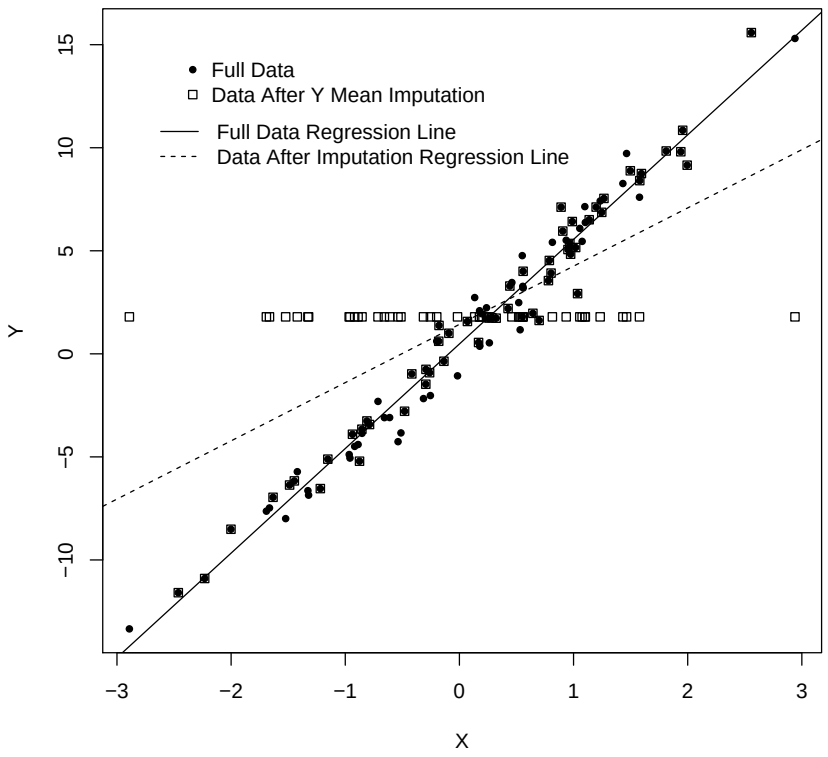

Figure 2.2: The effect of mean imputation on an MCAR outcome variable $\boldsymbol{Y}$ in a linear regression model.

Figure 2.2 shows the result of a simulation with 100 observations of $\boldsymbol{Y}$ and $\boldsymbol{X}$. $\boldsymbol{Y}$ is partially observed (45% missing) with a MCAR missing data mechanism. $\boldsymbol{Y}$ has been imputed using a simple mean imputation, shown with '□'. We can clearly see that the simple mean imputation remarkably reduces the slope of the regression line. Unless the missing Y_i value is close to the mean of $\boldsymbol{Y}$ then the simple mean imputation value is unlikely to be close to the unobserved value, thus causing bias.

Imputation of the regression mean can also lead to bias in the standard errors under MCAR. In the simplest setting, suppose we fully observe $\boldsymbol{X}$ which is linearly dependent to $\boldsymbol{Y}$ and $\boldsymbol{Y}$ is partially observed with a MCAR mechanism. The regression mean is found by regressing $\boldsymbol{Y}$ on $\boldsymbol{X}$.

$$\text{Regression Mean of } Y_i = \beta_0 + \beta_1 X_i$$

Estimates of β_0 and β_1 ($\hat{\beta}_0$, $\hat{\beta}_1$) are obtained and used to impute only the missing Y_i's with $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$. Figure 2.3 shows the imputation of the regression mean method applied to the previous simulation data in Figure 2.2.

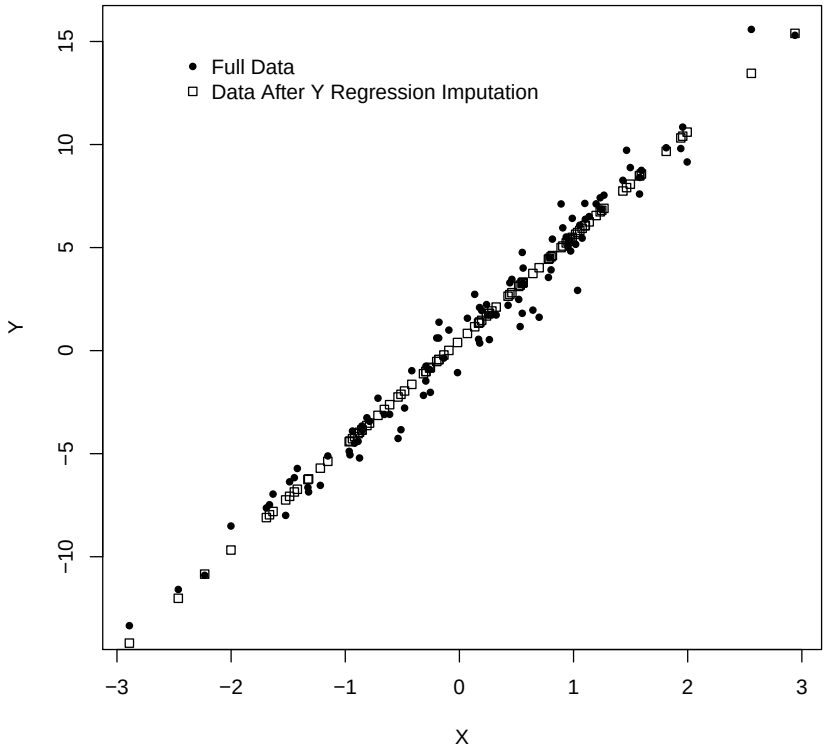

Figure 2.3: The effect of regression mean imputation on an MCAR outcome variable $\boldsymbol{Y}$.

By comparing Figures 2.2 and 2.3 it is clear that the regressed mean imputations are more plausible than the simple mean imputations. However the variability of the missing imputed data compared to the observed is greatly smaller. This will affect standard errors and subsequently p-values in analysis performed on the data. If we use the imputed data to regress $\boldsymbol{X}$ on $\boldsymbol{Y}$ this will also be biased.

Stochastic regression imputation is very similar to regression mean imputation. Stochastic regression imputation aims to correct the bias created by the reduced variability in regression mean imputation. Instead of imputing from $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ we impute from $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + S\epsilon_i$ where $\epsilon_i \sim N(0,1)$ and S^2 is the mean square error from the regression model $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$. This builds the variability of the observed data into the imputed values, giving better standard errors. This is better than regression mean imputation, but the resulting standard errors of parameter estimates are still too small in general, because we have not acknowledged the uncertainty in estimating $\hat{\beta}_0$ and $\hat{\beta}_1$. Of course we could do this, indeed this full stochastic imputation is a key step in the multiple imputation algorithm presented on page 57. However, as with all single imputation methods, the resulting imputed data set does not distinguish between the observed and imputed data. So, fitting the model of interest to this and calculating the standard errors of the parameters in the usual way gives standard error estimates that are too small.

Hot deck imputation (Andridge and Little, 2011) matches missing data records with observed data records by selected fully observed covariates. The method places records into subsets by selected covariates, and imputes missing values by the observed values within the subset. If only one match is found, the value of the observed variable is used to impute the missing value. If more than one match is found then the closest match could be taken (i.e. if weight

was missing and age could not be matched, take imputed value from the closest person by age). Another method could be to randomly select a value within the imputation subset or remove a selection covariate to possibly increase the subset size until one is found. If no match can be made then hot deck imputation cannot be preformed.

Last Observation Carried Forward (LOCF) is often used in longitudinal data. The missing value is imputed from the last observation to be seen for the particular individual record. The method is generally biased even under MCAR (Molenberghs *et al.*, 2004). The method however is popular due to its simplicity, and most commonly applied in medical studies (Pocock, 1983). The plausibility of repeat measurements stabilising is questionable; Heyting *et al.* (1993) points out that it is possible to see from the observed data whether LOCF is plausible. LOCF is sometimes claimed to preserve the Type I error (the probability of rejecting the null hypothesis incorrectly) in clinical trials when no treatment effect is present by maintaining the sample size. This is only true if there is no intervention effect and thus it is of limited use. In fact the method often seriously distorts the mean and covariance structure (Streiner, 2008) and is not founded on statistical principles, with Pocock (1996) noting "it is doubtful whether this [LOCF] actually answers a scientifically relevant question".

In summary, single imputation replaces missing observations to yield a single 'complete' dataset. This cannot provide an adequate estimate of the distribution of the missing data given the observed as there is no uncertainty built in. Fitting the model of interest to the resulting data can therefore produce substantial underestimation of the standard errors (Little and Rubin, 2002; Allison, 2001): as the method does not differentiate between imputed and original observations, it cannot account for the uncertainty in the imputations. A method which takes account of the original variability of the data and the uncertainty of the missing data imputation is

needed, and this is what multiple imputation sets out to do.

2.13 Multiple Imputation (MI)

We have noted in Section 2.12 with single imputation that the estimated standard error of the parameter is reduced, as fitting the model of interest to the imputed data does not differentiate between imputed and actual original observations.

By contrast, as MI replaces each missing value with a set of K plausible values, this generates K 'complete' datasets in each of which all missing observations have been imputed. Together these K datasets represent the distribution of missing given the observed data, which we need to average over for inference. The attraction of MI is that standard software can be used to fit the substantive model to each imputed dataset, and the rules for combining the results are simple and generic; these are known as Rubin's Rules (Rubin, 1987). In other words, the attraction of MI is that it does not require estimates of incomplete data quantities but only complete data quantities obtained by fitting the substantive model to each of the imputed datasets.

Another advantage of MI is the imputation model does not have to match the substantive model. This allows auxiliary variables to be added to improve the plausibility of the missing data mechanism (see Section 2.17 for restrictions on the imputation model choice) and also to provide information on the actual values of the missing data. The process of creating the imputation model is extremely important. One approach is to start with the substantive model and add auxiliary variables which not only predict missingness but the actual values of the

missing data (see Carpenter and Kenward (2007a) p.80 for details).

To obtain parameter estimates and inference through MI we thus need to perform three steps:

1. Imputation: Generate a set of K > 1 plausible values for $\boldsymbol{Z}^M = (\boldsymbol{Y}^M, \boldsymbol{X}^M)$ creating K complete datasets. Each imputed dataset is derived by drawing the imputed data from the posterior distribution of a Bayesian model for the missing given the observed data (imputation model).
2. Analysis: Fit our substantive model to each of the K datasets using standard methods.
3. Combination: Combine the results from the K analyses using Rubin's Rules (Rubin, 1987).

For illustration, we describe the implementation of steps 1-3 to the example (2.3.1). Suppose $\boldsymbol{X}$ is MAR given $\boldsymbol{Y}$. Estimates of $\boldsymbol{\beta}$ need to be calculated for each dataset, $k = 1, \ldots, K$, denoted $\hat{\boldsymbol{\beta}}_k$. The associated variance also needs to be calculated, call this $\boldsymbol{V}_k$. The combined estimate is then:

$$\bar{\boldsymbol{\beta}}_K = \frac{1}{K} \sum_{k=1}^{K} \hat{\boldsymbol{\beta}}_k \tag{2.13.1}$$

The combined variance results from the within and between imputation variances:

within:

$$\bar{\boldsymbol{W}}_{\boldsymbol{K}} = \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{V}_k$$

between:

$$\boldsymbol{B}_K = \frac{1}{K-1} \sum_{k=1}^{K} (\hat{\boldsymbol{\beta}}_k - \bar{\boldsymbol{\beta}}_K)^2$$

combined variance:

$$\mathbf{V_K} = \bar{\boldsymbol{W}}_K + \left(1 + \frac{1}{K}\right) \boldsymbol{B}_K.$$

Here $(1 + \frac{1}{K})$ is an adjustment for the fact that K is finite. If we assume that β is a scalar, then we can calculate the Student's t-value for testing $\beta = 0$:

$$\frac{\bar{\beta}_K}{\sqrt{V_K}}. \tag{2.13.2}$$

The value of (2.13.2) is compared to the t-distribution on v degrees of freedom where v comes from a Welch-Satterthwaite equation (Rubin and Schenker, 1986):

$$\nu = (K-1)\left(1 + \frac{\bar{W}_K}{(1+K^{-1})B_K}\right)^2$$

This works well for large datasets, a correction for smaller datasets is given by Barnard and Rubin (1999). Rubin's rules in matrix form can be found in Kenward and Carpenter (2009) page 477.

An attraction of MI is that it is relatively efficient and often only requires a few imputations to be completed (3 to 5 imputed data sets) to get a reliable result. Sometimes it is argued that it is appropriate to have more imputations. For instance Carpenter and Kenward (2007a) illustrate that more imputations can stabilise the results. The justification of a small number of imputations stems from the fact that the relative reduction in the variance is small when moving from 5 to 10 imputations. However because P-values /Z - statistics require dividing by the SE to get these accurate to 2 or 3 decimal places in order to ensure replaceability to the precision reported (i.e. so the Monte Carlo error on the p-value is small, below the

precision to which we wish to report the p-value) we need many more imputations (Carpenter and Kenward, 2007b). Another advantage of using MI is that after the data have been imputed the datasets can be used more than once, and used to fit a range of models congenial with the model of interest (as defined in section 2.17).

Rubin's combination rules are usually applied to data imputed under an MAR assumption; however, they can equally be applied to data imputed under an MNAR assumption. For inference under MNAR, we 'simply' need to impute data under an MNAR mechanism. As discussed in detail later, this can be done, for example, by using pattern-mixture models from the imputation model (e.g. Little and Yau (1996); Thijs *et al.* (2002); Carpenter and Kenward (2007b)) or alternatively approximated by re-weighting after MAR imputations, up-weighting imputations more plausible under MNAR (Carpenter *et al.*, 2007).

2.14 Markov Chain Monte Carlo Method

When the missing mechanism is a non-monotone pattern or the joint likelihood function is complex, it can be impossible to preform factorisation of the joint likelihood function into independent likelihood functions for the partially observed variables. This then prohibits imputation from independent univariate distributions. For example using the illustration monotone pattern in Figure 2.1, the independent univariate distributions for each variable are easy to create. The joint distribution is $[X_1, \ldots, X_5]$=$[X_5|X_1, \ldots, X_4][X_4|X_1, \ldots, X_3][X_3|X_1, X_2][X_2|X_1][X_1]$, because the missingness pattern is monotone we can estimate each of them validly using the observed data. If the missingness pattern is not monotone, this factorisation cannot be used. In this situation the Markov Chain Monte Carlo (MCMC) method can be applied to simulate

random draws (Gilks *et al.*, 1996; Schafer, 1999).

Assume we have p partially observed variables, $Y_1, Y_2, \ldots, Y_p$, let the missing data be represented by $\boldsymbol{Z}_m$, the observed data by $\boldsymbol{Z}_o$ and θ the parameter to be estimated. We wish to sample from the joint distribution for $\boldsymbol{Z}_m$, θ, conditional on the observed data, $\boldsymbol{Z}_o$.

$$P(\boldsymbol{Z}_m, \theta | \boldsymbol{Z}_o). \tag{2.14.1}$$

From a Bayesian perspective, the missing data are simply additional parameters. The MCMC algorithm starts with replacing $\boldsymbol{Z}_m$ with starting values, θ can then be drawn from the posterior distribution,

$$P(\theta | \boldsymbol{Z}_o, \boldsymbol{Z}_m). \tag{2.14.2}$$

Let the estimate for θ be represented by θ^t, where $t = 1, 2, \ldots$ is the iteration number. The next iterative step draws new values for $\boldsymbol{Z}_m$, call $\boldsymbol{Z}_m^{t+1}$ from the conditional predictive distribution,

$$\boldsymbol{Z}_m^{t+1} \sim P(\boldsymbol{Z}_m | \boldsymbol{Z}_o, \theta^t). \tag{2.14.3}$$

A new value of θ is then drawn,

$$\theta^{t+1} \sim P(\theta | \boldsymbol{Z}_m | \boldsymbol{Z}_o, \theta^t). \tag{2.14.4}$$

The random draws from the predicted distributions equations (2.14.3) and (2.14.4) can then be repeated in an iterative chain. The stationary distribution of the chain will be the joint distribution (2.14.1), under standard Metropolis Hastings sampling theory (Jackman, 2000). When the number of iterations t is large enough then θ^t should approximate the distribution

$P(\theta|\boldsymbol{Z}_o)$, likewise $\boldsymbol{Z}_m^t$ will approximate the distribution $P(\boldsymbol{Z}_m|\boldsymbol{Z}_o)$. The number of iterations t needs to be suitably large enough to converge to the stationary distribution to impute $\boldsymbol{Z}_m$ (Li, 1988), and also that the draws of the missing data ($\boldsymbol{Z}_m$) are independent, this is called a 'burn in' period. The speed of convergence will depend on the data, convergence time can generally be increased by mean-centering quantitative covariates. To make the multiple imputations of $\boldsymbol{Z}_m$ independent, successive iterations should not be used as they may be correlated. One method is to take a sub sample of the chain after the burn in period, for example at every d^{th} iteration where d is larger enough to stop correlation between iterations. Another method is to make multiple chains after the burn in period and use the final $\boldsymbol{Z}_m$ values.

2.15 Multiple Imputation using Chained Equations (MICE)

An alternative method to MCMC imputation is the Multiple Imputation using Chained Equations (MICE) method, this is also known as Fully Conditional Specification (FCS) as each partially observed variable is imputed from its full conditional distribution given all other variables. MICE incoropates the MCMC method but instead of explicitly modelling the joint model it instead uses univariate models for each partially observed variable in turn, conditional on all the others. If we had p partially observed variables $\boldsymbol{V}$ $(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_p)$, the MICE method would create p univariate models. The algorithm is:

a) A simple imputation is performed, for example imputation of the simple mean (see section 2.12), resulting in all missing observations being 'filled in'.

b) Imputations for $\boldsymbol{V}_1$ are reset to missing.

c) $\boldsymbol{V}_1$ is then regressed against all the other variables in the imputation model to find the estimated imputation regression coefficients, and their associated variance/covariance matrix.

d) The estimated coefficients and their variance/covariance matrix from stage c) are then used to draw parameter values for the model. These are used to generate stochastic imputations of each missing value in $\boldsymbol{V}_1$.

e) Stages b) to d) are repeated for each $\boldsymbol{V}_2, \ldots, \boldsymbol{V}_p$ in turn. This is called a 'cycle'. At the end of this stage all missing values in the data set have been replaced with estimations from regressions.

f) Stages b) to e) are then repeated, creating a number of cycles.

The choice for the regression model in stage c) should be dependent on $\boldsymbol{V}_1$'s characteristics. For example if $\boldsymbol{V}_1$ is binary then a logistic regression should be applied. Similarly if $\boldsymbol{V}_1$ is categorical, an ordinal or multinomial logistic regression is appropriate. Imputation software usually allows the user to specify which regression should be fitted, the default is often a linear regression.

The final cycle produces one imputed data set, the process is repeated a number of times to create multiple data sets which are then combined using Rubin's rules. The number of cycles is specified by the user, it is important that the value is higher enough for the algorithm to converge. The default number of cycles implemented by software is usually 10, however convergence should still be evaluated. Convergence can be tested by comparing the regression models at different cycles as discussed in He *et al.* (2010).

While MICE is flexible it has the disadvantage that the theoretical justification is not as sound as other imputation approaches (Azur *et al.*, 2011). There is no guarantee that the conditional distribution models will be consistent with the proper joint model, i.e. the conditional densities can be incompatible and thus there many be no stationary distribution to which the Gibbs sampler can converge to (Buuren *et al.*, 2006). However research by Brand (1999); Schafer and Graham (2002); Hughes *et al.* (2014) suggests that this is not a substantial issue in practice.

2.16 Multiple Multilevel Imputation (MMI)

Multilevel Models

Multilevel models (hierarchical models) are powerful tools for inference for hierarchical data. In effect, in "multilevel modelling the regression coefficients are also given a model whose parameters are estimated from the data" (Gelman, 2006). A good illustration of the importance of multilevel models can be observed in a primary school study (Bennett, 1976) which was amended to have a hierarchical structure to give unbiased inference (Aitkin *et al.*, 1981).

An example of a multilevel model, suppose we have a two level structure with i indexing units within the second level and j indexing level 1 units nested in level 2 units. Let $i = 1 \dots I_j$ and $j = 1 \dots J$ where J is the number of level 2 units and I_j is the number of level 1 units in unit j. A simple random intercept model and slope model is

$$y_{ij} = (\beta_0 + u_{0i}) + (\beta_1 + u_{1i})x_{ij} + e_{ij},$$

where

$$\begin{bmatrix} u_{0i} \\ u_{1i} \end{bmatrix} \sim N_2 \left(0, \boldsymbol{\Omega}_u = \begin{pmatrix} \sigma^2_{u0} & \\ \sigma_{u01} & \sigma^2_{u1} \end{pmatrix} \right),$$

$$e_{ij} = N(0, \sigma^2_e).$$

More generally for a p x 1 vector $\boldsymbol{x}_{ij} = (x_{1ij}, \ldots, x_{pij})^T$ which includes the constant, and a q x 1 vector $\boldsymbol{z}_{ij} = (z_{1ij}, \ldots, z_{pij})^T$, typically a subset of $\boldsymbol{x}_{ij}$. We can write:

$$\begin{aligned} y_{ij} &= \boldsymbol{x}^T_{ij}\beta_0 + \boldsymbol{z}^T_{ij}\boldsymbol{u}_i + e_{ij} \\ \boldsymbol{u}_i &\sim N_2(0, \boldsymbol{\Omega}_u) \\ e_{ij} &\sim N(0, \sigma^2) \end{aligned}$$

We can further extend this to include structure on the residuals, e_{ij}, if appropriate.

From a missing data perspective, the key point is if the x's are incomplete, we need an appropriate multilevel imputation model, otherwise we may have bias in both the point estimates and the standard errors (Carpenter *et al.*, 2012a). Such models have been described by Goldstein *et al.* (2009) and software by Carpenter *et al.* (2011a). In general, to retain congeniality between the substantive and imputation models the hierarchical structure of the models is required to be the same.

A flexible tool to apply multilevel multiple imputation is the software REALCOM. It was created at Bristol University by H Goldstein, J Rasbash, F Steele, C Charlton, H Browne and S Pollard. The software uses Markov Chain Monte Carlo (MCMC) estimations and can handle categorical data as both predictors and responses (Carpenter *et al.*, 2011a). It has the

capability to constrain elements of the imputation model and apply priors for both covariates means and variances.

2.17 Congeniality

As previously noted, multiple imputation does not require the imputation model and the substantive model to match. This advantage allows the imputation model to include auxiliary variables, not in the substantive model to increase the plausibility of the missing data mechanism assumption and improve the precision of the imputations. However, after adding such auxiliary variables the imputation and substantive models are no longer strictly congenial in the sense described by Meng (1994) outlined below, so Rubin's combinations rules are not guaranteed to hold (Robins and Wang, 2000).

To describe congeniality, we consider the example of (Carpenter and Kenward, 2013, p46), assuming we have obtained K imputed datasets under MAR (using $\boldsymbol{Y}_{obs}$ to predict $\boldsymbol{Y}_{Miss}$), to each of which we have applied the substantive model to estimate our parameter of interest β, and then combined the results for final inferences using Rubin's rules.

Now suppose that, as (Carpenter and Kenward, 2013, p46) say "Separate to this, assume that there exists a full Bayesian procedure for obtaining the posterior of $\boldsymbol{\beta}$, such that if in addition we were to use this full Bayesian procedure to impute the missing data, then the imputation distribution would be the same as the predictive distribution in the previous paragraph".

Models which do not follow these guidelines are described as uncongenial (Meng, 1994). An uncongenial imputation model typically arises when the imputation model contains more variables, i.e. auxiliary variables, than the substantive model. However, it also applies when the imputation model contains fewer variables than the substantive model. In the former setting we say the imputation model is *richer* than the substantive model if it contains nested within it a congenial imputation model (Carpenter and Kenward, 2013, p64-70). In the latter case we say the imputation model is *poorer* than the substantive model, e.g. the structure and/or variables in the substantive model is/are missing from the imputation model. The latter setting should be avoided as the validity of Rubin's variance estimator may not hold (Robins and Wang, 2000). In the former setting Rubin (1996); Meng (1994) and Schafer (2003) argue that uncongeniality is not a practically important issue; indeed it may even be beneficial as adding information to the imputation model which is not used in the substantive model will increase the efficiency of parameter estimates. The effect of *richer* uncongenial models is that Rubin's rules over estimates the sampling variability. The extent of this is dependent on the setting (Meng, 1994), in most applications it is negligible (Carpenter and Kenward, 2013).

We now move on to describing methods which are used for data with a missing not at random (MNAR) mechanism and discuss how they can be used for sensitivity analysis.

2.18 Missing Not At Random Missing Data

Analysing data under a Missing Not At Random (MNAR) mechanism is more complex. Broadly there are three types of model which can be applied, shared parameter (Little., 1995), selection and pattern mixture models. The shared parameter model uses latent variables (for

example random effects) to account for dependence between the reason for being missing and the missing observation value (Woolson and Clarke, 1984). Pattern mixture models impute the missing data with an assumed missing data distribution which may be different for each missingness pattern. Selection models include an explicit model for the probability of data being missing and jointly fit this with the substantive model. All three modelling methods can be applied with maximum likelihood, WinBUGS or IPW, but here we describe the pattern mixture and selection modelling approaches using the MI as we believe it is simpler and more flexible across a range of practical applications not least due to its relative computational simplicity.

2.19 Shared Parameter Modelling

Shared parameter modelling, makes the missing data mechanism and the observed data independent by conditioning on specified shared parameters. The shared parameters can be thought of as playing the role of confounders between the missing the observed data and the missing indicator. We outline this idea with our illustrative model example (2.3.1). Shared parameter modelling assumes Y_i is independent of R_i given a group of parameters ξ_i. Let $\boldsymbol{\delta}$ represent the parameters of the distribution of $\boldsymbol{R}$ thus the joint distribution is,

$$f(Y_i, R_i|X_i, \boldsymbol{\beta}, \boldsymbol{\delta}) = \int f\left(Y_i|\xi_i, X_i, \boldsymbol{\beta}\right) f\left(R_i|\xi_i, X_i\boldsymbol{\delta}\right) f(\xi_i) d\xi_i. \tag{2.19.1}$$

The model can be fitted using a package like WinBUGS. This method is ideal for data with an MNAR mechanism, it can also be used as a sensitivity analysis tool. However, the results can be sensitive to the choice of distribution for the random effects, about which the data has little

information. Further, it is relatively hard to explain the selection mechanism, particularly to non-statisticians, as it involves an adjusted dependence on latent variables. We therefore do not pursue this further here.

2.20 Pattern Mixture Modelling

Pattern mixture models specify a separate distribution for the missing data for each pattern. These are then averaged over the pattern frequency to obtain the marginal distribution. We outline this approach using the illustrative model example (2.3.1). Under the assumption of MAR we assume that the distribution of Y_i (partially observed) given X_i (fully observed) are the same whether or not Y_i is observed:

$$P(Y_i|X_i, R_i = 0) = P(Y_i|X_i, R_i = 1).$$

This relationship is exploited by multiple imputation under MAR by estimating the distribution of $P(Y_i|X_i)$ using the observed data distribution and then using this to impute the missing data. In MNAR pattern mixture modelling we change this relationship:

$$P(Y_i|X_i, R_i = 0) \neq P(Y_i|X_i, R_i = 1). \tag{2.20.1}$$

Typically, we estimate $P(Y_i|X_i)$ using the observed data, then change this before imputing the missing data to reflect our MNAR assumption. Having imputed the missing data, we fit the substantive model to each imputed dataset and apply Rubin's combination rules. The approach is called pattern mixture as it results in imputed data with a mixture of potentially different

imputation solutions across the different missing data patterns. As we elaborate in chapter 3, this approach is easy to explain and can be presented graphically to scientific collaborators. Pattern mixture modelling can be extended to stratified data i.e. treatment arms. It has also been applied to longitudinal data (Thijs *et al.*, 2002), which often suffers from missing values due to attrition.

2.21 Selection Modelling

Selection modelling jointly fits the model for the reason data is missing (the missingness mechanism) and the substantive model. It is an intuitive approach as it directly models the distribution of interest. We begin by discussing the Heckman selection model and then introduce another method which can be easily adapted to be approximated using a weighting method proposed by Carpenter *et al.* (2007).

The Heckman selection model (Heckman, 1976) consists of two stages, the first stage models the probability of an observation being observed with a probit model. Using the illustrative example (2.3.1) assume $\boldsymbol{Y}$ is partially observed and $\boldsymbol{X}$ is fully observed. Suppose the substantive model is:

$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + \varepsilon_i \tag{2.21.1}$$

We model the probability of being observed with the following Probit model,

$$\text{Prob}(R_i = 1 | \boldsymbol{Z}_i) = \Phi(\boldsymbol{Z}_i \boldsymbol{\alpha}), \tag{2.21.2}$$

where $\boldsymbol{\alpha}$ is a vector of unknown parameters, Φ is the cumulative distribution function of the standard normal distribution and $\boldsymbol{Z}$ represents explanatory variables which are assumed to predict $\boldsymbol{Y}$ being missing ($\boldsymbol{Z}$ may include a subset or all of the explanatory variables in $\boldsymbol{X}$). This estimate is then used to generate an inverse Mills ratio (Heckman, 1976),

$$E[Y|R=1] = \frac{\phi(\boldsymbol{Z}_i\boldsymbol{\alpha})}{1-\Phi(\boldsymbol{Z}_i\boldsymbol{\alpha})} = \delta_i. \tag{2.21.3}$$

The second stage of the Heckman selection model includes the inverse Mills ratio as an additional explanatory variable in the substantive model (2.21.1):

$$Y_i = \boldsymbol{X}_i^T\boldsymbol{\beta} + \gamma\delta_i + \varepsilon_i. \tag{2.21.4}$$

This method is often used when the missing data mechanism is MNAR. The Heckman selection model does not perform well when sample sizes are small, the method is also inconsistent if $\boldsymbol{X}$ and $\boldsymbol{Z}$ overlap significantly. Another problem is the covariance matrix created by the regression model in the final stage will be inconsistent, however estimates can be obtained from bootstrapping (Damien *et al.*, 2013, p.525).

Another possible selection model uses a Logit instead of the a Probit link. Thus the selection model becomes

$$\text{Logit } \ P(R_i=1) = \beta_0 + \beta_1 x_i + \delta y_i.$$

If $\delta = 0$ then the data are MAR and the substantive and observation probability models may be fitted separately. If $\delta \neq 0$ then the data are MNAR and observation probability models must be fitted jointly. A useful software to do this is using a package like WinBUGS. The value of δ may be difficult to find; elicitation of expert knowledge or exploration of different

values can be used to assess the sensitivity to the MAR assumption (Carpenter *et al.*, 2002). The difficulty is that δ represents the adjusted dependence of observing $\boldsymbol{Y}$ on the (potentially missing) value of $\boldsymbol{Y}$. This can be hard to explain to non-statistical experts when trying to elicit prior information as we discuss in Chapter 3.

As an alternative to fitting the full joint model in WinBUGS, Carpenter *et al.* (2007) propose a re-weighting method which approximates the selection model approach. The method uses multiple imputation to impute under the assumption of MAR, then as usual, the substantive model is fitted to each imputed dataset. We then use a re-weighted version of Rubin's rules which up-weights estimates that are more plausible under the MNAR mechanism. The approach can only be applied to local sensitivity analysis because we require the MNAR estimate to have non-zero probability under the MAR distribution. If the distributions are too far away from each other then no re-weighting could recover the MNAR estimate. To find the MNAR estimate we use:

$$\frac{\sum_{m=1}^{M} w_m \hat{\beta}_m}{\sum_{m=1}^{M} w_m}$$

Where $w_m \propto exp(-\delta \sum_{i=1}^{N} y_i)$, $\hat{\beta}_m$ represents a parameter estimate for each imputation, M is the number of imputations and δ the selection parameter which relates the logit of the probability that the variable is missing to the underlying value, after adjusting for dependence on the observed data; reflecting the difference between MAR and MNAR. This approach can be applied to both partially observed responses and or covariates; in this case we need a selection parameter for each partially observed variable, which is more difficult. A disadvantage of this approach is that it is only approximate, however it is much quicker and less prone to computational errors than fitting a selection model (Carpenter *et al.*, 2002). In Chapter 5 we develop this weighting method for small data sets.

2.22 Sensitivity Analysis Types

There are many forms of sensitivity analysis (Saltelli and Annoni, 2010) which we could have explored. Within this thesis we use two distinct classes. The first class which we call congenial sensitivity analysis, is not congenial in the sense of imputation as descried in Section 2.17. Instead, we mean that the modelling assumption for the substantive model hold under both the MAR and MNAR distributions considered. In the second class which we call uncongenial sensitivity analysis, the assumption og the substantive model are strictly violated by the data from the MNAR distribution, however we retain the substantive model, as we are interested in exploring the robustness of inferences from this model to the MNAR data. Both classes postulates an MNAR mechanism which represents the departure from the MAR assumption. The classes differ by the way the MNAR mechanism is chosen: in the congenial sensitivity analysis the mechanism is chosen to retain assumptions of the substantive model i.e. linearity even if we believe the missing data is not linear. In the second setting we ignore the assumption of the substantive model. This thesis looks mainly at congenial sensitivity but in Chapter 4 we apply uncongenial sensitivity analysis to look at extreme settings.

2.23 Summary

There is a wide range of literature on methodology for missing data, which is applied to an ever increasing range of data analyses. While we have demonstrated there is no definitive correct method to use we emphasis the flexibility and ease of use of MI approaches characteristics that are retained as we move to sensitivity analysis. Researchers are also increasingly using MI under

the assumption of MAR without exploring the sensitivity of inference to MNAR (Carpenter *et al.*, 2012b), we thus use this thesis to demonstrate and develop sensitivity methods, to show their importance and practical utility.

3

Estimating Cancer Survival from Registry Data: A practical framework for understanding the impact of missing data on conclusions

3.1 Introduction

In the statistical literature there are broadly two approaches to analyse data under the MNAR assumption: a selection model (e.g Carpenter *et al.*, 2007, 2011b) or pattern-mixture model (Little., 1995) (see Sections 2.21 and 2.20). A selection model contains a component which defines the probability of observations being missing and links this to the potentially missing variables. A pattern mixture approach creates a difference between the observed and missing distributions by specifying a model for each distinct pattern. Having specified this, estimations and inferences can be carried out using multiple imputation (MI). When applying pattern mixture modelling with MI, one approach is to estimate the Bayesian predictive distribution for imputation under MAR, but before imputing alter it by a random draw from a prior distribution. This prior represents the difference between MAR and MNAR in this pattern. The result is imputed data with a mixture of potentially different imputations across the different missing data patterns, hence the name pattern mixture. Carpenter and Kenward (2013) found that this approach is more readily understood by nonstatistically trained experts than others as it can be presented graphically, however it cannot be directly implemented using standard statistical software, which is possibly why MI is often not accompanied by pattern mixture sensitivity analysis. We searched publications for applications of either selection or pattern mixture modelling in epidemiology research but found no evidence of either approach being applied. The aim of this chapter is to perform contextually appropriate sensitivity analysis for a colorectal cancer epidemiology study using the pattern mixture approach, and eliciting prior information from experts. In so doing we hope to demonstrate the practical utility of this approach, and encourage its use in the future.

Thus we develop and apply a pattern mixture sensitivity analysis approach to our motivating colorectal cancer data, to explore in an accessible way the sensitivity of inferences to the MAR assumption made in the multiple imputation of a key variable. We achieve this by elicitation of information from experts through a questionnaire which quantifies what they believe the distribution of the missing data to be. This information should only come from experts which work/have good knowledge of the colorectal cancer field (Kadane and Wolfson, 1998; O'Hagan, 1998). The questionnaire information is used to create a Bayesian prior representing the missing distribution which we then use to impute the data under an MNAR assumption. This quantifies our uncertainty about the departure/sensitivity of the MAR assumption (Scharfstein *et al.*, 2003). To implement this approach, we wrote our own R code to perform the missing not at random multiple imputation using the elicited prior information. However, we begin by describing the data.

3.2 Colorectal Cancer Data

The motivating colorectal cancer data we use is from the National Cancer Data Repository (NCDR, 2009) and the Hospital Episode Statistics (HES, 2009) database. The data were collected to "assess the variation in risk adjusted 30 day postoperative mortality for patients with colorectal cancer between hospital trusts within the English NHS"(Morris *et al.*, 2011). Colorectal cancer is the fourth most common cancer in the UK, with 40,695 people being diagnosed in 2010 and with 15,659 bowel cancer deaths in 2011 (CR, 2013). To create the dataset to assess the risk, individuals who underwent a major resection for their diagnosed primary colorectal cancer in any NHS English hospital between January 1998 and December 2006, were identified in the Hospital Episode Statistics (HES) dataset and linked to the National Cancer

Data Repository (NCDR) dataset to extract more information. The final dataset contains information on patients demographics, Dukes' stage and 30 day postoperative mortality outcome. Dukes' cancer tumour stage records how far the cancer has spread with four stages from A to D. Stage A is the least severe meaning the cancer is only in the innermost lining of the bowel or slightly into the muscle, while stage D is the most severe meaning the cancer has spread to other areas of the body (NCIN, 2014).

The data consisted of 160,920 patients, of whom 10,704 (6.7%) died within 30 days after surgery. Data were complete for approximately 85% of the patients, thus we had 136,040 complete records. Missing observations were observed in three variables, Dukes' stage had 15% of its observations missing, quintile index for multiple deprivation (IMD) had 0.25% missing and the emergency admissions indicator (EAI) had 0.05% missing, see Appendix B for full list of variables and information.

The aim of the study was to investigate risk-adjusted surgical outcome for patients with colorectal cancer at a population level. The *substantive* model applied was a multilevel binary logistic regression model,

$$\begin{aligned} Logit\ P(MORT_{ijk}=1) &= \beta_0 + \beta_1 Age_{ijk} + \beta_2 YOD_{ijk} + \beta_3 Sex_{ijk} \\ &+ \beta_4 OT_{ijk} + \beta_5 Duke(B)_{ijk} + \beta_6 Duke(C)_{ijk} + \beta_7 Duke(D)_{ijk} \\ &+ \beta_8 IMD(2)_{ijk} + \beta_9 IMD(3)_{ijk} + \beta_{10} IMD(4)_{ijk} + \beta_{11} IMD(5)_{ijk} \\ &+ \beta_{12} CS(2)_{ijk} + \beta_{13} CS(3)_{ijk} + \beta_{14} CCS(1)_{ijk} + \beta_{15} CCS(2)_{ijk} \\ &+ \beta_{16} CCS(3+)_{ijk} + v_k + u_{jk}, \end{aligned} \tag{3.2.1}$$

where $MORT$ is postoperative mortality within 30 days, Age is age at diagnosis, YOD is Year Of Diagnosis, Sex (female), OT is Operation Type (emergency), Duke(B-D) are the corresponding Dukes' stages, IMD(2-5) are the corresponding Index for Multiple Deprivation, CS(2,3) are Cancer Site (Rectosigmoid and Rectum) and CCS(1-3+) are the corresponding Charlson Comorbidity Score. In the model i = patient ($i = 1, \ldots 160,920$), j = hospital trust ($j = 1, \ldots 151$), k = cancer network ($k = 1, \ldots 28$), $v_k \sim N(0, \sigma_v^2)$ and $u_{jk} \sim N(0, \Sigma_u^2)$. More details on the variables can be found in Appendix B.

3.3 Complete Records and Imputed Data Under MAR Analysis

We began by estimating the missing values under the assumption of MAR through multiple imputation by fully conditional specification (FCS) using Stata's user written program *ice* (StataCorp, 2011a; Carlin *et al.*, 2008). The program uses chained univariate imputation models (MICE) for each variable in turn, as described in Section 2.15 and Appendix B.2.

The imputation model was chosen to match the model used in (Morris *et al.*, 2011), to check reproducibility of inferences. The model was not hierarchical, so may not be congenial with the *substantive* model (see Chapter 2 Section 2.17). A likely consequence of this is that the hierarchical structure will not be correctly present in the imputed data, resulting in possible underestimation of variance components in the analysis. In Chapter 4 we discuss how to impute with a multilevel structure to retain congeniality with the *substantive* model.

In the imputation process we generated 10 imputed datasets under the assumption of MAR (with 10 cycles) based on $MORT$, sex, Median Annual Workload of the Trust ($MAWT$),

Dukes' stage, IMD, age at diagnosis, YOD, YO, CCS, OT, EAI, CS, hospital trust, and cancer registry. The imputation model contains three auxiliary variables not in the *substantive* model (MAWT, YO and EAI). The default for the *ice* program assumes that missing observations are MAR, thus the conditional distributions of partially observed variables given fully observed variables are the same, whether data are missing or observed. We applied the *substantive* model to the complete records and the MAR imputed data and compared the results, see Table 3.1.

	N=136,040 Complete Records Analysis			N=160,920 Multiple Imputation MAR		
Characteristic	AOR	(95% CI)	p Value	AOR	(95% CI)	p Value
Age at diagnosis (per 10 years)	2.11	(2.05-2.17)	<0.001	2.12	(2.07-2.17)	<0.001
Year of diagnosis (per advancing year)	0.99	(0.98-1.00)	0.005	0.97	(0.97-0.98)	<0.001
Sex						
Female	1.00			1.00		
Male	1.23	(1.18-1.29)	<0.001	1.21	(1.16-1.26)	<0.001
Operation						
Elective	1.00			1.00		
Emergency	2.61	(2.46-2.77)	<0.001	2.67	(2.53-2.82)	<0.001
Dukes' stage at diagnosis						
A	1.00			1.00		
B	1.28	(1.17-1.41)	<0.001	1.24	(1.14-1.35)	<0.001
C	1.53	(1.39-1.68)	<0.001	1.54	(1.42-1.68)	<0.001
D	2.64	(2.37-2.93)	<0.001	2.48	(2.25-2.73)	<0.001
IMD income category						
Most affluent	1.00			1.00		
2	1.04	(0.96-1.12)	0.321	1.03	(0.96-1.10)	0.429
3	1.13	(1.05-1.22)	0.002	1.11	(1.04-1.19)	<0.001
4	1.24	(1.15-1.34)	<0.001	1.21	(1.13-1.30)	<0.001
Most deprived	1.37	(1.26-1.49)	<0.001	1.32	(1.23-1.42)	<0.001
Cancer site						
Colon	1.00			1.00		
Rectosigmoid	0.83	(0.76-0.91)	<0.001	0.88	(0.82-0.96)	0.003
Rectum	0.92	(0.86-0.98)	0.008	0.94	(0.89-0.99)	0.018
Charlson comorbidity score						
0	1.00			1.00		
1	2.12	(1.99-2.26)	<0.001	2.05	(1.94-2.17)	<0.001
2	2.46	(2.26-2.68)	<0.001	2.43	(2.25-2.62)	<0.001
≥ 3	4.51	(4.06-5.01)	<0.001	4.39	(3.99-4.83)	<0.001

Table 3.1: Adjusted odds ratios (AOR) and associated 95% confidence intervals for death within 30 days of surgery for the complete records and imputed data with a 'Missing At Random' assumption based on 10 imputed datasets.

In Table 3.1 we can see that all covariates (apart from IMD category 2) are significantly associated with death within 30 days of surgery. We begin by focusing on the complete records. The odds of death were significantly higher for males (AOR=1.23), emergency operations (AOR=2.61) and age at diagnosis per 10 year change (AOR=2.11). In Dukes' stage we can see an increase in risk of death as the severity of the cancer stage increases (AOR's 1.28, 1.53 and 2.64). This increase is also seen in the IMD income with more deprived patients at greater risk and when the Charlson comorbidity score increases.

Analyses undertaken on the MAR imputed dataset showed similar inferences to the CR data subset. The largest absolute differences in AOR's are for Dukes' stage D (6.1% decreased from 2.64 to 2.48), cancer site rectosigmoid (6.0% increase from 0.83 to 0.88), IMD most deprived (3.6% decrease from 1.37 to 1.32) and Charlson comorbidity score of 1 (3.3% decrease from 2.12 to 2.05). The significance at the 5% level remains the same. We conclude the imputed data has not changed the inferences greatly, but has improved the power of the analyses as we have more observations, as can be seen by the narrowing of the confidence intervals. However this does not confirm that the MAR assumption in the multiple imputation is valid, we need to apply sensitivity analysis to the data to see how robust the inferences are to this assumption.

3.4 Pattern Mixture Approach

To apply sensitivity analysis we change the imputation distribution to represent an MNAR mechanism. We focused the sensitivity analysis on Dukes' stage, as the missing information in IMD and EAI are negligible in comparison. Our approach is as follows:

1. Find predictors for Dukes' stage being missing which are also strong predictors of Duke's stage.

2. Given each predictor from the previous step, we calculate the probability of being in each stage under the MAR assumption, using the data imputed under MAR.

3. The above probabilities are then given to experts in a questionnaire. We elicit information from the experts by saying, "given these probabilities in the observed data what do you think are the probabilities in the missing data?".

4. The estimated probabilities from the questionnaire are used to estimate the parameters of a Dirichlet distribution. Draws from the distribution are then used to impute under the MNAR assumption.

5. The *substantive* model is applied to the MNAR imputed data, inferences are compared to the MAR inferences to see how robust they are.

Elicitation of information to create a prior is not a new concept. It has been extensively used in other contexts to form Bayesian statistics, the process is explored and discussed in depth within White *et al.* (2007). We however could not find published elicitation used in sensitivity analysis for the MAR assumption within cancer epidemiology research, thus this is a novel approach in this setting which we aim to publish. We use probabilities in our elicitation instead of odds ratios after discussions with clinicians suggested that probabilities would be easier to communicate and hence yield more, and better informed responses.

We began with step one by finding possible predictors for Dukes' stage being missing. To do this we created a binary indicator for Dukes' stage being missing and used it as an outcome

in a logistic regression model, regressing other variables against it. The idea is then to use a subset of the statistically significant predictors to form a questionnaire to elicit information on the missing data distribution. Because this needs to be simple and accessible, we do not consider trust, network, year of diagnosis, year of operation and medium annual workload for a trust as covariates in the model. Note that the purpose is simply to identify key variables from the substantive model that are associated with the probability of data being missing. Therefore, whether or not the relationship of the probability of seeing stage with age, for example, is linear, is of secondary interest here.

Covariates in the logistic regression were removed in the usual backward stepwise manner (using a 1% level) with categorical variables with mixed significance being tested with a joint parameter test. The final model had 3 predictive explanatory variables, age at diagnosis, 30 day mortality and tumour site, see Table 3.2.

Covariates	AOR (CI)	P value
Age at Diagnosis (per 10 years)	0.93 (0.92-0.94)	<0.001
30 Day Mortality	1.90 (1.81-1.99)	<0.001
Tumour Site		
Colon	1.00	
Rectosigmoid	1.06 (1.01-1.12)	0.024
Rectum	1.51 (1.46-1.66)	<0.001

Table 3.2: Adjusted Odds Ratios (AOR) and their respective confidence intervals (CI) for predicting Dukes' stage being missing.

We can see in Table 3.2 that being dead 30 days after surgery increases (AOR=1.90) the chance of Dukes' stage being missing. Age at diagnosis is a protective effect (AOR=0.93),

thus younger patients are more likely to have Dukes' stage missing. In tumour site in respect to colon, only the rectum tumour site location predicts Dukes' stage being missing at the 1% level. Bearing in mind the discussion in the previous paragraph, we move forward with the two most predictive covariates for missing Dukes' stage which are age at diagnosis and 30 day postoperative mortality.

We now seek to elicit information on how the missing data differs from the observed data based on age at diagnosis and 30 day postoperative mortality. This will then be used to create a prior distribution for the missing data, which we will draw from to impute the missing values. Before doing this, however, we carry out a simulation study to review the performance of Rubin's rules in this setting.

3.5 Simulation Study

We build a simulation to look at the variability produced by Rubin's rules when data is imputed using an MAR and using a pattern mixture MNAR approach. We also wish to investigate if the confidence interval coverage is good and if the coefficients produced are biased. Our substantive model is:

$$Logit\ P\left(\text{Mortality} = 1\right) = \beta_0 + \beta_1 Age + \beta_2 Stage. \tag{3.5.1}$$

For simplicity we simulate observations of Age $\sim N(0,1)$ and Mortality $\sim B(0.65)$. A binary variable ($Stage$) representing a simplified Dukes' Stage was created from two distributions. As in the cancer data we create Age and $Mortality$ as fully observed and $Stage$ as partially observed. The first $Stage$ distribution represented the observed stage data ($stageob$), created

as:

$$Logit\ P(Stage = 1) = \delta_0 + \delta_1 Age + \delta_2 Mortality, \tag{3.5.2}$$

and the second represented the missing stage data ($stagemiss$), created as:

$$Logit\ P(Stage = 1) = \delta_0 + (\delta_1 + \alpha_1)Age + Mortality(\delta_2 + \alpha_2), \tag{3.5.3}$$

where $\delta_0 = 0.1$, $\delta_1 = 0.2$, $\delta_2 = 0.3$, $\alpha_1 = 1$ and $\alpha_2 = 1$. We then created a final stage variable using 60% of *stageob* and 40% of $stagemiss$. We began by apply the *substantive* model (3.5.1), to a large data set of 5 million observations drawn from the data generating mechanism. This is given in the 'true' parameter estimate in (3.5.1) which we use in the calculation of the empirical bias, variances and confidence interval coverage.

We perform the following steps to generate data and estimate parameters from not only the pattern mixture approach but also MAR MI and a complete records analysis.

- Simulated a new data set with 10,000 observations (called the $full\ data$).
- Fit the *substantive* model to the $full\ data$.
- Removed all records with stage drawn from $stagemiss$ to represent the *complete records data*.
- Fit the *substantive* model to the *complete record data*.
- Take the $full\ data$, make stage missing if stage was drawn from $stagemiss$ (called the *partially observed data*).

- Imputed the missing data in the *partially observed data* using multiple imputation under the assumption of MAR generating 5 imputed data sets (called the MAR *datasets*).
- Fit the *substantive* model to each MAR *dataset* and combine inferences with Rubin's rules.
- Imputed the missing data in the *partially observed data* using multiple imputation under an MNAR distribution generating 5 imputed datasets using draws from (3.5.3), the true missing data distribution (called the $MNAR$ *datasets*).
- Fit the *substantive* model to each $MNAR$ *dataset* and combine inferences with Rubin's rules.

The above steps were repeated 10,000 times and the averages of the inferences were calculated. We begin by looking to see if any of the models produced biased coefficient estimates:

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
True Value	0.32	-0.05	0.52
Full Data	0.31	-0.05	0.52
Complete Record	0.48	-0.01	0.25
MAR	0.48	-0.01	0.25
MNAR	0.32	-0.05	0.52

Table 3.3: Coefficients from the *substantive* model, shown for the true distribution values, full data, complete records, data imputed under MAR and data imputed under MNAR.

We can see in Table 3.3 that the complete records are biased for all coefficients. As the MAR data is only produced using the observed data, the inferences are also equally biased. The MNAR imputed coefficients are clearly less bias and are approximately the same as the full data coefficients. We are thus satisfied that the pattern mixture approach can produce unbiased

coefficients if the analysis about the MNAR mechanism is correct. We next investigate the confidence interval coverage which can be seen in Table 3.4.

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
Full Data	94.95	95.05	95.16
Complete Record	74.99	92.04	65.65
MAR	72.64	90.63	69.81
MNAR	96.59	95.26	97.73

Table 3.4: Confidence interval percentage coverage from the *substantive* model, shown for the full data, complete records, data imputed under MAR and data imputed under MNAR.

Table 3.4 shows the coverage for the full data is approximately 95% as expected. Coverage for the complete records and MAR imputation is badly biased as expected. Coverage is very slightly too large for the MNAR imputation. This is in line with theory, because the extra information given into the MNAR analysis makes the imputation model uncongenial. However the over coverage is practically negligible, and is arguably appropriate in a sensitivity analysis setting. This suggests that the pattern mixture approach under MNAR gives good confidence interval coverage. We lastly explore the empirical and theoretical variances, seen in Table 3.5.

	Empirical (Theoretical) for $\hat{\beta}_0$	Empirical (Theoretical) for $\hat{\beta}_1$	Empirical (Theoretical) for $\hat{\beta}_2$
Full Data	0.0105 (0.0105)	0.0047 (0.0047)	0.0188 (0.0191)
Complete Record	0.0442 (0.0163)	0.0095 (0.0075)	0.1033 (0.0301)
MAR	0.0417 (0.0141)	0.0063 (0.0045)	0.1056 (0.0326)
MNAR	0.0107 (0.0105)	0.0047 (0.0047)	0.0191 (0.0191)

Table 3.5: Empirical and theoretical variance from the *substantive* model, shown for the full data, complete records, data imputed under MAR and data imputed under MNAR.

We can see in Table 3.5 that after multiple imputation under MNAR using the pattern mixture approach, the empirical variance is similar to the theoretical variance. As expected the empirical is slightly larger than the theoretical consistent with the coverage results in Table 3.4. Compared to the full data, the MNAR variances are larger, particulary for $\hat{\beta}_2$, the coefficient of the partially observed variable. This reflects the inclusion of extra information.

We conclude that the pattern mixture approach works well and does not produce biased inferences. We thus move forward with the elicitation of the predictive covariates of Dukes' stage.

3.6 Prior Elicitation

As mentioned previously, to simplify the elicitation process we continue with the two most predictive covariates for missing Dukes' stage, age at diagnosis which we dichotomised (AGE) as 0 when the patient is less than or equal to 70 years old otherwise 1, and 30 day post-operative mortality ($MORT$). AGE was dichotomised as it would be extremely difficult to elicit information by year. The final model with AGE and $MORT$ as explanatory variables showed they are strong predictors of Dukes' Stage being missing (AOR 0.86 95% CI: 0.84 to 0.89 and AOR 1.80 95% CI: 1.71 to 1.89 respectively). We checked to see if AGE and $MORT$ are good predictors of Dukes' stage using a multinomial logistic regression. Both AGE and $MORT$ were strongly associated with the observed values of Dukes' stage (p <0.001). Thus not only are AGE and $MORT$ good predictive variables for missing Dukes' stage but they are also strongly associated with the underlying, unseen values. We moved on to look at the proportion of missing Dukes' stage data by AGE and $MORT$ in Table 3.6.

	30 Day Postoperative Mortality	
Age at Diagnosis	Alive	Deceased
Less than or equal to 70	0.16(11,453)	0.25(548)
Greater than 70	0.14(10,539)	0.22(1,894)

Proportions rounded to 2dp

Table 3.6: Proportion(frequency) of missing observations in Dukes' stage by dichotomised age at diagnosis and 30 day postoperative mortality.

Table 3.6 shows that the majority of the missing observations occurred when patients are dead 30 days after surgery and less than or equal to 70 years old. As AGE and $MORT$ are both binary variables, we label the cells in Table 3.6 by r for simplicity. Let $r = 1$ if $AGE = 0$ and $MORT = 0$, $r = 2$ if $AGE = 0$ and $MORT = 1$, $r = 3$ if $AGE = 1$ and $MORT = 0$ and $r = 4$ if $AGE = 1$ and $MORT = 1$.

To re-impute the data with an MNAR assumption we will need to specify the probability of being in each Dukes' stage given r. We begin by fitting a multinomial logistic model (an ordinal logistic regression was not appropriate due to violation of the proportional odds assumption) with $MORT$ and AGE as covariates.

$$\text{Log}\left(\frac{P(\text{Dukes' Stage}\, j)}{P(\text{Dukes' Stage}\, 1)}\right) = \alpha_j + \boldsymbol{\beta}_{j1} MORT + \boldsymbol{\beta}_{j2} AGE \qquad j = 2, 3, 4.$$

Hence:

$$\hat{P}_{ij} = P(\text{Dukes' Stage}_i = j) = \frac{\left(exp(\alpha_j + \sum_{k=1}^{2} \boldsymbol{\beta}_{jk} X_{ik})\right)^{(1-\delta_{1j})}}{1 + \sum_{h=2}^{4} exp(\alpha_h + \sum_{k=1}^{2} \boldsymbol{\beta}_{hk} X_{ik})}$$

Where $i = 1, \ldots, 160,920$ representing each patient, $j = 1, \ldots, 4$ denotes the number of categories of Dukes' stage, X is the data and δ_{1j} is 1 if $j = 1$ otherwise 0. Here $\boldsymbol{\beta}$ and α are

unknown parameters that can be estimated. Next we can calculate the probability of being in a stage given r as $\hat{P}_{rj}$:

$$\hat{P}_{rj} = E[\hat{P}_{ij}|r]$$

The probabilities $\hat{P}_{rj}$ were used to elicit priors $\hat{\pi}_{rj}$. To elicit information from experts about the departure from MAR, a natural approach is to seek prior information about the true probability for the missing observations for each Dukes' stage given r. To do this we created a questionnaire in Microsoft Excel so that it could be electronically distributed. The questionnaire is shown in Appendix B.3. The questionnaire was designed to be easy to complete. It is spilt into the 4 r categories, within each, a table of probabilities $\hat{P}_{rj}$ created from the MAR imputed data are shown. Alongside each table, a graph shows the probabilities. The user is invited to fill in their predicted probabilities for each r, the spreadsheet checks that the four stages within each r add to one (cell turns green) and plots the predicted probabilities on the same graph as the data probabilities. The questionnaires were sent out electronically and were accompanied by information on the data and sensitivity analysis concept (see Appendix B.3 to view supporting questionnaire literature).

The questionnaire was completed by 6 people, 4 of whom had seen the study data as they had worked on the Morris *et al.* (2011) publication. The other two experts were a cancer surgeon and a lecturer in cancer statistics. The individual probability responses from the questionnaire and the $\hat{P}_{rj}$ probabilities are graphically shown in Figure 3.1 (see Appendix B.4 for the numerical individual responses).

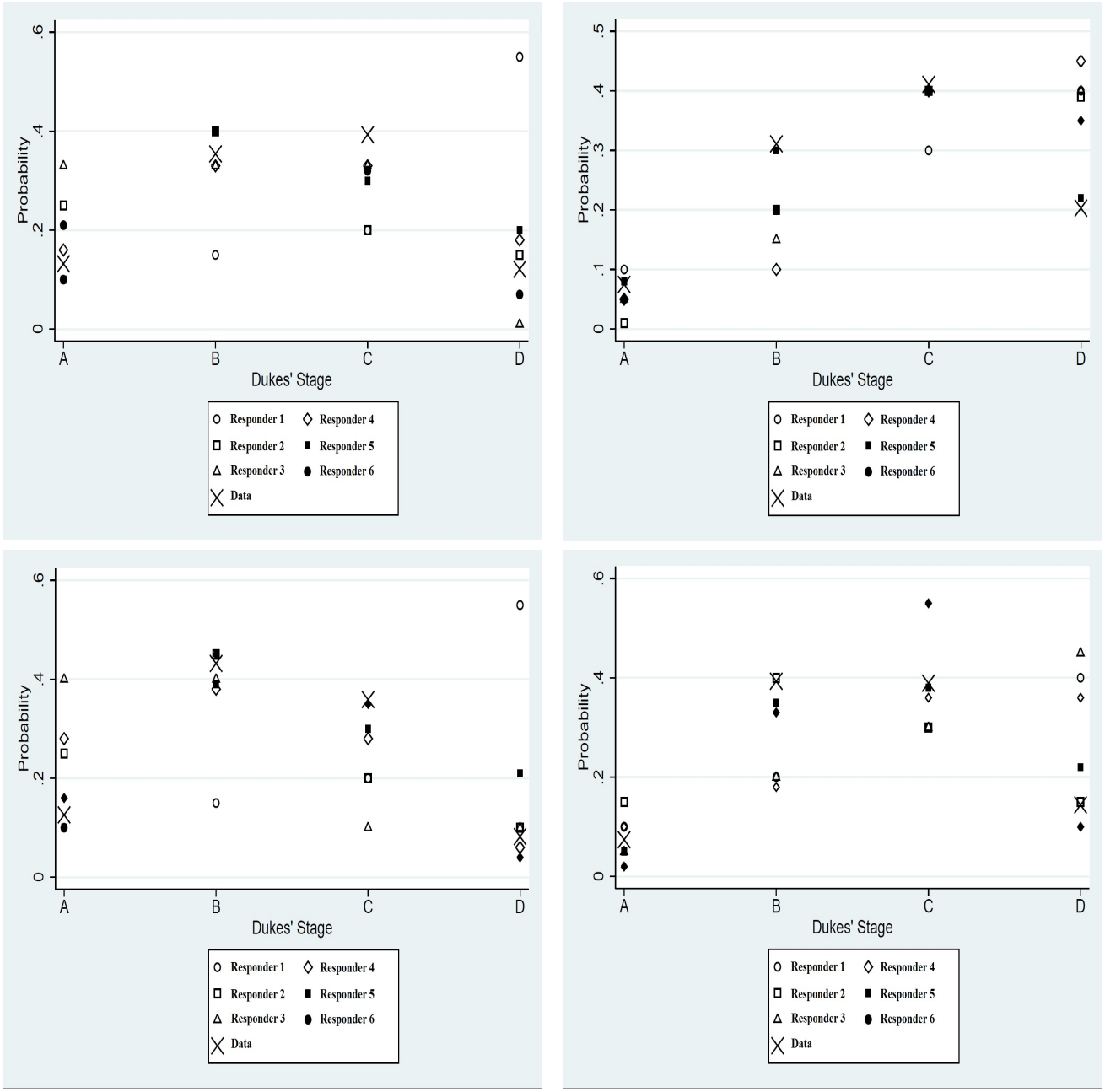

Figure 3.1: Individual responses from the questionnaire, 'Data' represents the MAR probabilities. Top left: patients alive after 30 days and less than or equal to 70 years old, top right: patients dead after 30 days and less than or equal to 70 years old, bottom left: patients alive after 30 days and greater than 70 years old and bottom right: patients dead after 30 days and greater than 70 years old. Data represents the $\hat{P}_{rj}$ probabilities from the MAR data.

Figure 3.1 shows that the individual responses are quite varied with no distinct differences in the way experts filled in the questionnaire. The number of experts who completed the questionnaire was small as many experts replied saying they did not have the knowledge to give estimates for the probabilities being requested. The advantage that 4 of our sample had worked directly with the data meant they had a good understanding of it, however this may mean that their probabilities were based on what they had seen in the data (data dependency) which may mean their elicited probabilities are unduly influenced by the published MAR analysis.

Once the prior beliefs were elicited we found the means $\hat{E}[P_{rjv}]$ (denoted as $\hat{\pi}_{rj}$) and variances $\hat{Var}[P_{rjv}]$ (denoted as $\hat{V}_{rj}$) of the opinions, where $r = 1, \ldots 4$, $j = 1, \ldots 4$ and $v = 1, \ldots, K$, K is the number of experts. The means $\hat{\pi}_{rj}$ and variances $\hat{V}_{rj}$ from elicited probabilities for each Dukes' stage and r can be observed in Table 3.7.

Characteristic	Dukes' Stage probabilities* under MAR	Elicited Dukes' stage probabilities mean*(variance)
Alive 30 days after surgery and age ≤ 70		
Dukes' stage A	0.13	0.21 (0.008)
Dukes' stage B	0.35	0.32 (0.010)
Dukes' stage C	0.39	0.28 (0.005)
Dukes' stage D	0.12	0.19 (0.045)
Dead 30 days after surgery and age ≤ 70		
Dukes' stage A	0.08	0.05 (0.001)
Dukes' stage B	0.31	0.17 (0.002)
Dukes' stage C	0.41	0.38 (0.002)
Dukes' stage D	0.20	0.40 (0.001)
Alive 30 days after surgery and age > 70		
Dukes' stage A	0.13	0.24 (0.013)
Dukes' stage B	0.43	0.37 (0.016)
Dukes' stage C	0.36	0.23 (0.009)
Dukes' stage D	0.08	0.17 (0.046)
Dead 30 days after surgery and age > 70		
Dukes' stage A	0.07	0.08 (0.003)
Dukes' stage B	0.39	0.26 (0.010)
Dukes' stage C	0.39	0.36 (0.012)
Dukes' stage D	0.14	0.29 (0.025)

*Probabilities rounded to 2dp

Table 3.7: Dukes' stage probabilities given age and mortality status 30 days after the patient operation, for the data estimated under a MAR assumption and the elicited mean(variance) from the questionnaires.

We can see from Table 3.7, that the elicited probabilities for Dukes' stage D given any r, are larger than the imputed MAR probabilities. This suggests that the experts on average believe the probability of being in Dukes' stage D is higher than estimates derived from the observed data. The elicited probabilities have decreased Dukes' stage B and C given any r and Dukes'

stage A has increased except when the patient is dead 30 days after surgery and less than or equal to 70 years old.

For each r, we modelled the elicited data with a Dirichlet distribution (Teh *et al.*, 2006) because it is a conjugate prior for the multinomial distribution. The Dirichlet distribution model is represented as:

$$f(\pi_{r1}, \pi_{r2}, \pi_{r3}, \pi_{r4}; \gamma_{r1}, \gamma_{r2}, \gamma_{r3}, \gamma_{r4}) = \frac{1}{D(\boldsymbol{\gamma})} \prod_{j=1}^{4} \pi_{rj}^{\gamma_{rj}-1}.$$

Where $\pi_{rj} > 0$, $\gamma_{rj} > 0$ and $\sum_{j=1}^{4} \pi_{rj} = 1$. $D(\boldsymbol{\gamma})$ is the normalising constant which is a multinomial beta function expressed in terms of a gamma function:

$$D(\boldsymbol{\gamma}) = \frac{\prod_{j=1}^{4} \Gamma(\gamma_{rj})}{\Gamma\left(\sum_{j=1}^{4} \gamma_{rj}\right)}$$

The mean of the Dirichlet distribution is:

$$E[\pi_{rj}] = \frac{\gamma_{rj}}{\sum_{s=1}^{4} \gamma_{rs}} = \frac{\gamma_{rj}}{S_r}. \tag{3.6.1}$$

The denominator, S_r defines the precision of the distribution. When scalar S_r is large (much greater than 1), draws of π_{rj} parameters are likely to be similar to their expectation, $E[\pi_{rj}]$, thus the distribution is more concentrated. When S_r is small (less than 1) π_{rj} parameter distribution is more diffuse.

The Dirichlet variance of π_{rj} is:

$$Var[\pi_{rj}] = \frac{\gamma_{rj}(S_r - \gamma_{rj})}{S_r^2(S_r + 1)}. \tag{3.6.2}$$

We wished $Var[\pi_{rj}]$ to be approximately equal to the empirical variance of the elicited data, $\hat{V}_{rj}$, so when we draw from the Dirichlet distribution, the variance is similar to the responses in the questionnaire. We estimate S_r to achieve this, using the method of moments. We used a range of values for S_r in (3.6.1) (with $\hat{E}[\pi_{rj}] = \hat{\pi}_{rj}$) which were then substituted into (3.6.2) to find the Dirichlet distribution variances $Var[\pi_{rj}]$. We then plotted these alongside the elicited data to help choose S_r. Figure 3.2 shows an example of finding S_3 (the group of patients who are alive at 30 days and older than 70).

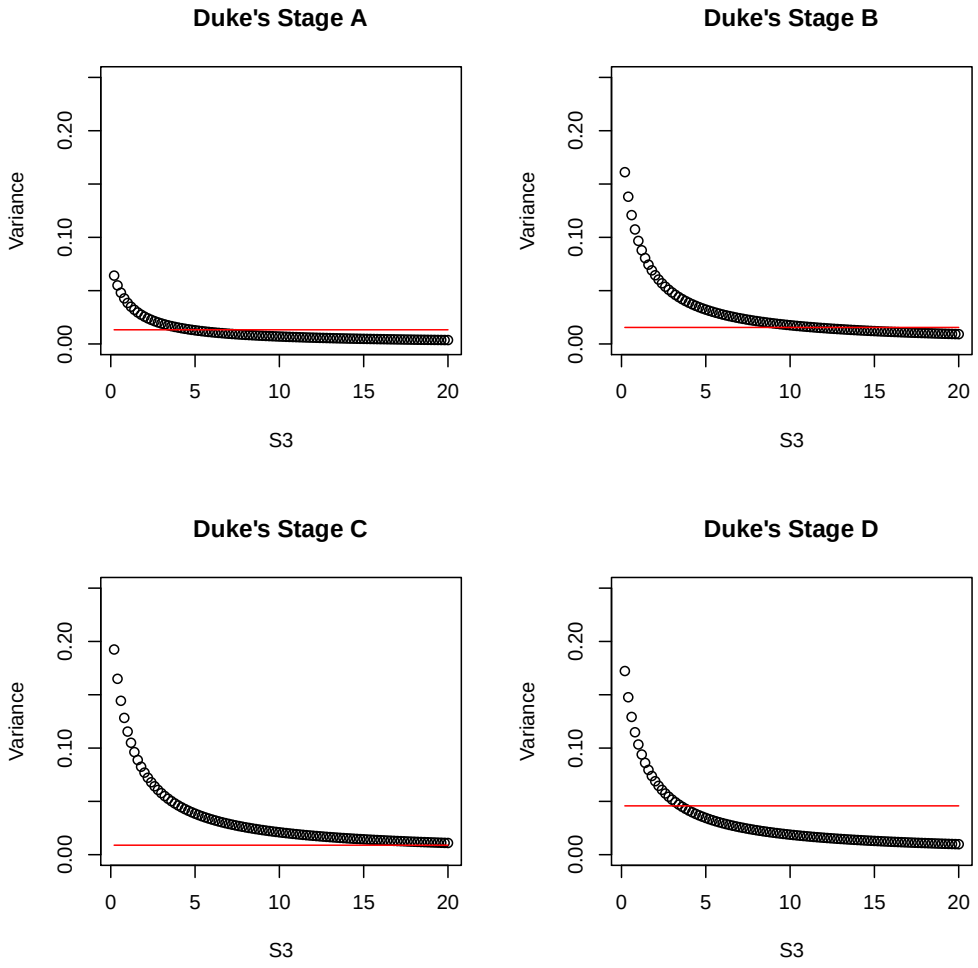

Figure 3.2: Dukes' Stage Dirichlet variances $\hat{V}ar[\pi_{3j}]$ (circle points) and empirical variance of questionnaire responders $\hat{V}_{3j}$ (horizontal lines) for varying S_3 based on patients alive at 30 days and greater than 70 years old.

The solid lines in Figure 3.2 represent the empirical variances ($\hat{V}_{3j}$) of the estimates from the questionnaire responses. We wish the Dirichlet variance $\hat{V}ar[\pi_{rj}]$ to be approximately equal to the questionnaire variance $\hat{V}_{rj}$. Clearly this is not achieved by any single S, however taking the average ($E[\hat{S}_{3j}]$) of the $\hat{S}_3$'s when $\hat{V}ar[\pi_{3j}]$ equals the solid line $\hat{V}_{3j}$ will solve (3.6.2) and (3.6.1).

Using this information we found the approximate parameters $\hat{\gamma}_{rj}$ for the Dirichlet distribution.

	Parameters for the Dirichlet distribution by Dukes' stage				Precision S
r	A	B	C	D	
1	4.31	6.62	5.67	3.94	20.55
2	2.51	8.21	18.35	19.22	48.30
3	3.01	4.62	2.86	2.15	12.64
4	2.50	7.80	10.78	8.70	29.78

Values rounded to 2dp

Table 3.8: Gamma parameters ($\hat{\gamma}_{rj}$) and S for the Dirichlet distribution by Dukes' stage, age at diagnosis and 30 day post surgery mortality.

We used these to draw the imputed data, which is now imputed under MNAR. The *substantive* model was fitted to each MNAR imputed data set and the results combined for inference using Rubin's combination rules. The results were then compared to the inferences from the MAR imputed data.

3.7 Results

The results from the multivariable analyses examining the adjusted odds of death within 30 days of surgery are shown in Table 3.9.

	Multiple Imputation MAR			Multiple Imputation MNAR		
Characteristic	AOR	(95% CI)	p Value	AOR	(95% CI)	p Value
Age at diagnosis (per 10 years)	2.12	(2.07-2.17)	<0.001	2.08	(2.03-2.13)	<0.001
Year of diagnosis (per advancing year)	0.97	(0.97-0.98)	<0.001	0.98	(0.97-0.98)	<0.001
Sex						
Female	1.00			1.00		
Male	1.21	(1.16-1.26)	<0.001	1.21	(1.16-1.26)	<0.001
Operation						
Elective	1.00			1.00		
Emergency	2.67	(2.53-2.82)	<0.001	2.76	(2.61-2.91)	<0.001
Dukes' stage at diagnosis						
A	1.00			1.00		
B	1.24	(1.14-1.35)	<0.001	1.15	(1.04-1.26)	0.005
C	1.54	(1.42-1.68)	<0.001	1.29	(1.18-1.42)	<0.001
D	2.48	(2.25-2.73)	<0.001	1.82	(1.66-2.00)	<0.001
IMD income category						
Most affluent	1.00			1.00		
2	1.03	(0.96-1.10)	0.429	1.03	(0.96-1.10)	0.417
3	1.11	(1.04-1.19)	<0.001	1.12	(1.04-1.19)	0.002
4	1.21	(1.13-1.30)	<0.001	1.22	(1.14-1.30)	<0.001
Most deprived	1.32	(1.23-1.42)	<0.001	1.32	(1.23-1.42)	<0.001
Cancer site						
Colon	1.00			1.00		
Rectosigmoid	0.88	(0.82-0.96)	0.003	0.88	(0.82-0.96)	0.003
Rectum	0.94	(0.89-0.99)	0.018	0.91	(0.86-0.96)	0.001
Charlson comorbidity score						
0	1.00			1.00		
1	2.05	(1.94-2.17)	<0.001	2.06	(1.95-2.19)	<0.001
2	2.43	(2.25-2.62)	<0.001	2.42	(2.24-2.61)	<0.001
≥ 3	4.39	(3.99-4.83)	<0.001	4.35	(3.96-4.79)	<0.001

Table 3.9: Multivariable analyses showing the adjusted odds ratios (AOR) and associated 95% confidence intervals for death within 30 days of surgery for the 'Missing At Random' missing data assumption and 'Missing Not At Random' missing data assumption, based on 10 imputations.

Table 3.9 shows that results from MAR and MNAR are broadly similar. The significance of variables at the 95% level has not changed. The largest absolute differences in AOR's can be observed for the Dukes' stages. For Dukes' stage B, the AOR decreases 7.3% from 1.24 to 1.15, stage C decreases 16.2% from 1.54 to 1.29 and stage D decreases 26.6% from 2.48 to 1.82. It is important to note that the AOR's from the MAR imputation for Dukes' stage C and D are not within the confidence interval for the corresponding Dukes' stage under MNAR, suggesting the AOR's are significantly different however the direction of risk and significance remains the same. The imputation under the assumption of MNAR reduces the effect of Dukes' stage on 30 day postoperative mortality, by contrast, the AOR or elective vs emergency surgery has increased by 3.4%. This suggests that, if the experts views are correct, Dukes' stage is a less important predictor than MAR would suggest, while emergency surgery is a more important predictor. We also investigated the variance elements of the multivariable analyses for both the MAR and MNAR imputed data, this can bee seen in Appendix B.5. The between and within variability is comparable between the MAR and MNAR, which is consistent with the results in Table 3.9.

3.8 Discussion and Conclusion

Data conclusions

This chapter was motivated by a colorectal cancer dataset collected and analysed by Morris *et al.* (2011). They handled the missing data by imputation under MAR. Using the same *substantive* and imputation model which was applied in the article, we replicated their results. We then moved a step further, carrying out a pattern mixture sensitivity analysis by eliciting

information though a questionnaire from experts for a prior. We validated the approach with a simulation study which showed the MNAR approach in this setting gave good confidence interval coverage and unbiased inferences.

We had 6 responses to our questionnaire which we modelled using a Dirichlet distribution to impute the data under the MNAR distribution. We then applied the *substantive* model to each MNAR imputed data set and combined the results using Rubin's rules. We believe that the elicited prior is reliable as 4 of the 6 responders had very good knowledge of the data, however more expert responses would have strengthened our conclusions.

Broadly speaking, we found the results were similar to the MAR analysis. However, the effect of Dukes' stage was reduced under MNAR, with estimates under MAR for Dukes' stage C and D now fully outside the 95% MNAR confidence interval.

General conclusions on methodology

The method we have proposed, and the associated software to implement it, allows us to take into account informatively missing data in the analysis. This allows the sensitivity of the inferences to the MAR assumption to be evaluated. The main limitation to the method is the difficulty in obtaining prior information that is reliable and accurate. First, we found not all experts were comfortable with giving their opinion in line with White *et al.* (2007) and secondly it is unclear who exactly is an appropriate expert. We used electronic communication to elicit information from experts; however face to face meetings would almost certainly have aided this process as it would give more freedom to question and clarify concepts. Thus the accuracy of the information cannot be ascertained, more research into the reliability and methodology of

collecting prior information is needed. Further research into the number of experts who need to be consulted should also be done, as it is not currently clarified in literature.

Where prior information is not available, we may consider a tipping point analysis. This moves the data further and further away from MAR in a particular direction, until key inferences from the substantive model change. Experts can then discuss whether the degree of departure from MAR needed to change the substantive results is plausible. A further point of interest concerns the standard error of the parameter estimates under MNAR, and how they compare to those obtained under MAR. Ideally we would like them to be greater under MNAR, reflecting the uncertainty relative to MAR. However, this is not guaranteed under our approach - it depends on the variability of the imputation distribution derived from expert opinion relative to that under MAR. Fortunately, looking at the standard errors in Table 3.9 this does not appear to be an issue here.

Lastly, neither our MAR or MNAR imputation has allowed for the multilevel structure (multilevel by trust and registry) of the data. This is because we wished to reproduce the inferences in Morris *et al.* (2011), who did not use multilevel multiple imputation. In the next chapter we use a multilevel multiple imputation with a different example, however we note that an elicitation process such described here could readily be used in the multilevel setting.

Multiple imputation software is becoming more accessible and the default assumption for the missing data is MAR. If inferences are sensitive to this, those to whom the research is presented (e.g. journal readers) should be aware of this. Broadly, the greater the number of missing observations, the greater the proportion of missing information and the more sensitive the inferences are to the imputation assumptions. As the MAR assumption is untestable, ad-

ditional analysis to explore the sensitivity of inferences to the MAR assumption are desirable. However this extra work can be time consuming, so a limitation of our method is the time it takes to elicit the prior information possibly putting analysts off. However an advantage of this approach is that it is computationally simple once the prior has been elicited. Unlike under the selection modelling approach, the missingness mechanism model does not need to be jointly modelled with the substantive model, which can be computationally demanding.

Further, having imputed the data under MNAR we can readily fit a logistic model to the imputed data to understand the selection implication of our pattern mixture assumption. Also, as discussed above, if elicitation is difficult, a tipping point analysis may be considered.

In our setting we only selected two predictive variables, age at diagnosis and 30 day post surgery mortality to frame the elicitation process. This was done to simplify the questionnaire. However in some situations this simplification would not be appropriate, making the elicitation of information difficult.

Of course, this approach is no substitute to collecting data fully. We do however believe that the process of eliciting information helps raise awareness of the potential loss of information and possible bias caused by missing data. This is turn leads to more emphases minimising missing data in future study designs.

In summary, we believe the approach described here is a computationally feasible, accessible and practical approach to sensitivity analysis within epidemiology. We hope it may find application in the area where missing data are often an issue, because the data were collected for direct clinical need, not with research in mind.

4

Sudden Infant Death Syndrome: Multilevel Multiple Imputation and Sensitivity Analysis

4.1 Introduction

In this chapter we consider data from an individual patient meta analysis of five multi-centre case control studies of Sudden Infant Death Syndrome (SIDS). The data are multilevel (babies nested in centres) and information on two key variables is missing for some centres. In this chapter we describe the multilevel multiple imputation we performed for the primary analysis in Carpenter *et al.* (2013), together with a series of sensitivity analyses to explore the robustness of the conclusions about the risk of bedsharing to assumptions about the missing data. The lead author of the paper is Professor Robert Carpenter, he proposed the study, collected the data and selected the statistical analysis models. My role was to investigate the missing data, create an imputation model and impute the data. Professor James Carpenter independently repeated the imputation to validate the results and I preformed sensitivity analysis to investigate the robustness of the inferences to the MAR assumption used in the imputation procedure.

Sudden infant death syndrome, also known as cot death, is defined as an infant death whose post-mortem examination could not explain the cause of death (Lullaby Trust 2014). In 2011 8% of all infant deaths were accounted as SIDS. Currently SIDS occur at a rate of 0.34 deaths per 1,000 live births in England and Wales. This rate decreases to 0.19 per 1,000 when children are born inside marriage and increases to 0.91 per 1,000 live births when the baby was registered by the mother only outside of marriage (Lullaby Trust 2014). Guidelines to reduce SIDS vary by country and thus the literature does not show definitive risks. The risk of SIDS when bed sharing is a controversial topic and thus a multicentre case-control study was undertaken to estimate the associated risk of bed sharing with SIDS in relation to other factors.

This chapter first introduces the data and then presents the scientific model of interest (fitted in Carpenter *et al.* (2013)). It next discusses the multiple imputation under MAR that we performed and concludes with a range of sensitivity analyses to explore the robustness of inferences about the risk of bed sharing to the missing at random assumption.

4.2 Sudden Infant Death Syndrome Data

The SIDS data set came from 5 independent case-control data sets:

- European case control study, (ECAS), 1992 - 1996 (Carpenter *et al.*, 2004)
- Scottish study, 1996 - 2000 (Tappin *et al.*, 2005)
- New Zealand study, 1987 - 1990 (Mitchell *et al.*, 1992)
- Irish study, 1994 - 2003 (McGarvey *et al.*, 2006)
- German study, (GeSID), 1998 - 2001 (Findeise *et al.*, 2004)

Data were collected at more than one centre in some of the studies. The aim of the combined data study was to analyse risk factors for SIDS cases and to explore the risk associated with bed sharing. The data consisted of N=6151 babies from 19 centres, of which 1472 were SIDS cases and 4679 were controls of a similar age who did not have a SIDS event (see individual studies for details of the control selection information). Due to differing study protocols, variables

collected within studies varied and thus some risk factors (alcohol intake in the previous 24 hours of questioned night, and drug abuse after birth) which we wished to explore and adjust for were not present in every study. The Scottish study (Tappin *et al.*, 2005), New Zealand study (Mitchell *et al.*, 1992) and GeSID study (Findeise *et al.*, 2004) had close to 100% missing for alcohol and drug intake, see Table 4.3.

The data contained 3 fully observed variables, baby age (days), centre reference and case/control indicator. There are 15 partially observed variables, the frequency and percentage of missing data for each variable is shown in Table 4.1:

Variable	Missing (%)
Sex	18 (0.29)
Bed Shared[A] (Yes/No)	58 (0.94)
Race (European Yes/No)	18 (0.29)
Other Room[B]	89 (1.45)
Mother Smokes	61 (0.99)
Partner Smokes	135 (2.19)
Birth Weight	141 (2.29)
Mother Drank Alcohol ($>$ 2 units, Yes/No)[C]	3,768 (61.26)
Mother Used Drugs After Birth (Yes/No)[D]	3,717 (60.43)
Bottle Fed (Yes/No)	48 (0.78)
Married or Cohabiting (Yes/No)	11 (0.18)
Number Of Live Births[E]	48 (0.78)
Mothers Age	34 (0.55)
Sleeping Position Left On Last Occasion[F]	97 (1.58)
Matched By Sex[G]	18 (0.29)

[A]Bed shared defined as baby sleeping in bed with one or two people on the night in question. [B]Baby slept in a separate room from the mother. [C]In last 24 hours. [D]Illegal drugs. [E]Parity. [F]1) Supine, 2) Side or 3) Prone. [G]1) Female, 2) Male matched or 3) Male un-matched.

Table 4.1: Frequency and percentage of missing data by variable ($N = 6151$).

Table 4.1 shows that the variables '*mother drank alcohol*' (in the 24 hours prior to death for cases or interview for controls) and '*mother used drugs after birth*' contain the highest proportion of missing values. To focus on these key variables, we removed the partially observed records for all variables other than '*mother drank alcohol*' and '*mother used drugs after birth*'. The data still contained 4322 controls (over 92%) and 1,305 cases (over 88%). The '*mother drank alcohol*' indicator now had 3,510(62%) missing observations and '*mother used drugs after birth*' 3,469(62%).

4.3 Model of Interest

The model of interest (MOI) we applied in Carpenter *et al.* (2013) is a multilevel logit regression model with bed sharing random across centres to take in to account the significant interaction between bed sharing and centre. Other random effects for parameters which had significant adjusted odds ratios with centre were considered, but were dropped because they did not materially affect the parameter estimates, and made model fitting more difficult. The main interest of the analysis was to investigate the associated risk between bed sharing and SIDS after adjusting for possible confounders. The final MOI is:

$$\text{Logit P(baby } i \text{ from centre } j \text{ is a case)} = \boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \boldsymbol{Z}_{ij}^T w_j \tag{4.3.1}$$

where $i = 1, \ldots, 5627$ (indexes babies), $j = 1, \ldots, 19$ (indexes centres) and the fixed covariate effect for baby i in centre j are denoted by $\boldsymbol{X}_{ij}$. $\boldsymbol{X}_{ij}$ contains the following main effect indicator/categorical variables; bed shared, bottle fed, smoking (grouped), centred age, other room, birth weight (grouped), married or cohabiting, mother's age (grouped), number of live births

grouped, race and matched by sex indicator. The fixed effects also contained the following interactions: smoking and bed shared (interaction A), position left in and bed shared and baby age (interaction C), mother drank alcohol and bed shared (interaction D), mother used drugs after birth and bed shared yes (interaction E), mother used drugs after birth and bed shared no (interaction F) (see Table 4.2 for more information on covariates, interactions and grouped covariates). Note that interaction B is not included in the MOI, this is discussed later.

Created Variable	Information
Smoking Grouped	A categorical variable with 4 groups; 1) neither mother or partner smoke, 2) only partner smokes, 3) only mother smokes and 4) both smoke.
Birth Weight Grouped	A categorical variable with 4 groups; 1) birth weight<2000, 2) 2000≤birth weight<2500, 3) 2500≤birth weight<3500 and 4) birth weight≥3500.
Mother Age Grouped	A categorical variable with 5 groups; 1) age≥31, 2) 26≤age<31, 3) 21≤age<26, 4) 19≤age<21 and 5) age<19.
Number Of Live Births Grouped	A categorical variable with 5 groups; 1) 1 birth, 2) 2 births, 3) 3 births, 4) 4 births and 5)≥ 5 births.
Smoking Grouped, Bed Shared Interaction (A)	A categorical variable with 7 groups; 1) neither mother or partner smoke and no bed share, 2) partner smokes not bed share, 3) mother smokes not bed share, 4) both smoke not bed share, 5) partner smokes and bed shares, 6) mother smokes and bed shares and 7) both smoke and bed share.
Centred Age	Baby's age centred on 6 months (age - 182).
Centred Age, Bed Shared Interaction (B)	The variable represents bed shared×centred age unless baby is older than 6 months then equal to 0.
Position Left In, Bed Shared, Baby Age Interaction (C)	A categorical variable with 7 groups; 1) supine+(baby age<3m & bed shared), 2) side + not bed shared & baby age<3m, 3) prone + not bed shared & baby age<3m, 4) side + not bed shared & baby age≥3m, 5) prone + not bed shared & baby age≥3m, 6) side + bed shared & baby age≥3m and 7) prone + bed shared & baby age≥3m.
Mother Drank Alcohol, Bed Share Interaction (D)	A categorical variable with 3 groups; 1) mother drank alcohol≤2 units, 2) mother drank alcohol≥2 units & not bed share and 3) mother drank alcohol≥2 units & bed share.
Mother Used Drugs After Birth, Bed Shared No Interaction (E)	A binary variable; 1) not bed shared and drug taking and 0) all other combinations.
Mother Used Drugs After Birth, Bed Shared Yes Interaction (F)	A binary variable; 1) bed shared and drug taking and 0) all other combinations.

Table 4.2: Description of created grouped variables and interactions used in the analyses.

The interactions (seen in Table 4.2) were considered so that we could explore the hypothesis that the associated risks are larger if you bed share and behave/have these characteristics. The random effects for baby i in centre j are denoted by $\boldsymbol{Z}_{ij}$; $\boldsymbol{Z}_{ij}$ contains a constant and bed shared indicator. We also define $w_j \sim N(0, \boldsymbol{\Omega}_w)$, where the level 2 covariance matrix $\boldsymbol{\Omega}_w$ is

$$\boldsymbol{\Omega}_w = \begin{bmatrix} \sigma^2_{wC} & \sigma_{wCB} \\ \sigma_{wCB} & \sigma^2_{wB} \end{bmatrix}$$

with C representing the level 2 random effect for the constant and B the level 2 random effect for the bed sharing covariate.

To explore the data further we look at how '*mother drank alcohol*' (*alcohol*) and '*mother used drugs after birth*' (*drug*) vary by centre and bed sharing.

		Mother Drank Alcohol		*Mother Used Drugs After Birth*	
		Bed shared		Bed shared	
Study	Centre	No N(%)	Yes N(%)	No N(%)	Yes N(%)
ECAS	1	12 (66.7)	6 (33.3)	1 (100.0)	0 (0.0)
ECAS	2	6 (85.7)	1 (14.3)	0 (0.0)	0 (0.0)
ECAS	3	16 (69.6)	7 (30.4)	1 (100.0)	0 (0.0)
ECAS	4	2 (66.7)	1 (33.3)	1 (100.0)	0 (0.0)
ECAS	5	1 (100.0)	0 (0.0)	0 (0.0)	0 (0.0)
ECAS	6	1 (100.0)	0 (0.0)	0 (0.0)	0 (0.0)
ECAS	7	1 (100.0)	0 (0.0)	0 (0.0)	0 (0.0)
ECAS	8	0 (0.0)	0 (0.0)	1 (100.0)	0 (0.0)
ECAS	9	0 (0.0)	1 (100.0)	0 (0.0)	0 (0.0)
ECAS	10	3 (100.0)	0 (0.0)	0 (0.0)	0 (0.0)
Irish	18	76 (62.3)	46 (37.7)	6 (50.0)	6 (50.0)

Table 4.3: Separate frequencies/percentages for variables *alcohol* and *drug*, cross classified by bed sharing, study and centre.

We can see in Table 4.3 that *alcohol* and *drug* users generally did not bed share. We can also see how sparse the data are with only 180 observed *alcohol* users and 16 observed *drug* users.

4.4 Exploring the Missing Data

The model of interest (MOI) is the starting place for exploring the missing data. We observed that both variables with missing data *alcohol* and *drug* are used in the MOI, as part of interactions. It is clear in the data that if *alcohol* is missing then *drug* is very likely also to be missing. There was 2,116 babies who have both *alcohol* and *drug* observations, and 3,468 babies with both missing. Only 42 babies had *alcohol* missing but *drug* observed and 1 baby with *alcohol* observed and *drug* missing. We created two binary indicators, representing if *alcohol* was observed or missing and if *drug* was observed or missing. The indicators where applied in a chi squared test, which suggested a very strong association between missingness ($p < 0.001$), thus the reason for *alcohol* and *drug* being missing is likely to be connected.

There also appeared to be a strong association between *alcohol* or *drug* being missing and which study/centre the baby came from, as can be seen in Table 4.4.

Study	Centre	Missing *alcohol* Frequency (%)	Missing *drug* Frequency (%)
ECAS	1	0 (0.0)	0 (0.0)
ECAS	2	0 (0.0)	0 (0.0)
ECAS	3	0 (0.0)	0 (0.0)
ECAS	4	0 (0.0)	0 (0.0)
ECAS	5	17 (37.0)	2 (4.4)
ECAS	6	3 (4.6)	0 (0.0)
ECAS	7	0 (0.0)	0 (0.0)
ECAS	8	3 (7.14)	0 (0.0)
ECAS	9	18 (56.3)	0 (0.0)
ECAS	10	0 (0.0)	0 (0.0)
ECAS	11	0 (0.0)	0 (0.0)
ECAS	12	0 (0.0)	0 (0.0)
ECAS	13	0 (0.0)	0 (0.0)
ECAS	14	0 (0.0)	0 (0.0)
ECAS	15	0 (0.0)	0 (0.0)
Scottish	16	1,823 (100.0)	1,823 (100.0)
New Zealand	17	357 (99.7)	357 (99.7)
Irish	18	3 (0.4)	1 (0.13)
GeSID	19	1,286 (99.9)	1,286 (99.9)
Total		3,510 (62.4)	3,469 (61.7)

Table 4.4: Frequency and percentage of missing '*mother drank alcohol*' and '*mother used drugs after birth*' observations by study and centre.

Table 4.4 clearly shows that the majority of missing observations are from the Scottish, New Zealand and GeSID studies. This is because these studies did not set out to record information on mother's *alcohol* or *drug* intake. This suggests the missing pattern is mainly created by study and not a specific association with other variables. This is consistent with the different data collection protocols in different studies. The studies with the majority of missing observations only have one centre, thus centre and study are interchangeable for them. For this reason and because the MOI (4.3.1) clusters by *centre* not study, we move forward discussing

missing in terms of *centre*. As long as *centre* is used in the model of interest then a complete records analysis will not be biased. To see why, consider the following argument:

Let $\boldsymbol{R} = 1$ represent a complete record and $\boldsymbol{R} = 0$ an incomplete record. The outcome data $\boldsymbol{Y}$ is fully observed and the covariates $\boldsymbol{X}$ are partially observed. If the probability of a complete record is just a function of $\boldsymbol{X}$ then:

$$P(\boldsymbol{R} = 1 \mid \boldsymbol{Y}, \boldsymbol{X}) = P(\boldsymbol{R} = 1 \mid \boldsymbol{X}).$$

Then

$$\begin{aligned} P(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{R} = 1) = \frac{P(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{R} = 1)}{P(\boldsymbol{X}, \boldsymbol{R} = 1)} &= \frac{P(\boldsymbol{R} = 1 \mid \boldsymbol{Y}, \boldsymbol{X}) P(\boldsymbol{Y}, \boldsymbol{X})}{P(\boldsymbol{R} = 1 \mid \boldsymbol{X}) P(\boldsymbol{X})} \\ &= \frac{P(\boldsymbol{Y}, \boldsymbol{X})}{P(\boldsymbol{X})} = P(\boldsymbol{Y} \mid \boldsymbol{X}). \end{aligned}$$

Thus the regression relationship estimated in the complete records validly represent those in the population the data are drawn from. Imputation under MAR is thus likely to give broadly similar point estimates to the complete records analysis. However, this may nethertheless be very beneficial as over half of the data set will be removed in a complete records analysis and thus it should markedly increase the accuracy of the estimates. In order not to introduce bias when imputing under the MAR assumption the imputation model needs to be congenial (see section 2.17) with the structure in the substantive model.

We wish to show graphically (as we did in Carpenter *et al.* (2013)) the adjusted odds ratio for bed sharing over the first three months. The MOI does not include an interaction between bed sharing and age. We thus include an extra interaction covariate (interaction B

as described in Table 4.2), as a fixed effect in the MOI model. We call this new model the Adjusted Odds Ratio Model of Interest (AOR MOI).

4.5 Initial Imputation Exploration

For the imputation model to be congenial we require it to have the structure of the substantive model and account for the important factors in the missing data mechanism. We thus began with our imputation models similar to the AOR MOI (as it is more complex than the MOI). As we have two variables with missing data (*alcohol* and *drug*) we need a bivariate response imputation model and we also need to apply a joint modelling approach to the multiple imputation to handle the multilevel structure. The first imputation models were thus:

$$\text{Logit P(baby } i \text{ from centre } j \text{ with } alcohol \text{ drunk)} = \boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \boldsymbol{Z}_{ij}^T u_j \tag{4.5.1}$$

$$\text{Logit P(baby } i \text{ from centre } j \text{ with } drug \text{ Taken)} = \boldsymbol{X}_{ij}^T\boldsymbol{\alpha} + \boldsymbol{Z}_{ij}^T v_j \tag{4.5.2}$$

where $u_j \sim N(0, \boldsymbol{\Omega}_u)$ and $v_j \sim N(0, \boldsymbol{\Omega}_v)$. The level 2 covariance matrix $\boldsymbol{\Omega}$ is,

$$\begin{pmatrix} \boldsymbol{\Omega}_u \\ \boldsymbol{\Omega}_v \end{pmatrix} = N\left[0, \boldsymbol{\Omega} = \begin{pmatrix} \sigma_{uC}^2 & \sigma_{uC}\sigma_{uB} & \sigma_{uC}\sigma_{vC} & \sigma_{uC}\sigma_{vB} \\ \sigma_{uC}\sigma_{uB} & \sigma_{uB}^2 & \sigma_{uB}\sigma_{vC} & \sigma_{uB}\sigma_{vB} \\ \sigma_{uC}\sigma_{vC} & \sigma_{uB}\sigma_{vC} & \sigma_{vC}^2 & \sigma_{vC}\sigma_{vB} \\ \sigma_{uC}\sigma_{vB} & \sigma_{uB}\sigma_{vB} & \sigma_{vC}\sigma_{vB} & \sigma_{vB}^2 \end{pmatrix}\right]$$

where C represents the random effect level 2 constant and B the random effect level 2 bed shared covariate. The fixed effects for baby i in centre j are denoted by $\boldsymbol{X}_{ij}$. $\boldsymbol{X}_{ij}$ contains

the same covariates as the MOI (4.3.1) with the indicator for case/control and Interaction B added. Another alteration to the AOR MOI was we chose to use the continuous variables to impute with (i.e. mothers age instead of mothers age grouped), to retain the information lost within the grouping procedure. The covariates for the random effects for baby i in centre j are denoted by $\boldsymbol{Z}_{ij}$ and are the same as in (4.3.1).

Alcohol and *drug* are not included directly in the AOR MOI, but through interactions with *bed shared*. Thus *Alcohol* appears in one interaction (interaction D) and *drug* appears in two (interaction E and F). These interactions will contain missing observations when either *alcohol* and/or *drug* are missing, thus they cannot be used in the imputation models. However as *bed shared* is a covariate in both the *alcohol* and *drug* imputation models, then the interaction effect will be imputed if the interaction between *bed shared* and case/control studies is included in the imputation model or cases and controls are imputed separately and *bed shared* is a covariate. To demonstrate this point consider following the example where the model of interest is:

$$\text{Logit}(\boldsymbol{Y}) = \beta_0 + \beta_1 \boldsymbol{X} + \beta_2 \boldsymbol{Z} + \beta_3 \boldsymbol{X}\boldsymbol{Z}, \tag{4.5.3}$$

where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are binary. If $\boldsymbol{X}$ is missing our imputation model should be,

$$\text{Logit}(\boldsymbol{X}) = \delta_0 + \delta_1 \boldsymbol{Y} + \delta_2 \boldsymbol{Z} + \delta_3 \boldsymbol{Y}\boldsymbol{Z}. \tag{4.5.4}$$

Thus the relationship between $\boldsymbol{Y}$ and $\boldsymbol{X}$ for different levels of $\boldsymbol{Z}$ is preserved. This demonstrates that in our SIDS imputation model, as long as we have *bed shared* included and impute separately for cases and controls, the imputation model will be congenial with the AOR MOI.

The imputation models were first investigated by trying to fit a simplified version (without the

hierarchical structure) to the observed data. To make computation easier, the data was split into cases and controls, thus removing the indicator from the fixed effects covariates matrix ($\boldsymbol{X}_{ij}$). It is also preferable to impute the data separately as the relationship between variables among cases and variables among controls maybe quite different, and we want to maintain this. We call these models without hierarchical structure *simplifed exploratory models*. As a first step these *simplifed exploratory models* were fitted in STATA to the complete records for *alcohol* and *drug* separately. The results are shown in Table 4.5 and Table 4.6.

	Cases		Controls	
Covariate	AOR	SE	AOR	SE
Bed Shared	7.01	10.15	1.38	1.25
Bottle Fed	1.57	0.56	1.75	0.49
Interaction A				
Neither mother or partner smoke, no bed sharing				
Partner smokes only, no bed sharing	0.34	0.18	1.37	0.63
Mother smokes only, no bed sharing	0.04	0.05	0.55	0.58
Both mother and partner smoke, no bed sharing	0.41	0.26	1.03	0.33
Partner smokes only, bed shared	0.37	0.20	0.60	0.33
Mother smokes only, bed shared	1.43	0.97	0.25	0.28
Both mother and partner smoke, bed shared	0.54	0.65	1.00	Omitted
Age Centred at 6m	0.54	0.16	1.18	0.21
Interaction B: Centred age, bed shared	3.35	1.89	0.47	0.32
Other Room	0.69	0.27	1.23	0.34
Interaction C				
Supine, baby<3m, bed shared				
Side, baby<3m, no bed sharing	2.69	2.34	2.65	0.84
Prone, baby<3m, no bed sharing	6.08	5.96	0.93	0.60
Side, baby≥3m, no bed sharing	6.08	5.96	0.93	0.60
Prone, baby≥3m, no bed sharing	9.66	7.45	2.77	0.99
Side, baby≥3m, bed shared	2.56	3.02	2.28	1.91
Prone, baby≥3m, bed shared	4.77	6.06	2.03	2.73
Birth weight	4.32	4.74	6.23	5.81
Married or Cohabiting	1.38	0.29	0.97	0.21
Mothers Age	1.02	0.32	0.87	0.38
Number of Live Births	1.86	0.51	1.87	0.45
Race	0.89	0.60	1.00	Omitted
Matched By Sex				
Female				
Male matched	2.77	0.87	0.77	0.21
Male unmatched	1.26	0.51	0.85	0.26
Constant	0.00	0.00	0.01	0.01

Table 4.5: Adjusted odds ratios (AOR) and standard errors (SE) for the *Simplified exploratory* '*mother drank alcohol*' imputation model fitted to cases and controls separately. Interaction C is '*Position Left In* (Supine, side, prone) with *Bed Shared* and *Baby Age* (Less than or greater than 3 months).

Covariate	Cases AOR	Cases SE	Controls AOR	Controls SE
Bed Shared	3.43e+07	1.35e+11	1.00	Omitted
Bottle Fed	1.00	Omitted	2.34E-41	2.27E-38
Interaction A				
Neither mother or partner smoke, no bed sharing				
Partner smokes only, no bed sharing	0.06	0.12	7.4E+250	1.7E+253
Mother smokes only, no bed sharing	1.00	Omitted	1.00	.
Both mother and partner smoke, no bed sharing	8.31	17.10	1.00	.
Partner smokes only, bed shared	4.47	11.24	1.00	.
Mother smokes only, bed shared	1.00	Omitted	1.00	.
Both mother and partner smoke, bed shared	1.00	Omitted	1.00	.
Age Centred at 6m	0.01	0.02	2.6E+123	8.3E+125
Interaction B: Centred age, bed shared	58.06	202.84	1.00	Omitted
Other Room	1.56	2.31	1.00	Omitted
Interaction C				
Supine, baby<3m, bed shared				
Side, baby<3m, no bed sharing	1.00	Omitted	1.00	Omitted
Prone, baby<3m, no bed sharing	1.00	Omitted	1.00	.
Side, baby≥3m, no bed sharing	3.02e+06	1.19e+10	1.00	.
Prone, baby≥3m, no bed sharing	1.00	Omitted	1.00	.
Side, baby≥3m, bed shared	1.00	Omitted	1.00	.
Prone, baby≥3m, bed shared	1.00	Omitted	1.00	.
Birth weight	0.24	0.19	1.73E+48	2.09E+50
Married or Cohabiting	9.80	14.56	1.00	Omitted
Mothers Age	0.07	0.12	2.76E+14	1.02E+16
Number of Live Births	1870.08	9746.67	9.05E-24	.
Race	1.00	Omitted	1.00	Omitted
Matched By Sex				
Female				
Male matched	6.71	9.02	1.00	Omitted
Male unmatched	1.00	Omitted	1.00	Omitted
Constant	1.71e-07	0.00	0.00	.

Table 4.6: Adjusted odds ratios (AOR) and standard errors (SE) for the *Simplified exploratory* '*mother used drugs after birth*' imputation model fitted to cases and controls separately. The controls model did not converge. Interaction C is '*Position Left In* (Supine, side, prone) with *Bed Shared* and *Baby Age* (Less than or greater than 3 months).

We see from Tables 4.5 and 4.6, that neither the *alcohol* or *drug* models fit the data well. The *drug* model did not converge for the controls, giving unreliable inferences and hence, would in turn give unreliable imputations. The *drug* model for the cases, also had difficulty converging resulting in unrealistic extremely large adjusted odds ratios. For some covariates the *alcohol* models have also produced large adjusted odds ratios (greater than 5) for some covariates, with *race* and group 7 of interaction A being omitted in the control model. We saw in Table 4.3 that we have very few observations of *alcohol* or *drug* use. This explains why the models struggle to converge, as many categories will be empty or nearly empty.

Since we are unable to fit single level logistic models to the data we will also be unable to fit a multilevel imputation model. Thus we reduced the number of covariates, removing covariates which made the models unstable. The *drug* models (for both the cases and controls) were so unstable due to the sparse data that the fixed effects was reduced to just a constant with no other covariates. The fixed effects in the *alcohol* model were also reduced: the $Matched\ By\ Sex$ indicator was replaced by a *sex* indicator, *race* was omitted, the $Smoking\ Grouped$, $Bed\ Shared\ Interaction$ was replaced with $Mother\ Smoked$ indicator and $Partner\ Smoked$ indicator. For the cases the $Position\ Left\ In$, $Bed\ Shared$, $Baby\ Age\ Interaction$ was left as the original but for the controls two categories were combined (side bed shared & baby age$\geq$3m and prone bed shared & baby age$\geq$3m were combined). The new models were checked to make sure that they were equivalent to the original simplified exploratory models using a likelihood ratio test (*lrtest* function in STATA (StataCorp, 2011b)). The *alcohol* models were found to be equivalent with a likelihood ratio test of $p = 0.4357$ for cases models and $p = 0.7500$ for controls models. The test could not be applied to the new *drug* model as the original simplified exploratory model did not converge. We now move forward with our simplified model adding the random effects back in to create a multilevel imputation model. The estimates of

the simplified multilevel model can be seen in Table 4.7.

Covariate	**Cases**			**Controls**		
Drug Model						
Fixed Effect	**AOR**	**SE**	**p > |z|**	**AOR**	**SE**	**p > |z|**
Constant	0.00	0.01	0.000	0.00	0.00	0.000
Random Effects	**Estimate**	**SE**		**Estimate**	**SE**	
Bed Shared	0.00	0.88		0.00	2.27	
Constant	1.53	1.13		0.00	1.44	
Alcohol Model						
Fixed Effect	**AOR**	**SE**	**p > |z|**	**AOR**	**SE**	**p > |z|**
Bed Shared	2.52	1.79	0.192	1.87	1.32	0.378
Bottle Fed	1.01	0.40	0.977	1.40	0.42	0.263
Age Centred at 6m	0.73	0.23	0.320	1.43	0.26	0.051
Interaction B: Centred age, bed shared	2.47	1.44	0.123	0.52	0.36	0.339
Interaction C						
Supine, baby<3m, bed shared						
Side, baby<3m, no bed sharing	0.38	0.21	0.076	1.46	0.71	0.433
Prone, baby<3m, no bed sharing	0.05	0.05	0.005	1.00	1.08	1.000
Side, baby≥3m, no bed sharing	0.44	0.28	0.199	0.96	0.33	0.901
Prone, baby≥3m, no bed sharing	0.39	0.21	0.073	0.91	0.53	0.872
Side, baby≥3m, bed shared	1.51	1.05	0.553	0.16	0.18	0.102
Prone, baby≥3m, bed shared	0.78	0.99	0.844	NI[1]		
Other Room	0.80	0.32	0.576	1.32	0.37	0.333
Mother Smokes	3.67	1.37	0.000	1.31	0.36	0.316
Partner Smokes	1.36	0.66	0.530	1.73	0.50	0.060
Birth weight	1.36	0.29	0.145	0.92	0.20	0.700
Married or Cohabiting	0.76	0.26	0.409	0.80	0.36	0.613
Mothers Age	1.50	0.43	0.151	1.58	0.41	0.076
Live Births	0.26	0.29	0.234	0.15	0.20	0.140
Sex	2.24	0.68	0.007	0.77	0.18	0.264
Constant	0.01	0.01	0.000	0.01	0.01	0.000
Random Effects	**Estimate**	**SE**		**Estimate**	**SE**	
Bed Shared	0.00	0.33		0.48	0.76	
Constant	0.67	0.26		1.02	0.80	

Table 4.7: Adjusted odds ratios (AOR) and standard errors (SE) for the *Simplified exploratory models* fitted to case and control data separately. Interaction C is '*Position Left In* (Supine, side, prone) with *Bed Shared* and *Baby Age* (Less than or greater than 3 months). Prone, baby≥3m, bed shared was Not Included (NI) in the model.

We can observe in Table 4.7 that all adjusted odds ratios are now realistic and no covariates are omitted by STATA in order to fit the model. We next take these models and use them as a bivariate response model and apply a joint modelling approach to multiply impute the missing *alcohol* and *drug* data consistent with the multilevel structure.

4.6 Multilevel Multiple Imputation

We moved forward with these models by fitting the bivariate model in the REALCOM software (Goldstein *et al.*, 2013). In order to do this, we had to modify the MATLAB code for REALCOM to extend it to incorporate prior information to allow MNAR imputation, as described below. First, though, we give an overview of the REALCOM software.

The REALCOM software uses a joint multivariate normal modelling approach through the Bayesian estimation method Markov Chain Monte Carlo (MCMC). The MCMC method generates random draws from a specified posterior distribution. The software requires a burn in to be specified, this is a set number of iterations whose data will not be used in the imputed datasets. This is so the Markov chains have time to stabilise, so that it draws from the true distribution (see Section 2.14). The number of iterations after burn in must also be specified and also the iteration at which the imputed datasets should be output. For example we could set burn in as 100, iterations as 1000 and outputted datasets at 200, 400, 600, 800 and 1000. The program would run 100 iterations then the iteration calculator would reset to 0 and start incrementing again. At the 200^{th} iteration, one imputed dataset would be created, another will be created at 400 etc creating 5 imputed datasets. The process stops after the 1000^{th} iteration. REALCOM handles binary and categorical variable responses by creating latent variables for

each category, thus in our setting two latent normal variables are created to represent the 2 binary responses *alcohol* and *drug*. For example in the cases the imputation model is thus changed to,

$$\text{Alcohol Drunk}_{ij} = P((\boldsymbol{S}_{ij}^T\boldsymbol{\beta} + \boldsymbol{Z}_{ij}^T u_j + e_{i,1}) > 0) \tag{4.6.1}$$

$$\text{Alcohol Not Drunk}_{ij} = P((\boldsymbol{S}_{ij}^T\boldsymbol{\beta} + \boldsymbol{Z}_{ij}^T u_j + e_{i,1}) < 0) \tag{4.6.2}$$

$$\text{Drug Used}_{ij} = P((\boldsymbol{U}_{ij}^T\boldsymbol{\alpha} + \boldsymbol{Z}_{ij}^T v_j + e_{i,2}) > 0) \tag{4.6.3}$$

$$\text{Drug Not Used}_{ij} = P((\boldsymbol{U}_{ij}^T\boldsymbol{\alpha} + \boldsymbol{Z}_{ij}^T v_j + e_{i,2}) < 0) \tag{4.6.4}$$

where u_j and v_j remain similar to the parameters in (4.5.1), $\boldsymbol{S}_{ij}$ represents the fixed effects for *alcohol* cases and $\boldsymbol{U}_{ij}$ represents the fixed effects for *drug* cases listed in Table 4.7. $\boldsymbol{Z}$ represents the random effects which are the same as those in (4.3.1) i.e. *bedshared* and *centre*. The covariance matrix between the latent normal models is,

$$\begin{pmatrix} e_{i,1} \\ e_{i,2} \end{pmatrix} = N\left[0, \boldsymbol{\Omega}_e = \begin{pmatrix} \sigma_A^2 & \sigma_A\sigma_D \\ \sigma_A\sigma_D & \sigma_D^2 \end{pmatrix}\right],$$

where σ_A represents the variance for *alcohol*, and σ_D the variance for *drug*. We put a uniform prior on the covariance matrix, which is the default in REALCOM. When we began fitting the models with varying number of iterations and burn in, it became clear that the model was not fitting well (coefficient estimates greater than $10x^{20}$ were obtained). We believed the problem was associated with the sparse nature of the data. With over 60% observations missing in both *alcohol* and *drug*, some centres contain very few observations. When there are no observations and/or all covariates are zero in one or more of the level 2 centres, the corresponding latent normal variable will become increasingly negative at each iteration of the MCMC algorithm.

While this should not in general cause a problem for estimating the centre level covariance matrix, the SIDS data contains many centres with this characteristic as seen in Table 4.4.

To address this issue, we decided to place a prior on the level 2 covariance matrix $\mathbf{\Omega}$ with the aim of adding sufficient information so that the matrix does not increase towards infinity. We applied an inverse Wishart prior as this is the conjugate prior for the covariance matrix under a multivariate Gaussian distribution. Another advantage of using an inverse Wishart is that the posterior distribution can be computed analytically making it easy to sample from. The probability density function of an inverse Wishart $T \sim W^{-1}(\Psi, m)$ is:

$$f(T) = \frac{|\Psi|^{\frac{m}{2}}}{|T|^{\frac{m+p+1}{2}} 2^{\frac{mp}{2}} \Gamma_p(\frac{m}{2})} exp\left[-\frac{1}{2} tr(\Psi T^{-1})\right] \tag{4.6.5}$$

where tr is the trace, p is the dimensions of T, m is the degrees of freedom ($m > p-1$) and $\Gamma_p(\frac{m}{2})$ is a multivariate gamma function:

$$\Gamma_p(\frac{m}{2}) = \pi^{\frac{1}{2}(\frac{p}{2})} \prod_{j=1}^{p} \Gamma(\frac{m}{2} + \frac{(1-j)}{2}). \tag{4.6.6}$$

We let the prior T be:

$$\begin{pmatrix} 0.01 & 0 & 0 & 0 \\ 0 & 0.01 & 0 & 0 \\ 0 & 0 & 0.01 & 0 \\ 0 & 0 & 0 & 0.01 \end{pmatrix},$$

which will constrain the off diagonal elements of the level 2 covariance matrix $\mathbf{\Omega}$ to 0. This is to reduce the complexity of the problem by making the *alcohol* and *drug* models a priori

independent at level 2. The prior T was applied to $\mathbf{\Omega}$ with varied degrees of freedom m. The results (with 1000 burn in and 100,000 iterations) of the models are shown in Appendix (C). The tables clearly show the need for a prior on $\mathbf{\Omega}$. When no level 2 prior is used the coefficients and covariances grow extremely large. A prior with just 4 degrees of freedom ($m = 4$) seems to be sufficient to constrain the coefficients and covariances of the posterior distribution to sensible values.

The new models with their priors (identical priors used for cases and controls) were applied in REALCOM, creating 20 imputed datasets (10 controls and 10 cases). To create the imputed data we applied the software program separately (with different starting random seeds to make sure the MCMC draws were independent) to the data 20 times, with a burn in of 50,000 and which then outputted a imputed dataset after 5000 iterations. The chains for each covariate were examined to check convergence, an example can be observed in Figure 4.1.

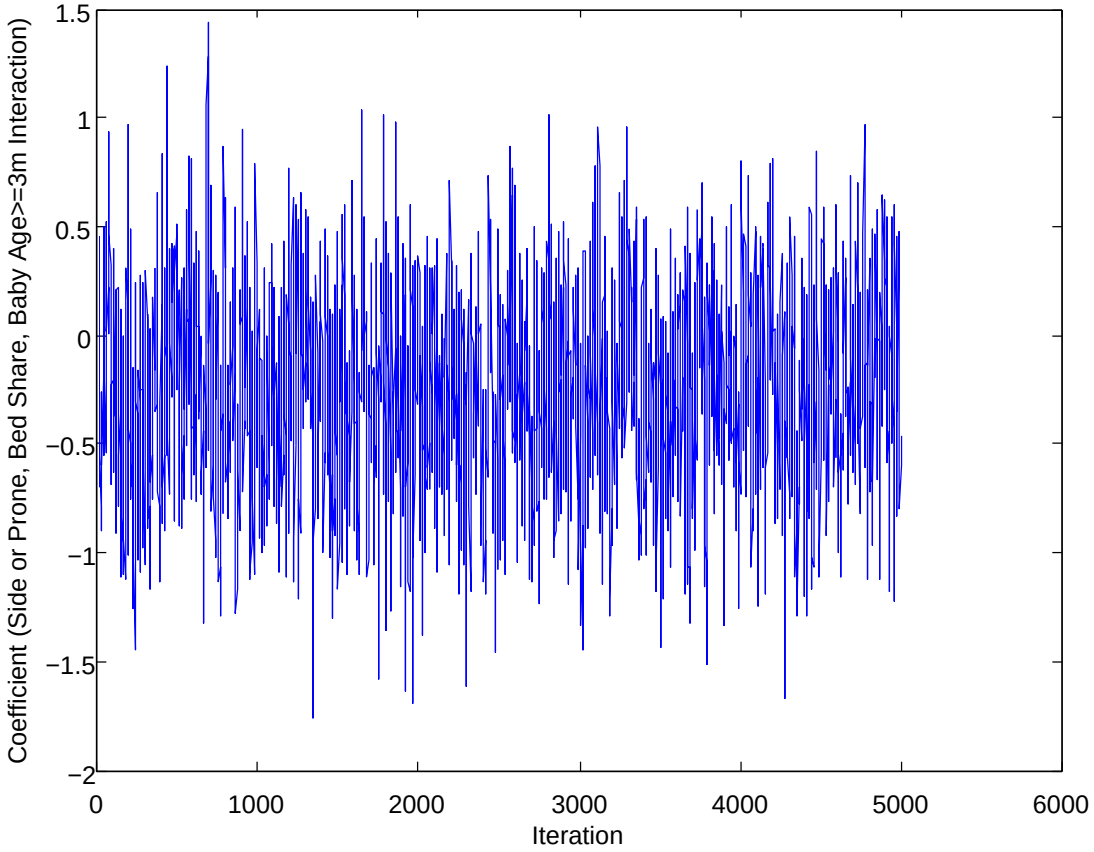

Figure 4.1: Example coefficient chain of 5,000 interactions created by the MCMC algorithm after a 50,000 burn in from the case data for the coefficient representing that a baby was left on its side/prone was bedsharing and greater or equal to 3 months old.

Figure 4.1 shows a typical 5,000 iteration chain from the MCMC sampler after 50,000 burn in iterations (other chains were similar), we can observe that the coefficient achieved approximate stability. The case and control imputation data sets were combined and arranged in a STATA *ice* format. The *mi* command was applied with the MOI to create inferences from the imputed data sets by fitting the MOI to each imputed dataset and then combing the results for the final inference using Rubin's rules (Rubin, 1976) (see section 2.13). The results of the imputation analysis are shown in Table 4.8 together with the complete records analysis.

Fixed Effects	AOR MAR (CR)	SE MAR (CR)	95% CI MAR (CR)
Bed Shared	2.42 (3.36)	0.84 (1.85)	1.23 - 4.78 (1.14 - 9.88)
Bottle Fed	1.53 (1.33)	0.15 (0.23)	1.26 - 1.85 (0.95 - 1.86)
Int A:P smokes only, no BS	1.06 (1.30)	0.15 (0.30)	0.81 - 1.38 (0.83 - 2.03)
Int A:M smokes only, no BS	1.41 (1.06)	0.21 (0.29)	1.05 - 1.89 (0.62 - 1.80)
Int A:Both smoke, no BS	2.79 (2.99)	0.33 (0.61)	2.22 - 3.51 (2.01 - 4.46)
Int A:P smokes only, BS	2.58 (1.81)	0.92 (1.00)	1.28 - 5.18 (0.61 - 5.35)
Int A:M smokes only, BS	6.24 (4.48)	2.62 (3.21)	2.73 - 14.23 (1.10 - 18.21)
Int A:Both smoke, BS	9.24 (13.43)	3.18 (7.51)	4.71 - 18.16 (4.49 - 40.19)
Age centred at 6m	1.00 (0.99)	0.00 (0.00)	1.00 - 1.00 (0.99 - 1.00)
Other Room	2.34 (2.83)	0.23 (0.49)	1.93 - 2.84 (2.02 - 3.97)
Int C:Side, baby<3m, no BS	1.50 (1.82)	0.23 (0.46)	1.10 - 2.03 (1.10 - 3.00)
Int C:Prone, baby<3m, no BS	10.34 (10.49)	1.77 (3.03)	7.39 - 14.47 (5.96 - 18.47)
Int C:Side, baby≥3m, no BS	1.37 (1.24)	0.19 (0.30)	1.04 - 1.81 (0.78 - 1.99)
Int C:Prone, baby≥3m, no BS	7.50 (8.78)	1.08 (2.05)	5.66 - 9.93 (5.55 - 13.89)
Int C:Side, baby≥3m, BS	0.37 (0.29)	0.12 (0.14)	0.20 - 0.69 (0.11 - 0.77)
Int C:Prone, baby≥3m, BS	1.68 (0.56)	0.77 (0.59)	0.68 - 4.14 (0.07 - 4.48)
Int D:Alcohol≥2 units, no BS	2.61 (2.91)	0.79 (0.82)	1.40 - 4.87 (1.67 - 5.06)
Int D:Alcohol≥2 units, BS	4.98 (5.52)	3.17 (2.54)	1.29 - 19.17 (2.23 - 13.63)
Int E:Drug use, no BS	5.53 (3.84)	4.96 (3.60)	0.93 - 33.09 (0.61 - 24.04)
Int F:Drug use, BS	6.4e+4 (3.4e+6)	4.2e+8 (3.3e+9)	0.00 - . (0.00 - .)
2000 ≤ BW <2500	1.72 (1.78)	0.16 (0.26)	1.44 - 2.06 (1.33 - 2.38)
2500 ≤ BW <3500	4.45 (5.72)	0.82 (1.83)	3.11 - 6.38 (3.06 - 10.70)
BW ≥3500	10.59 (16.19)	2.43 (6.67)	6.75 - 16.61 (7.22 - 36.32)
Married or Cohabiting	2.08 (2.53)	0.25 (0.56)	1.65 - 2.63 (1.63 - 3.92)
26 ≤ MA <31	1.99 (2.16)	0.22 (0.38)	1.61 - 2.47 (1.52 - 3.06)
21 ≤ MA <26	3.16 (3.52)	0.40 (0.74)	2.47 - 4.05 (2.34 - 5.32)
19 ≤ MA <21	8.90 (7.04)	1.72 (2.43)	6.10 - 13.00 (3.58 - 13.86)
MA <19	11.02 (16.02)	2.52 (8.59)	7.04 - 17.24 (5.60 - 45.83)
2 Live Births	2.54 (2.70)	0.28 (0.49)	2.04 - 3.15 (1.90 - 3.85)
3 Live Births	4.31 (4.83)	0.59 (1.08)	3.30 - 5.62 (3.12 - 7.48)
4 Live Births	6.09 (5.13)	1.08 (1.62)	4.31 - 8.61 (2.77 - 9.51)
5+ Live Births	8.62 (11.67)	1.69 (4.69)	5.86 - 12.67 (5.31 - 25.66)
Race	1.34 (2.80)	0.17 (1.02)	1.04 - 1.73 (1.37 - 5.72)
Male matched	1.61 (1.59)	0.18 (0.31)	1.30 - 1.99 (1.09 - 2.34)
Male unmatched	0.79 (0.75)	0.10 (0.15)	0.61 - 1.02 (0.51 - 1.10)
Constant	0.01 (0.00)	0.00 (0.00)	0.00 - 0.01 (0.00 - 0.01)

Random Effects	Estimate MAR (CR)	SE MAR (CR)	95% CI MAR (CR)
Standard Deviation Bed Shared	0.47 (0.80)	0.22 (0.34)	0.19 - 1.18 (0.35 - 1.82)
Standard Deviation Constant	0.77 (0.66)	0.15 (0.66)	0.52 - 1.13 (0.40 - 1.11)

Table 4.8: Adjusted odds ratios (AOR) and standard errors (SE) for Complete Records (CR, N=2116) and imputed (N=5627) inferences from the MOI (thus no interaction B). BS=Bed Sharing, Int=Interactions, P=Partner and M=Mother. See Table 4.2 for reference groups and interaction details.

We can see from Table 4.8 the point estimates from fitting the MOI to the imputed data are similar to the complete records. We expected to observe this as the MOI contains the covariate centre which is the key predictor of an observation being missing. By including centre as a covariate, under the assumption that the data are missing completely at random within each centre, the missingness mechanism does not depend on the case/control status (i.e. the response in the substantive model). Thus a complete records analysis should be unbiased. The largest differences can be observed in interaction F (*Mother Used Drugs After Birth* with *Bed Shared*), the extremely large OR reflects that nearly all babies were cases if they had a positive indicator for interaction F. We can also see that the standard errors are smaller in the MAR imputed data than the complete records. We have gained power mostly because MI brings back into the MOI the observed data on records with one or more missing values which were excluded from the complete record analysis.

To summarise the results, we reproduce the graphical summaries we presented in Carpenter *et al.* (2013), using our multilevel imputed data (the publication used single level imputation). We graphically represent the bed sharing SIDS risk rate by age using the AOR MOI (which includes interaction B, as discussed on page 113. To create the graph we calculated the AOR (on the log scale) for bed sharing at two week intervals up to 26 weeks adjusting for all the other covariates within the AOR MOI. For example the coefficient for 2 weeks would be calculated using the coefficients from the AOR MOI as:

$$\begin{aligned}\text{Week 2 Coefficient} = {} & \text{Bed Shared Coefficient} \\ & - ((182 - (2 * 7)) \times \text{Bed Shared Age Centred Interaction (B) Coefficient}) \quad (4.6.7)\end{aligned}$$

In (4.6.7), the value of 182 is used as *Bed Shared Age Centred Interaction* (B) is centred at 182 days (6 months); 2 as we want week 2 and 7 as there is 7 days in a week. The coefficients are then converted into AOR's on the log scale and are graphed over the 26 week period, this can be seen in Figure 4.2.

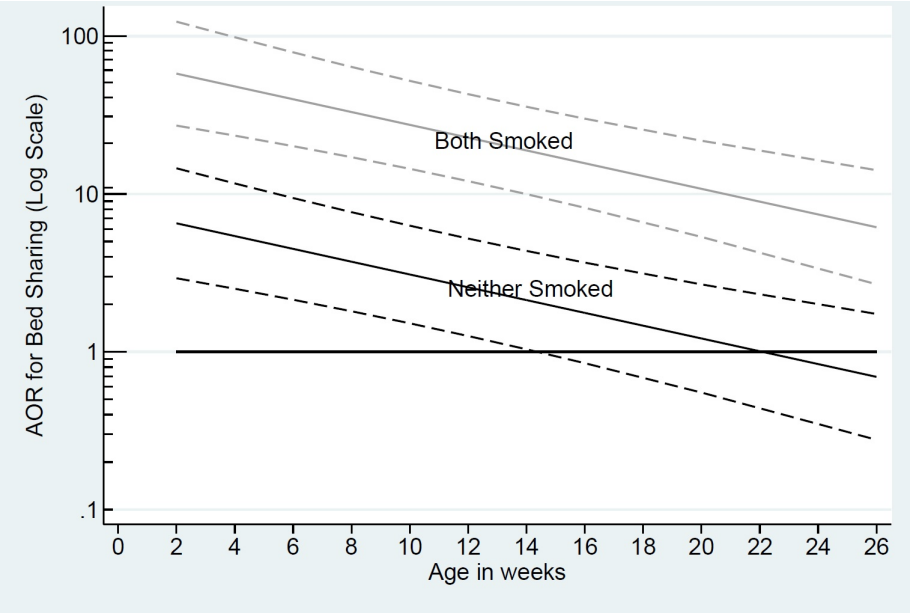

Figure 4.2: Adjusted ORs (AORs; log scale) with 95% confidence interval (dotted lines) for Sudden Infant Death Syndrome by age for bed sharing breast-fed infants, when neither parent smoked and both smoked versus comparable infants sleeping supine in the parents room. AORs are also adjusted for feeding, sleeping position when last left, where last slept, sex, race, and birth weight, mothers age, parity, marital status, *alcohol* and *drug* use. Results for data from the multilevel MI under the MAR assumption.

The results from our imputation model, which is slightly more general than that used in the analyses reported in the paper by Carpenter *et al.* (2013) are nevertheless very similar. We see in Figure 4.2 that there is a log-linear downward tend in the AOR for bed sharing in the first 6 months when neither or both parents smoke and babies of smoking parents have a higher risk of SIDS than non-smoking parents. Bed sharing is a statistically significant risk factor at the 5% level for SIDS up to approximately 14 weeks when both parents do not smoke. However,

the risk declines with age and is not significantly different after 14 weeks.

The multilevel multiple imputation process which we have applied was very computationally and methodologically complex and encountered a number of problems which we had to work hard to overcome. In particular, we believe our strategy of starting with 'full' univariate models for *alcohol* and *drug* use, then simplifying these so they could be estimated, and finally using an inverse Wishart prior to allow imputation consistent with the multilevel structure, is a statistically principled approach to imputing these data. We hope that this will be of value to other researchers faced with similar problems.

In this imputation we have assumed the missing data mechanism is missing at random. However as this is a strong untestable assumption the conclusions are contestable, as we witnessed in letters in response to the Carpenter *et al.* (2013) paper. To support our findings we need to explore the robustness of the inferences to the alternative assumption that the missing data is missing not at random. We tested the robustness through a series of pattern mixture sensitivity analyses (introduced in Section 2.20).

4.7 Sensitivity Analysis

In this section we investigate the robustness of inferences to departure from the the MAR assumption in the AOR MOI. We focused our sensitivity analysis on the cases only, the controls remained imputed under MAR. We do this to address the concerns raised by critics in response to Carpenter *et al.* (2013). The main concern raised by critics was that the *alcohol* and *drug* use data that are missing in cases may not be like the observed data (not MAR) and hence

the risk of bed sharing could be explained by *alcohol* and/or *drugs* if they are imputed under an MNAR mechanism.

We applied two types of sensitivity analysis to investigate the sensitivity of the inferences to the MAR missing data assumption in the imputation of *alcohol* and *drug*. The first looked at congenial sensitivity analysis (see Section 2.22) through a pattern mixture approach using a prior on the level 1 coefficients in the imputation model and their respective variances. In the second we applied an uncongenial sensitivity analysis to see how robust the inferences are to extreme departures from the MAR assumption. In a practical situation a researcher may wish to begin with the uncongenial sensitivity analysis that we describe later in this chapter. The uncongenial sensitivity analysis (as described) was far easier and quicker to implement than the congenial sensitivity analysis. If inferences are robust to extreme uncongenial sensitivity analysis settings then they will hold for weaker congenial settings. However if inferences from the extreme uncongenial sensitivity settings are not robust, then the researcher should also implement more realistic congenial sensitivity analysis. These could be informed by prior elicitation, as in the previous chapter.

The congenial sensitivity analysis was applied using a pattern mixture approach through multilevel multiple imputation. The approach we used was to choose a prior which takes the association between *alcohol* and *drug* with bed sharing seen in the observed data and alters it, so that the imputed data has a different association with bed sharing, hence a MNAR mechanism. We focus on bed sharing as the focus of the Carpenter *et al.* (2013) was to assess the associated risk of bed sharing with SIDS, thus by focusing on bed sharing we can see how robust our main bed sharing inference is to the departures from the MAR assumption.

To apply the approach needed, we had to modify the REALCOM software in MATLAB. The REALCOM software uses latent normal variables (equivalent to probit models) to handle binary variables. Hence the informative prior for imputing *alcohol* or *drug* needs to be applied on the probit scale in the software. In particular our approach is applied to the latent normal models within (4.6.1) through a multivariate normal conjugate prior distribution, where the prior is applied to bed sharing coefficients in the latent normal imputation model. Let μ represent the posterior mean, Λ the posterior precision, μ_d the Probit (*bed shared*) coefficient from the data with Λ_d associated precision and μ_p as the Probit coefficient prior (for *bed shared*) with Λ_p associated prior precision. The multivariate normal conjugate prior distribution is as follows:

$$\mu = (\Lambda_p + \Lambda_d)^{-1}(\Lambda_p\mu_p + \Lambda_d\mu_d) \tag{4.7.1}$$

$$\Lambda = (\Lambda_p + \Lambda_d) \tag{4.7.2}$$

By applying the prior we change the parameters of the imputed model and hence the imputed values so the missing data distribution used to impute the data is no longer the same as the observed data distribution. This creates a Missing Not At Random (MNAR) mechanism. After the data were imputed the AOR MOI is fitted to the new data sets in turn, the results are combined using Rubin's rules and the inferences are compared to the MAR inferences.

To apply the uncongenial sensitivity analysis we impute the data under different, extreme, MNAR scenarios. The extreme sensitivity analysis settings involved changing all the MAR imputed values for certain groups of cases/controls to take a particular value. However, the remaining missing values stay multiply imputed under MAR. We created 10 scenarios and applied them to *alcohol* and *drug* use. For example, the 10 *alcohol* scenarios and descriptions are shown in Table 4.9.

Scenario Reference	Set up (all other imputed values retain MAR values)
A0	All imputed, alcohol=0
A1	All imputed, alcohol=1
A0Ca	All imputed cases, alcohol=0
A1Ca	All imputed cases, alcohol=1
A0CaB	All imputed bed sharing cases, alcohol=0
A1CaB	All imputed bed sharing cases, alcohol=1
AB0	All imputed bed sharing, alcohol=0
ABCa1Co0	All imputed bed sharing: cases alcohol=1, controls alcohol=0
ABCa0Co1	All imputed bed sharing: cases alcohol=0, controls alcohol=1
AB1	All imputed bed sharing, alcohol=1

Table 4.9: Description of extreme sensitivity analysis settings. *Alcohol*=0 represents less than 2 units drunk within the observation 24 hours, *alcohol* = 1 equal or more than 2 units drunk.

The 10 imputed data sets created in each scenario are each fitted with the AOR MOI, and the results are combined using Rubin's rules. Then the inferences are compared to those from the MAR imputed data to explore robustness of the results.

The main focus of our sensitivity analysis is to explore how robust the calculated associated risk between bed sharing and SIDS is in infants with parents that both smoke and parents that neither smoke evaluated over time. We apply both the uncongenial and congenial sensitivity analysis approaches separately in the *alcohol* and *drug* models. We begin with the congenial sensitivity analysis of the *alcohol* imputation.

4.8 Congenial Sensitivity Analysis, Imputing Alcohol Under MNAR

To apply the congenial sensitivity analysis through the multivariate normal prior we must first take the Probit coefficient for *bed shared* from the MAR imputation model μ_d and the associ-

ated precision Λ_d. The MAR imputed *bed shared* Probit coefficient was -0.47 (4.58 precision) which is approximately -0.78 (2.75 precision) on the Logit scale. We began by looking at doubling the *bed shared* coefficient on the Logit scale and halving its variance for the *alcohol* imputation model (*drug* imputations remain MAR, unchanged). Though this is arbitrary, it does represent a substantially stronger association than that seen in the observed data; a doubled coefficient is arguably at the extreme end of what we would see in practice.

We thus require the posterior *bed shared* coefficient to be $\mu = -1.55$ (Logit scale) and have variance of $\Lambda = 5.49$ (Logit scale). By taking (4.7.1) and (4.7.2), we can rearrange and substitute values to find the required prior to give us the requested *bed shared* posterior values:

$$\Lambda_p = (\Lambda - \Lambda_d) = 2.75 \tag{4.8.1}$$

$$\mu_p = \frac{\mu(\Lambda_p + \Lambda_d) - \Lambda_d \mu_d}{\Lambda_p} = -2.33. \tag{4.8.2}$$

This prior will increase the probability of being imputed as drinking less than 2 units of alcohol as it increases the *bed shared* coefficient and reduces it variance in the latent *alcohol* model which REALCOM uses:

$$\text{Probit P(baby } i \text{ from centre } j \textit{ alcohol}{=}0) = \boldsymbol{X}_{ij}^T \boldsymbol{\beta} + \boldsymbol{Z}_{ij}^T w_j. \tag{4.8.3}$$

The MNAR imputation model produced a *bed shared* Logit coefficient of -1.56 (5.63 variance), which is close to the requested posterior (-1.55 (5.49 variance)). The original case/control MAR imputed data had on average 5264(93.54%) as drinking less than 2 units of alcohol over all 10 imputed data sets. The MNAR prior increased this amount marginally to 5269 (93.64%). To test the MNAR imputation prior method we continued by applying a different prior which

halved the coefficient ($\mu = -0.39$ Logit scale) and quartered the variance ($\Lambda = 10.99$ Logit scale). The final case and control MNAR imputed dataset had on average 5268(93.62%) as drinking less than 2 units of alcohol which is also an increase from the MAR frequency of 5264(93.54%). This suggested that the prior was not working how we envisaged, the prior was altering the coefficient to the specified posterior but this was not having the hypothesised effect on the proportions of *alcohol* use. Closer investigation showed the level 2 covariance matrix was increasing to reflect the increased variance being introduced to the centres where *alcohol* was mostly observed. This was over-riding our intention with the prior.

To overcome this problem we created an indicator variable for the three centres which had approximately no *alcohol* observations (centres 16, 17 and 19 see Table 4.4). An interaction between this indicator and *bed shared* (called the *indicator interaction*) was included in the *alcohol* model (4.8.3). The prior was then applied to the *indicator interaction* and not the *bed shared* coefficient. The *indicator interaction* coefficient will only add or subtract from the model when a baby shared the bed and came from either centre 16, 17 or 19. This creates the MNAR mechanism as it changes the contribution of *bed sharing* to the *alcohol* model, specifically changing it from the MAR contribution only in centres were the data was generally not observed.

As the *indicator interaction* has a linear predictor using a Probit link with the value and variance of zero in the initial MAR imputation, we do not require equations (4.8.1) and (4.8.2) as the posterior will be equal to the prior. The advantage of this method is we can keep the variance the same as the MAR imputation and only change the *bed shared* mean (something equations (4.8.1) and (4.8.2) would not allow in the previous setting). This makes life easier because we do not have to hypothesise a suitable MNAR variance.

We applied the *indicator interaction* prior with different mean values for the linear predictor on the Probit scale (*Probit prior mean*) separately four times within the *alcohol* imputation model (cases only, controls remain MAR). The following scenarios were investigated:

- Posterior of *indicator interaction Probit prior mean* equal to 2.0 with 0 variance. This will increase the number of *alcohol* observations imputed as drinking less than 2 units of alcohol.
- Posterior of *indicator interaction Probit prior mean* equal to 0.5 with 0 variance. This will also increase the number of *alcohol* observations imputed as drinking less than 2 units of alcohol but to a lesser extent than the first scenario.
- Posterior of *indicator interaction Probit prior mean* equal to -0.5 with 0 variance. This will decrease the number of *alcohol* observations imputed as drinking less than 2 units of alcohol.
- Posterior of *indicator interaction Probit prior mean* equal to -2.0 with 0 variance. This will decrease the number of *alcohol* observations imputed as drinking less than 2 units of alcohol, even greater than the previous scenario.

The 4 scenarios were applied in the same manner as the MAR imputation using 10 separate runs of the software (different starting seeds used) applied with a burn in of 50,000 iterations outputting a dataset after another 5,000 iterations. The data sets were then combined and merged back with the MAR imputed controls. The average proportions of *alcohol* from the 4 scenarios over the 10 imputed data with the controls was calculated to check the prior was creating the desired effect. The proportions are shown in Table 4.10.

Probit prior mean (Variance=0)	Logit scale	Mother Alcohol<2 Units N(%)	Mother Alcohol≥2 Units N(%)
2.0	3.3	5295(94.09)	332(5.91)
0.5	0.8	5271(93.66)	357(6.34)
MAR (No Prior)		5264(93.54)	363(6.46)
-0.5	-0.8	5258(93.44)	369(6.56)
-2.0	-3.3	5202(92.45)	425(7.55)

Table 4.10: Average frequency and percentages of imputed *alcohol* values over all 10 imputed datasets merged with the MAR control data using different *indicator interaction* priors. The *indicator interaction* priors are shown on the Probit and Logit scale.

The first two rows of Table 4.10 (*Probit prior mean* 2.0 and 0.5) will reduce the probability of imputing *alcohol* use when the baby *bed shared* and belonged in one of the three specified centres (centre 16, 17 and 19). The last two rows (*Probit prior mean* -0.5 and -2.0) will affect the *alcohol* model (4.8.3) similarly, but resulting in an increase in the probability of imputing *alcohol* use. It is clear that the *indicator interaction* prior is having the derived effect. As the *Probit prior mean* increases, the proportion of imputed low/non *alcohol* drinkers also increases. As the *Probit prior mean* decreases, we observed a decrease in the proportion of low/non *alcohol* drinkers. We can observe that the overall proportions change fairly modestly as we move away from imputing under MAR.

To investigate if the main conclusions from our Carpenter *et al.* (2013) paper are robust, we apply the AOR MOI to the newly imputed MNAR data and plot the corresponding graph. We started with the positive *Probit prior means* in the *indicator interaction* priors, which decreases the probability of drinking more than 2 units of alcohol. The AOR graph for an *indicator interaction Probit prior mean* of 2.0 is shown in Figure 4.3.

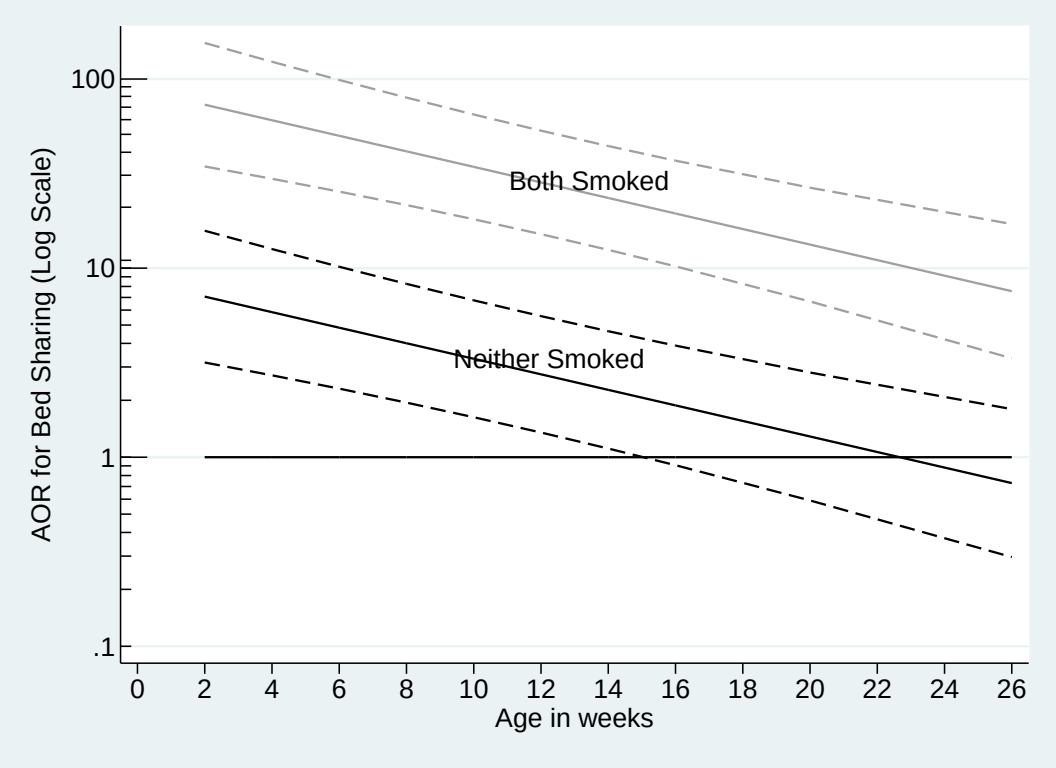

Figure 4.3: Adjusted ORs (AORs; log scale) with 95% confidence interval (dotted lines) for Sudden Infant Death Syndrome by age for bed sharing breast-fed infants, when neither parent smoked and both smoked. Imputed with an MNAR bed share, centre *indicator interaction* prior with a *Probit prior mean* of 2.0 and variance of 0.

By comparing Figure 4.3 with Figure 4.2, we can observe that the MNAR mechanism has not changed the inferences greatly. The risk of SIDS has slightly increased for both the smoking and none smoking groups. Closer investigation shows that the adjusted odds ratio for *bed shared* at 6 months in the AOR MOI (4.3.1) has increased towards one (MAR AOR = 0.69 to MNAR AOR = 0.73), thus increasing the risk of bed sharing (the standard errors remain similar).

Under the MNAR *Probit prior mean* of 2.0 we create more observations in the cases where the

mother is drinking less than 2 units when bed sharing than the MAR data. This has the effect of reducing (MAR AOR = 5.1 (SE 3.4) to MNAR AOR = 2.5 (SE 2.1)) the risk associated between bed sharing and drinking (interaction D). This results in an increase in risk with bed sharing and SIDS. As the risk increase is small it suggests the inferences are not sensitive to an MNAR mechanism which increases the probability of being imputed as drinking less than 2 units of alcohol (see Appendix C for the results with the weaker indicator prior with $\mu = 0.5$).

We continued by applying the AOR MOI to the data created by the two negative indicator prior means ($\mu = -0.5$ and -2.0). The AOR graph for $\mu = -2.0$ is shown in Figure 4.4.

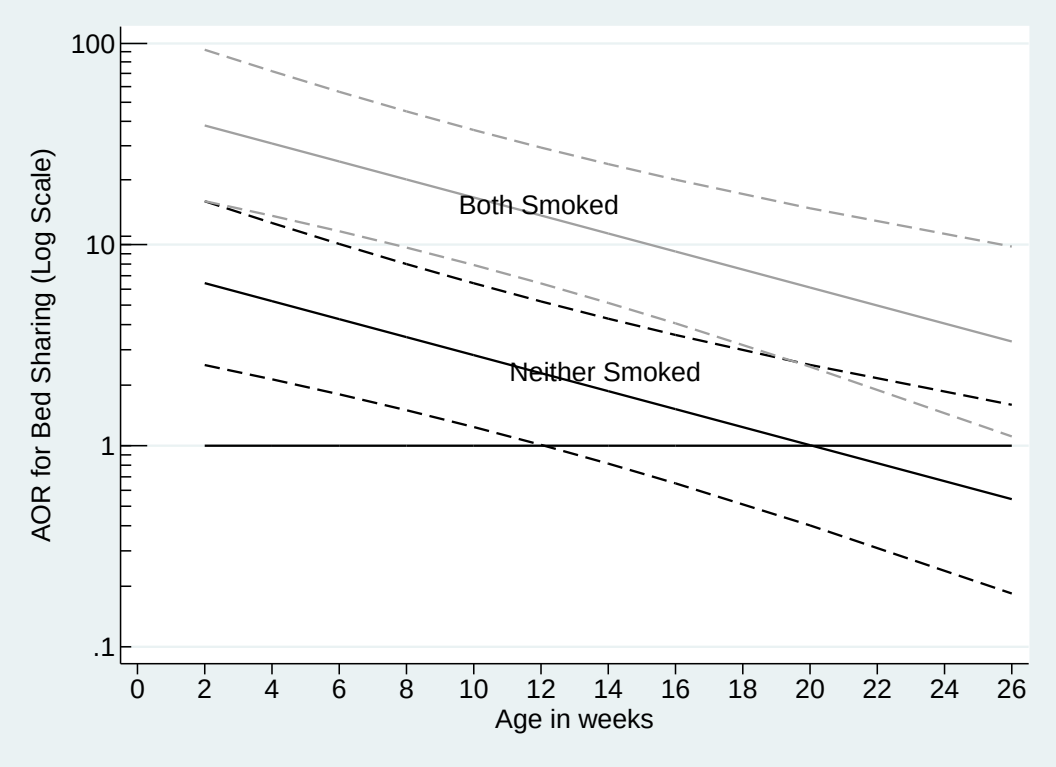

Figure 4.4: Adjusted ORs (AORs; log scale) with 95% confidence interval (dotted lines) for Sudden Infant Death Syndrome by age for bed sharing breast-fed infants, when neither parent smoked and both smoked. Imputed with an MNAR bed share, centre *indicator interaction Probit prior mean* of -2.0 and variance of 0

Comparing Figure 4.4 and Figure 4.2, we can observe the risk of *bed sharing* is decreased slightly. When the *indicator interaction Probit prior mean* is -2.0 we are increasing the probability of the mother of a case being imputed as drinking 2 or units of alcohol. The *bed shared* AOR has decreased from 0.69 MAR to 0.54 MNAR (the standard errors are similar). The reason for this is because the association between alcohol and bed sharing (interaction D) AOR in the AOR MOI has increased greatly from 5.11 MAR to 19.33 MNAR, this reduces the direct risk between *bed shared* and SIDS (increases protective effect). Figure 4.4 and Figure 4.2 are very similar, we thus conclude the AOR MOI inferences appear to be robust to the MNAR mechanism (*indicator interaction Probit prior mean* of -2.0) applied to the imputation of *alcohol* values in cases.

In this section we have applied congenial sensitivity analysis through pattern mixture using a multivariate normal prior using 4 scenarios to the *alcohol* model. The inferences in the AOR MOI appear to be robust to the departure of MAR in our scenarios, we now move on to investigate the imputation of *drug* using a similar method.

4.9 Congenial Sensitivity Analysis, Imputing Drug Under MNAR

We use the same approach as the last section in the *drug* model, however unlike the *alcohol* MNAR imputation we can only apply an *indicator* prior based on *centre* and not on the *bed shared* selected *centre* interaction, as *bed shared* is not a covariate in the *drug* imputation model after it was simplified to be fitted (see Table 4.7). The original coefficient for the constant produced by the MAR imputation in the *drug* imputation model was 3.98 (8.48 precision) on the Logit scale. We again applied the same four *indicator* priors which we used on the *alcohol*

model, the first two (μ=2.0 and μ=0.5) will have the effect of adding to the constant for the three specified centres. The last two have the opposite effect of subtracting from the constant (μ=-0.5 and μ=-2.0) for the specified centres, the variance of the indicator priors will be zero ($\Lambda = 0$). The resulting proportions of *drug* imputed over all 10 imputations for the cases and controls (controls remained unaltered) are shown in Table 4.11

Probit prior mean (Variance=0)	Logit scale	Did Not Take Drugs N(%)	Took Drugs N(%)
2.0	3.3	5608(99.66)	19(0.34)
0.5	0.8	5600(99.51)	27(0.49)
MAR (No Prior)		5600(99.51)	28(0.49)
-0.5	-0.8	5578(99.14)	49(0.86)
-2.0	-3.3	5360(95.26)	267(4.74)

Table 4.11: Average frequency and percentages of imputed *drug* values over all 10 imputed datasets merged with the MAR control data using different *indicator* priors. The *indicator* priors are shown on the Probit and Logit scale.

We can observe in Table 4.11 that if we add to the imputation model (by setting the *interaction* prior to 0.5 or 2.0), we increase the probability of being imputed as a non drug user. Likewise if we subtract (by setting the *interaction* prior to -0.5 or -2.0) from the constant we decrease the probability of being a non drug user. This suggests the *indicator* prior is working as expected. We continued by fitting the AOR MOI to the MNAR imputed data to investigate the robustness of the inferences to the MAR assumption. We began with the addition priors ($\mu = 2.0$ and $\mu = 0.5$), we see the $\mu = 2.0$, $\Lambda = 0$ prior in Figure 4.5.

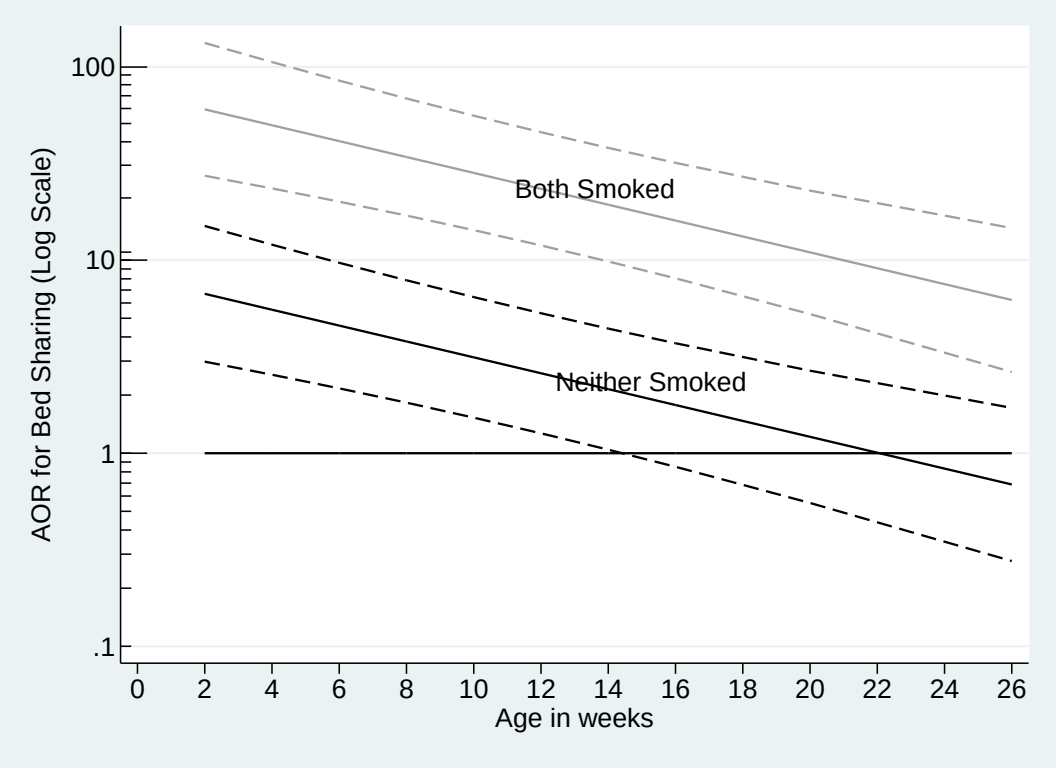

Figure 4.5: Adjusted ORs (AORs; log scale) with 95% confidence interval (dotted lines) for Sudden Infant Death Syndrome by age for bed sharing breast-fed infants, when neither parent smoked and both smoked. Imputed with an MNAR indicator prior ($\mu = 2.0$, $\Lambda = 0$).

We can see Figure 4.5 and Figure 4.2, look very similar. The AOR's have changed very little, *bed shared* and *bed shared age* interactions (B) are approximately the same as the MAR and the AOR for both parents smoking has increased slightly from 8.88 under MAR to 9.02 under MNAR (standard errors approximately the same). This effect is also observed when the *indicator* prior is weaker $\mu = 0.5$ (see Appendix C). We continued by applying the AOR MOI to the substraction *indicator* prior which increases the number of mothers imputed as taking drugs, see Figure 4.6 for when $\mu = -2.0$ and $\Lambda = 0$.

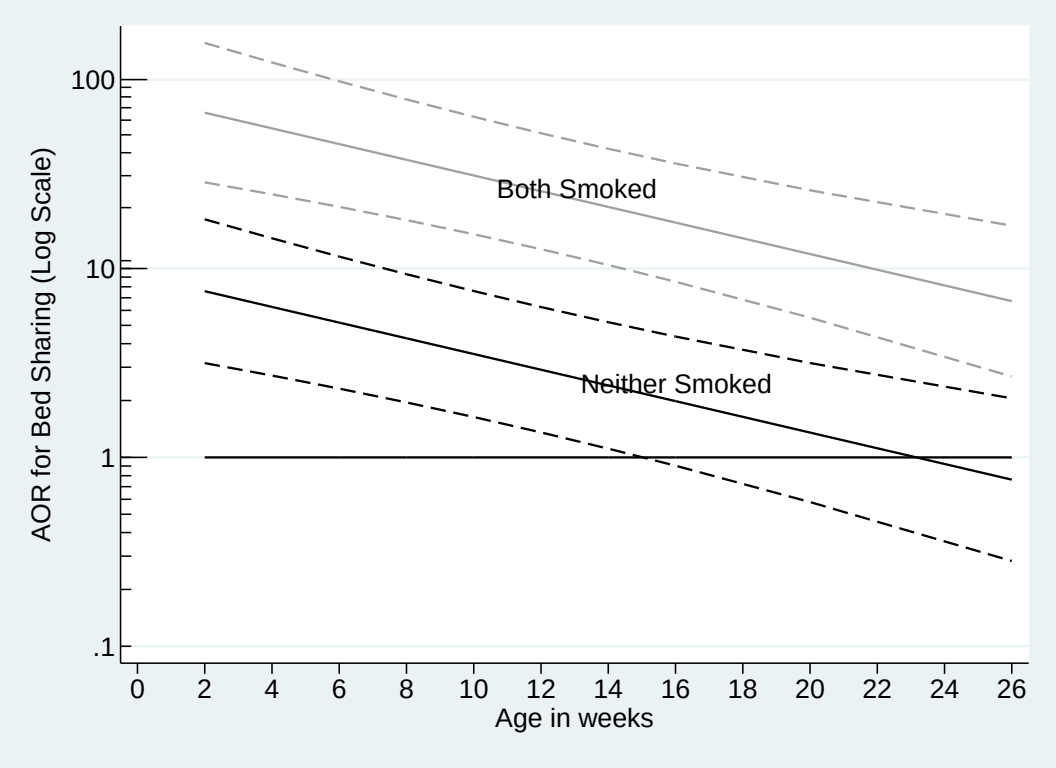

Figure 4.6: Adjusted ORs (AORs; log scale) with 95% confidence interval (dotted lines) for Sudden Infant Death Syndrome by age for bed sharing breast-fed infants, when neither parent smoked and both smoked. Imputed with an indicator MNAR prior ($\mu = -2.0$, $\Lambda = 0$).

Again by comparing Figure 4.6 with Figure 4.2, we can see the MNAR mechanism has not changed the SIDS risk greatly when *bed shared*. There is an increase in SIDS risk created by a increase in the *bed shared* AOR (MAR = 0.69 to MNAR = 0.76 standard errors approximately the same) but this is still approximately significant to 14 weeks for the 'neither smoke' group. This suggests the inferences are robust in the AOR MOI model to the departure of MAR in the imputation of *drug* under our congenial sensitivity analysis scenarios. Next we look at sensitivity analysis by considering extreme settings which break the congeniality between the imputation models and the AOR MOI.

4.10 Uncongenial Sensitivity Analysis, Imputing Alcohol Under MNAR

We now move on to discuss the uncongenial sensitivity analysis we applied to the *alcohol* model, which seeks to assess how robust the inferences are to extreme departures from the MAR assumption. The method is uncongenial sensitivity analysis as the (implicit) imputation model is inconsistent with the linear trend assumption of the AOR MOI. We applied the 10 MNAR scenarios detailed in Table 4.9 to the MAR imputed data, Again imputed values that are not altered, retain their MAR imputed values. The average proportion of observations imputed to each *alcohol* category over all 10 imputed data sets can be observed in Table 4.12.

Scenario Reference	Mother Alcohol<2 Units N(%)	Mother Alcohol≥2 Units N(%)
MAR	5264(93.54)	363(6.46)
A0	5447(96.80)	180(3.20)
A1	1937(34.42)	3690(65.58)
A0Ca	4554(80.23)	1073(19.07)
A1Ca	5344(94.97)	283(5.03)
A0CaB	5156(91.63)	471(8.37)
A1CaB	5295(94.11)	332(5.89)
AB0	5310(94.37)	317(5.63)
ABCa1Co0	5171(91.90)	456(8.10)
ABCa0Co1	5061(89.95)	566(10.05)
AB1	4922(87.48)	705(12.52)

Table 4.12: The mean frequency and percentages of imputed *alcohol* values from different scenarios (described in Table 4.9) over 10 imputed cases/control datasets.

We see from Table 4.12, that the proportion of mothers imputed to drinking less than 2 units ranges from approximately 34% to 97% over the cases and controls. To see the effect of changing the proportions in the *alcohol* categories we applied the AOR MOI of interest to the data separately for each scenario. We graphed each covariate for each scenario, in Figure 4.7

we can see the results for the *bed shared* covariate (all covariate graphs may be observed in Appendix C).

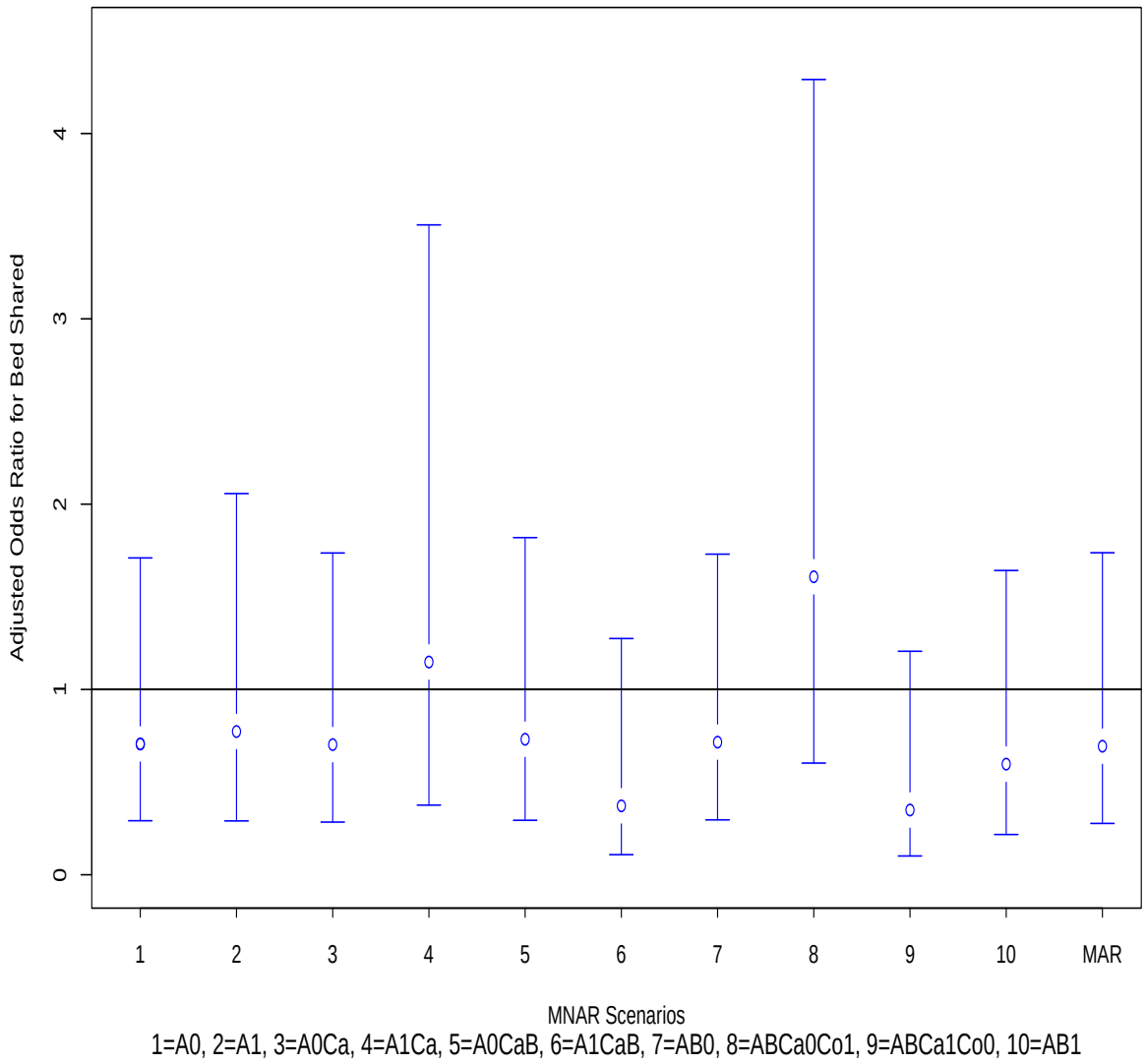

Figure 4.7: Adjusted odds ratio's at 6 months with confidence intervals from the adjusted odds ratio model of interest for the *bed shared* coefficient from the 10 *alcohol* scenarios and the MAR results.

We see in Figure 4.7 that the $ABCa0Co1$ setting affects the *bed shared* covariate the most. The adjusted odd ratio is larger than the MAR confidence interval, however the MAR confidence still contains the point estimate suggesting that the difference is not significant. The AOR's for scenarios $A0$, $A1$, $A0Ca$, $A0CaB$, $AB0$ and $AB1$ remain similar to the MAR AOR. Scenario $A1Ca$ which imputes all cases to drinking more than two units (controls remain MAR) and scenario $ACa0Ca1$ have changed the AOR *bed sharing* so that it is no longer a protective effect for SIDS as seen in the MAR imputation. The reason why the *bed shared* AOR is quite robust to the extreme MNAR scenarios is because of the interaction covariate *bed shared* with *alcohol* absorbing the most of the MNAR impact. The *bed shared* with *alcohol* fluctuates from the MAR value of 5.11, between 0.05 for the $ABCa0Co1$ scenario to 131.40 in the $ABCa1Co0$ (see Appendix C). Other variables also fluctuate absorbing the MNAR changes to the imputation making the *bed sharing* covariate reasonably stable between scenarios.

We examined the effect of the largest departure from the MAR AOR for *bed shared* (which we can see in Figure 4.7 was the $ABCa0Co1$ setting) on our main inference *bed sharing* risk on SIDS, using the AOR graph technique seen in Figure 4.2. The graph can be seen in Figure 4.8.

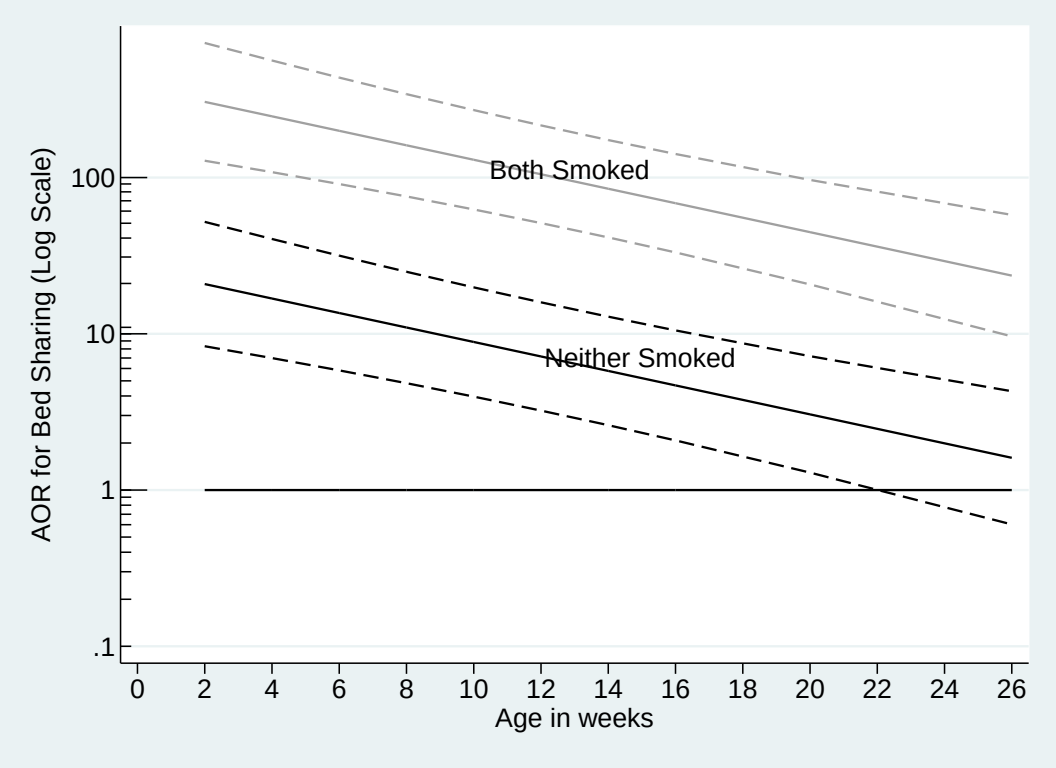

Figure 4.8: Adjusted ORs (AORs; log scale) with 95% confidence interval (dotted lines) for Sudden Infant Death Syndrome by age for bed sharing breast-fed infants, when neither parent smoked and both smoked. Imputed with the extreme sensitivity analysis setting when all controls that bed shared are imputed as drinking 2 or more units of alcohol and all bed sharing cases are imputed as drinking less than 2 units, non bed sharers retain MAR imputations ($ABCa0Co1$).

By comparing Figure 4.8 with Figure 4.2, we can observe that the MNAR mechanism has increased the association between bed sharing and the risk of SIDS. When neither parents smoked the time when the AOR is significant has increased from approximately 14 weeks under MAR to 22 weeks. The AOR shows that babies with smoking mothers and partners still have a higher risk of SIDS even when imputed with an MNAR mechanism. The extreme sensitivity analysis suggests that the inferences are robust to even extreme departures from MAR in the *alcohol* imputation.

4.11 Uncongenial Sensitivity Analysis, Imputing Drug Under MNAR

We now apply the uncongenial sensitivity analysis to the *drug* model using similar extreme scenarios as described in Table 4.9 (scenario references changed i.e. $ABCa0Co1$ becomes $DBCa0Co1$). We can observe the effect of the scenarios on *drug* category proportions in Table 4.13.

Scenario Reference	Did Not Take Drugs N(%)	Took Drugs N(%)
D0	5611(99.72)	16(0.28)
D1	2142(38.07)	3485(61.93)
D0Ca	4829(85.82)	798(14.18)
D1Ca	5608(99.66)	19(0.34)
D0CaB	5466(97.13)	161(2.87)
D1CaB	5602(99.55)	25(0.45)
DB0	5602(99.56)	25(0.44)
DBCa1Co0	5466(97.14)	161(2.86)
DBCa0Co1	5359(95.24)	268(4.76)
DB1	5223(92.82)	404(7.18)
MAR	5600(99.51)	27(0.49)

Table 4.13: Mean frequency and percentages of imputed *drug* values from different scenarios (similar to those described in Table 4.9) over all 10 imputed datasets.

We can see in Table 4.13 that proportion of mothers not taking illegal drugs ranges from approximately 38% to 100%. Again we apply the AOR MOI to each scenario separately and graphed each covariate for each scenario. In Figure 4.9 we can see the results for the *bed shared* covariate (other covariate graphs may be observed in Appendix C).

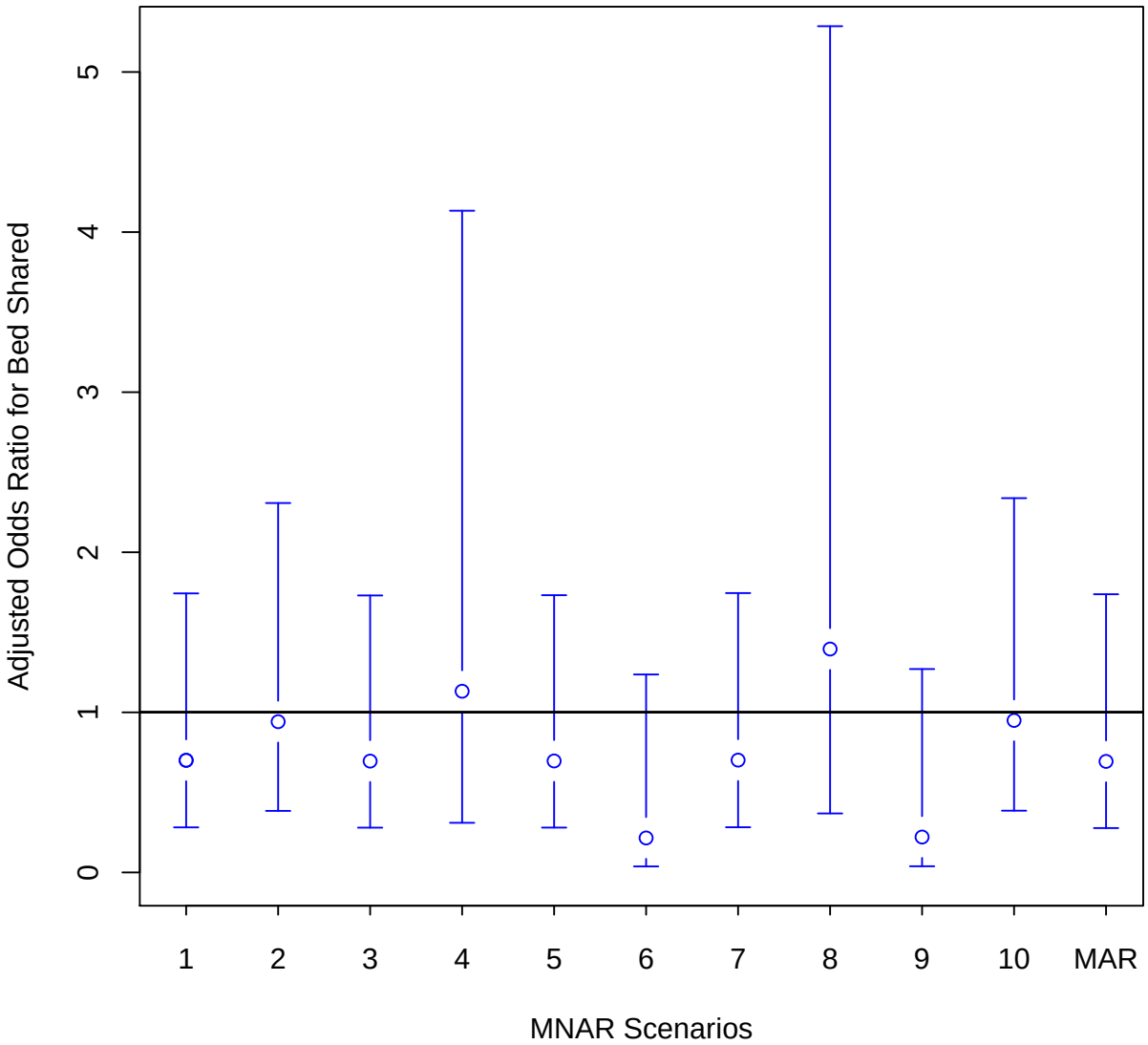

Figure 4.9: Adjusted odds ratio's at 6 months with confidence intervals from the adjusted odds ratio model of interest for the *bed shared* coefficient from the 10 *drug* scenarios and the MAR results.

We see in Figure 4.9 $DBCa0Co1$ setting affects the *bed shared* covariate the greatest changing

the AOR from 0.69 under MAR to 1.40. Both the $DBCa0Co1$ and $D1Ca$ scenarios have removed the protective affect of bed sharing. Scenarios $D1CaB$ and $DBCa1Co0$ have increased the protective effect of bed sharing outside of the MAR AOR confidence interval. We investigate the effect of the changing the *bed shared* covariate graphically, looking at the $DBCa0Co1$ setting first as it created the largest differences and then $DBCa1Co0$ second as it lowered the AOR the greatest. Figure 4.10 shows the first setting when all bed sharing cases are imputed non drug users and all bed sharing controls as drug users, non-bed sharing babies retain their MAR values.

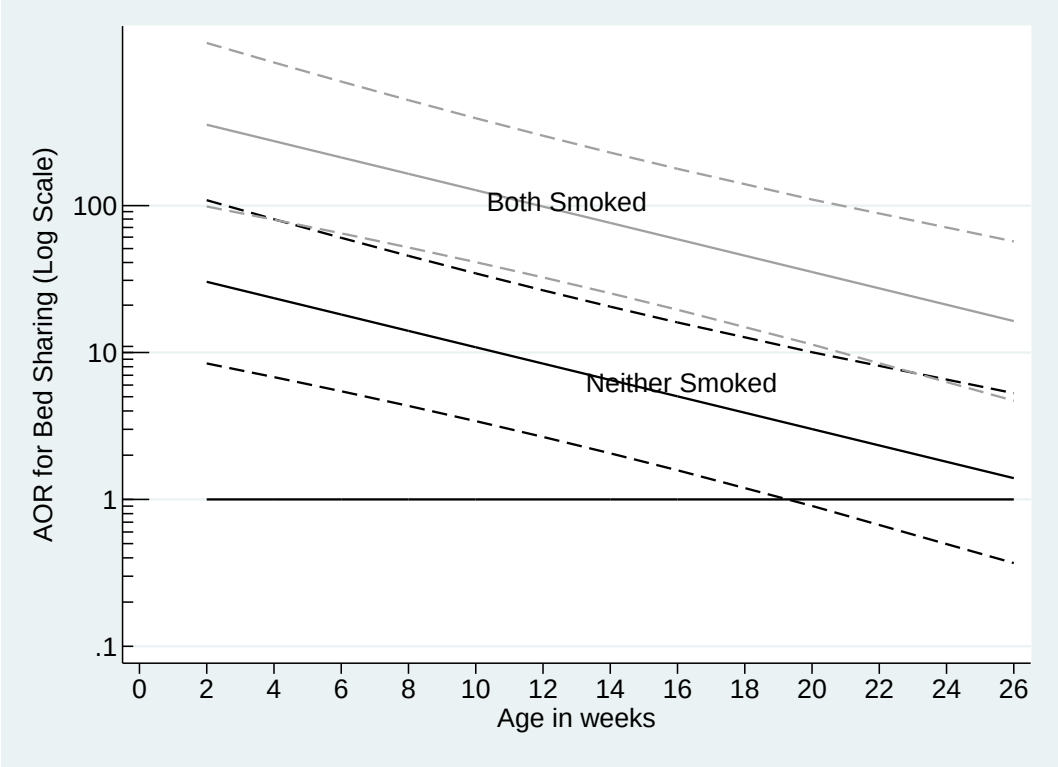

Figure 4.10: Adjusted ORs (AORs; log scale) with 95% confidence interval (dotted lines) for Sudden Infant Death Syndrome by age for bed sharing breast-fed infants, when neither parent smoked and both smoked. Imputed with the extreme sensitivity analysis setting where bed sharing cases were imputed as non-drug users and bed sharing controls were imputed as drug users, non-bed sharing babies retained MAR values ($DBCa0Co1$).

We see by comparing Figure 4.10 to the MAR Figure 4.2, that the risk of SIDS has increased. When neither parents smoked the time when the AOR is significant has increased from approximately 14 weeks under MAR to 19 weeks. The general downward trend of SIDS risk in both groups is similar to the MAR trend.

Next we looked at the $DBCa1Co0$ when all bed sharing cases are imputed as drug users, bed sharing controls are imputed as non-drug users and non-bed sharing babies retain their MAR values. The graph can been seen in Figure 4.11.

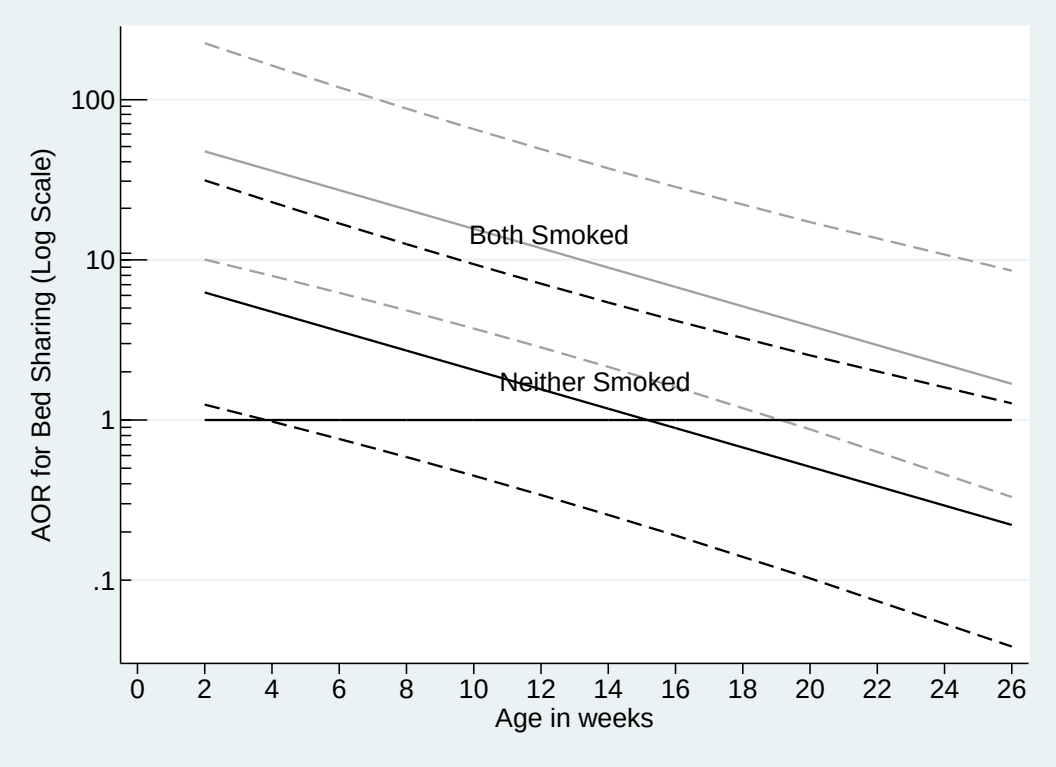

Figure 4.11: Adjusted ORs (AORs; log scale) with 95% confidence interval (dotted lines) for Sudden Infant Death Syndrome by age for bed sharing breast-fed infants, when neither parent smoked and both smoked. Imputed with the extreme sensitivity analysis setting where bed sharing cases were imputed as drug users and bed sharing controls were imputed as non-drug users, non-bed sharing babies retained MAR values ($DBCa1Co0$).

We can clearly see that the risk of bed sharing has decreased in both smoking and non-smoking groups compared to the MAR AOR graph (Figure 4.2). The risk is now only significant for approximately the first three weeks in the non-smoking group, however the general downward trend of SIDS risk in both groups is similar to the MAR trend. The two smoking categories confidence intervals have now crossed suggesting there is a reduced effect of parental smoking in this setting.

By comparing Figure 4.10 and Figure 4.11 we can see the covariate changes in the AOR MOI have increased and decreased the SIDS risk greatly. The time period when the non smoking parents risk is significant, ranges from approximately 3 weeks to 22 weeks. It is important to note, that even in these very unlikely extreme scenarios we still observe bed sharing as a risk for SIDS in both smoking groups. Thus our main conclusion that bed sharing is a risk for SIDS when the baby is young in the Carpenter *et al.* (2013) paper, is robust to all but the most extreme departures from the MAR assumption.

4.12 Discussion and Conclusions

In this chapter we have investigated the robustness of inferences to both congenial sensitivity analysis though multivariate priors and uncongenial sensitivity analysis using extreme settings. We saw the MNAR effect on the *drug* imputation was larger than the *alcohol* imputation. We begin by discussing the study limitations, and then our findings in relation to the data, we then move on to discuss the methodological conclusions.

Study Conclusions

All five studies used in the Carpenter *et al.* (2013) analysis were case control studies. While case control studies have their limitations — not least the difficulty in identifying the appropriate controls, given sudden infant death syndrome is relatively rare, they are the only practical approach. Due to the sensitive nature of infant death, great care needs to be taken over the way that questions about alcohol and drug use are asked, particularly for cases. Therefore, even in situations where such data are relatively complete, it may well be useful to conduct sensitivity analyses to systematic bias in the reporting of alcohol and drug use. Such analyses could be performed in a number of ways. However a relatively natural approach would be to apply the method developed here, but now imputing both observed and missing data assuming missing not at random.

This chapter has used the same AOR MOI as Carpenter *et al.* (2013). The model incorporates the age of the child as a linear term and hence our imputation and sensitivity analysis have also used a linear term. The original paper describes how the model was derived, and in particular looks into the data in some detail to show that the Logit-linear model for age is not an "out of sample" extrapolation. A personal communication with Carpenter (2015), confirmed that he found no evidence of a non-linear relationship between age and bed-sharing. Overall, the model fit in the original paper was very satisfactory (see Carpenter *et al.* (2013), appendix). The focus of the discussion following the publication of the paper was whether it was appropriate to impute the missing alcohol and drug data, and whether the assumption of missing at random made in the paper was driving the risk associated with bed sharing. This has therefore been the focus of the sensitivity analysis.

First, though we had to perform multilevel multiple imputation for the missing data under MAR, consistent with the multilevel structure of the data. As described in Section 4.2, this was not straightforward, because of drug use in particular is relatively rare. Multilevel multiple imputation showed a risk of bed sharing, for both smokers and non smokers for babies under 14 weeks, as summarised in Figure 4.2.

Following this, we applied the congenial sensitivity analysis by placing priors on the relevant parameters in the imputation model. This required the REALCOM software to be adapted to incorporate the prior. We focused on the cases to assess the sensitivity of the AOR MOI inferences to the MAR assumption. In particular, we wanted to asses whether the risk of bed sharing was maintained if the probability of alcohol use in the imputation model for missing alcohol among bed sharing cases was increased. Of course, we could get a similar effect by decreasing the probability of alcohol use when imputing missing alcohol use data in the controls.

We first applied the congenial sensitivity analysis to the *alcohol* imputation model. Increasing or decreasing the bed sharing coefficient in the *alcohol* imputation model had only a small effect on the bed sharing AOR in the AOR MOI. The change was small because the association between *alcohol* and bed sharing (interaction D) AOR in the AOR MOI absorbs the effect. We thus concluded the bed sharing inference from the AOR MOI in these settings is robust to departures from the MAR assumption in the imputation of *alcohol* values for cases.

We repeated similar congenial sensitivity analysis for the *drug* imputation model. While the effect was larger than the congenial sensitivity analysis on the *alcohol* imputation model, the bed sharing risk inference from the AOR MOI to the departure of MAR in the imputation of *drug* was robust under the settings we considered.

We then preformed uncongenial sensitivity analysis for both the *alcohol* and *drug* model separately, looking at extreme scenarios. The scenario that affected the *bed shared* AOR the greatest in the *alcohol* imputation was when all SIDS cases who *bed shared* was set to not drinking more than 2 units of alcohol and all controls who *bed shared* was set to drinking more than 2 units of alcohol (non bed sharing records remained imputed under MAR). In this setting the adjusted odds ratio is above one unlike the MAR adjusted odds ratio, however the MAR confidence still contains the MNAR point estimate suggesting that the difference is not significant (see Figure 4.7). We explored this graphically in Figure 4.8 and observed that the time when the AOR is significant has increased from approximately 14 weeks under MAR to 22 weeks. The AOR shows that babies with smoking mothers and partners still have a higher risk of SIDS even when imputed with an extreme MNAR mechanism. The extreme sensitivity analysis suggests that the inferences are robust to even extreme departures from MAR in the *alcohol* imputation.

Similarly in the *drug* imputation scenarios, when all SIDS cases that *bed shared* were set to non *drug* use and all controls that *bed shared* were set to *drug* use (non bed sharing records remained imputed under MAR), the *bed shared* AOR changed the most from 0.69 under MAR to 1.40. We showed in Figure 4.10 and Figure 4.11 that the covariate changes in the AOR MOI have increased and decreased the SIDS risk greatly, with the time period when the non smoking parents risk is significant ranging from 3 to 22 weeks. However, it is important to note, that even in these very unlikely extreme scenarios we still observe bed sharing as a risk for SIDS in both smoking groups. The extreme sensitivity analysis suggests that the inferences are robust to even extreme departures from MAR in the *drug* imputation.

In both the congenial and the uncongenial sensitivity analysis settings we see that the bed share inference is more sensitive to departures from MAR in the imputation of *drug* than *alcohol*. This is probably because *drug* use has a smaller prevalence than *alcohol* drinking so that by increasing or decreasing the number of imputed mother drug users the effects are greater. In all scenarios the downward SIDS rate trend by age is retained and in all examples bed sharing increases the risk of SIDS for young babies. In the more realistic settings when the multivariate prior is applied the alterations to the AOR's are negligible, suggesting the inferences are robust to plausible departures from the MAR assumption. This suggests our findings in Carpenter *et al.* (2013) are robust to departures away from the MAR assumption. Our next step is to write a rebuttal letting in response to the criticism of the imputation in Carpenter *et al.* (2013) demonstrating the robustness of inferences which we have shown here.

Methodological Conclusions:

This section has demonstrated two different sensitivity analysis approaches. We begun with the congenial sensitivity analysis, using multivariate priors with chosen values to reflect plausible modest changes to the *bed shared* covariate on the Probit scale for the centres where *alcohol* and *drug* were not collected. To create more realistic priors we could have elicited information about the MNAR mechanism from experts. We also applied the priors on *alcohol* and *drug* separately; for a more comprehensive congenial sensitivity analysis these could have been applied simultaneously. However our approach allows us to understand the effect of each, on its own. The steps we undertook to apply the uncongenial sensitivity analysis were complex and time consuming. However, in situations where inferences are expected to be sensitivity to extreme, uncongenial sensitivity analysis (as was the case here), the additional complexity of congenial sensitivity analysis provides a way to explore the effect of contextually plausible

departures from MAR. This can be formed by expert opinions, as discussed in Chapter 3.

To implement our approach, we had to adapt the MATLAB code for REALCOM, to extend it to handle multivariate priors. If this feature was made generally available in the REALCOM software, sensitivity analysis of this kind would be much more straightforward for practitioners. To make the process feasible for readers who cannot engage in programming we therefore suggest that future releases of REALCOM include the multivariate prior option.

In the latter part of the chapter we demonstrated uncongenial sensitivity analysis by proposing extreme imputation settings. While the settings are highly unrealistic it gives insight into the range of inferences which can be created. In a practical situation a researcher may wish to begin with the extreme uncongenial sensitivity analysis due to its ease of application. If inferences are found to be robust, then the congenial sensitivity analysis (describe in this chapter) is not required to be preformed as it will also be robust. However if inferences from the extreme uncongenial sensitivity settings are not robust, then the researcher should also implement more realistic congenial sensitivity analysis. These could take the form of searching for the "tipping point"— i.e. the parameter values for the MNAR prior at which the MAR inferences significantly change.

In both sensitivity analysis approaches we focused on the cases, as debate in the SIDS community has focused on the possibility of different alcohol and drug habits in this group. Nevertheless the method can easily be applied to the controls. However, changing many elements together increases the difficulty of understanding the reasons why inferences are/or are not robust to departures from MAR.

5

Sensitivity Analysis via Re-Weighting after Multiple Imputation

Assuming Missing At Random: Issues with Small Datasets

5.1 Introduction

Multiple imputation (MI) is often applied under the assumption that data are Missing At Random (MAR). However, this assumption is untestable and may be inappropriate. It is thus good practice to perform sensitivity analysis to explore how robust inferences are to the plausible departure from the MAR assumption. One approach for this is selection modelling, where a model for the probability of observing data (including dependence on underlying, but unseen values) is fitted concurrently with the substantive model of interest. However, this is often relatively difficult to perform, hence a proposal by Carpenter *et al.* (2007) to approximate this by re-weighting estimates obtained after multiple imputation under MAR. The method exploits the advantages of MI to assess robustness of inference to departure from the assumption of MAR.

In this chapter, we describe our investigation into the performance of the re-weighting method in small samples. This included writing R code to implement the method and perform the simulation studies. While the Carpenter *et al.* (2007) method showed promising results in simulation studies, Carpenter *et al.* (2007) noted some erratic behaviour in small samples. The purpose of this chapter is to identify the cause of this and address it.

The chapter is structured as follows. In the first section we describe the approach by Carpenter *et al.* (2007), and reproduce their simulation results. This confirms the method may be biased in small datasets. We then describe and explore two possible sources of this bias. These are, first, that the distribution of the observed data may be poorly modelled by the MAR imputation model, and second that the MAR imputation distribution may have too large a

variance relative to the true distribution of the observed data. We explore the first issue in Section 5.3. We do this with a tractable probit selection model. Then, in the Section 5.4, we investigate the second issue. We show this is the cause of the problem and propose an approximate correction/solution. Through our consideration of the method behaviour in small datasets,we are led to a proposal for extending the approach to non-local sensitivity analyses.

5.2 Sensitivity Analysis by Re-Weighting after MI under MAR

We begin by describing the re-weighting method proposed by Carpenter *et al.* (2007). This uses importance sampling (e.g. Ripley, 1987). Importance sampling draws from a distribution but over weights the important region (hence the name importance sampling) of the distribution that we truly want to sample from. For example if we wish to find $\mu = E[g(\boldsymbol{x})]$ where $P(\boldsymbol{x} \in A) \approx 0$ where A is a region of the distribution $\boldsymbol{x}$ (for example the tail of distribution), then with a simple Monte Carlo sample from $\boldsymbol{x}$ we are unlikely to have many observations from region A. To overcome this we sample from a distribution of $\boldsymbol{x}$ which over weights region A. Afterwards we have to adjust our estimate of μ to reflect the fact that $\boldsymbol{x}$ has not been sampled from the actual distribution of interest. This adjustment involves the *importance weights*. In our sensitivity analysis setting the *importance weights* are defined as the ratio of our target MNAR distribution to the MAR distribution which we have imputed under.

To illustrate, consider $(X_i,\ Y_i)$, $i = 1, \ldots, n$, and suppose our parameter of interest is θ, estimated by $\hat{\theta}(\boldsymbol{X}, \boldsymbol{Y})$. Suppose that $\boldsymbol{Y}$ is partially observed and $\boldsymbol{X}$ is fully observed, with a selection mechanism:

$$\text{logit}P(\text{observe}\, Y_i) = f(X_i) + \delta Y_i, \tag{5.2.1}$$

for some unspecified $f(X_i)$. Thus, if $\delta = 0$, $\boldsymbol{Y}$ is MAR dependent on $\boldsymbol{X}$.

Suppose we impute missing $\boldsymbol{Y}$ under MAR, with Y_i^m, $m = 1, \ldots, M$, as the m^{th} imputed value of Y_i. As usual, we form the imputed datasets ($\boldsymbol{Y}_i^m$, $\boldsymbol{X}$), $m = 1, \ldots, M$, and estimate the parameter of interest in each, giving $\hat{\theta}_1, \ldots, \hat{\theta}_M$. Then our MAR estimate of θ from MI is:

$$\hat{\theta}_{MAR} = \frac{1}{M} \sum \hat{\theta}_m.$$

(Carpenter *et al.*, 2007, p.262) show that under selection model (5.2.1), θ can be estimated by:

$$\hat{\theta}_{MNAR} = \sum \frac{w_m \hat{\theta}_m}{\sum w_m},$$

where because observing Y_i^m are independent of observing $Y_{i'}^m$ ($i \neq i'$), the log weights for imputation m are a linear combination of the imputed data:

$$w_m \propto exp\left(- \sum_{i=1}^{n} \delta Y_i^m \right). \tag{5.2.2}$$

Note that if $\delta = 0$, $\hat{\theta}_{MAR} = \hat{\theta}_{MNAR}$ as it should.

Carpenter *et al.* (2007) reported promising results with moderately large datasets with this method, but showed over compensation in some small datasets. This is confirmed in the following simulation which reproduces the simulation results of Carpenter *et al.* (2007).

Simulation Study

We draw $n = 100$ pairs of observations, (X_i, Y_i) from:

$$N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right),$$

and take the selection model:

$$\text{logit}P(\text{observe}\, Y_i) = X_i + \delta Y_i, \tag{5.2.3}$$

following from (5.2.1), with $f(X_i) = X_i$ and $\delta = 1$. This makes half the values of Y missing on average. For each pair, the probability of observing Y_i,

$$p_i = \frac{e^{(X_i+Y_i)}}{1+e^{(X_i+Y_i)}}, \tag{5.2.4}$$

was calculated, and a uniform $[0,1]$ variable u_i was generated. Y_i was set to missing if u_i was greater than p_i which made approximately 50% of the data set missing. The data were thus MNAR. To show the effect of miss-specifying the missing data mechanism, we calculate the marginal mean of $\boldsymbol{Y}$, which is 0, using the following three methods:

1. Method 1: A complete record analysis in which only observed Y_i's were used; this method assumes MCAR.

2. Method 2: MI creating M imputed datasets from a regression of $\boldsymbol{Y}$ on $\boldsymbol{X}$ under the assumption that the data is MAR, dependent on $\boldsymbol{X}$.

3. Method 3: Re-weighting the M imputed estimates from method 2 as described above, using weights given by (5.2.2), with $\delta = 1$.

The process was replicated 1000 times with various values of M. The average estimate of $E[\boldsymbol{Y}]$ was found across 1000 replications.

		Number of Imputations M				
		5	10	50	100	1000
Method 1: MCAR (No Imputations)	0.491					
Method 2: MAR		0.322	0.323	0.328	0.327	0.325
Method 3: MNAR		0.194	0.163	0.087	0.063	-0.012

Table 5.1: Simulation results showing the marginal mean of $\boldsymbol{Y}$, $E[\boldsymbol{Y}]$, from methods 1-3 described in the text. Results represent the average estimate of $E[\boldsymbol{Y}]$ across 1000 replications based on 100 observations.

In Table 5.1 all three methods show an upward bias from the true mean of zero, except one result from method 3 with $M = 1000$ imputations. Method 1 has the largest bias. The bias observed for method 2 does not converge to zero with more imputations as the MAR assumption is incorrect. This demonstrates the effect of miss-specifying the missing data mechanism. Method 3 removes substantially more bias than the other two methods. As the number of imputations increases, the amount of bias removed increases, with 98% removed when 1000 imputations are used.

We next extended the simulation to explore the performance of the method with small datasets. Thus the simulation was repeated with $n = 20, 40, \ldots, 100$.

	Number of Imputations M				
n	5	10	50	100	1000
100	0.19	0.16	0.09	0.06	-0.01
80	0.19	0.14	0.05	0.02	-0.05
60	0.17	0.11	0.02	-0.02	-0.13
40	0.12	0.06	-0.04	-0.11	-0.23
20	0.02	-0.08	-0.21	-0.27	-0.45

Table 5.2: Simulation results showing the marginal mean of $\boldsymbol{Y}$, $E[\boldsymbol{Y}]$, from method 3 described in the text. Results represent the average estimate of $E[\boldsymbol{Y}]$ across 1000 replications, with varying sample size n and number of imputations M.

Table 5.2 shows the results for method 3. For small n, the results are clearly not converging on 0 as M increases; they appear to be increasingly negative.

The following sections of this chapter explore two possible theories for why the estimates do not converge to zero. We believe that the problem lies within the importance sampling which underpins this approach. Central to the importance sampling argument is that the estimated MAR imputation distribution, say $f(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{R} = 1)$, is a good approximation to the true distribution of the observed data, say $g(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{R} = 1)$, where $\boldsymbol{R} = 1$ if $\boldsymbol{Y}$ is observed, otherwise $\boldsymbol{R} = 0$.

Since the re-weighting method involves up-weighting samples from the tail of $f(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{R} = 1)$, we hypothesise first that, when the sample size n is small, approximating g by choosing f to be a normal distribution may cause the bias. If the g distribution is not normal then we could be over weighting the observations draw from the f distribution and thus reducing the estimate too much creating the negative values seen in Table 5.2. We explore this using the analytically tractable probit selection model in the next section.

5.3 Estimated MAR Distribution

In this section we investigate whether the bias observed in small samples is caused by modelling the distribution of the observed data with the normal distribution.

We consider a probit selection model, as this allows us to calculate the true distribution of the observed data. We derive a probit selection model from the bivariate normal distribution defined as:

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ \mu_x \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right). \tag{5.3.1}$$

Where $f(X|y)$ is:

$$f(X|y) \sim N(\mu_x + \rho y, (1-\rho^2)). \tag{5.3.2}$$

We suppose that x is an unobserved latent variable which determines the missingness in Y. Y is observed if and only if $x > 0$. Then the distribution of the observed data is given by:

$$P(Y|x>0) = \frac{\Pr(Y) \times \Pr(x>0|Y)}{\Pr(x>0)}. \tag{5.3.3}$$

From (5.3.3) we calculated $\Pr(Y|x>0)$ and $\Pr(Y|x \leq 0)$ under model (5.3.1) (see Appendix D.1):

$$P(Y|x>0) = \frac{\phi(y)}{\Phi(\mu_x)}\Phi\left(\frac{\mu_x + \rho y}{\sqrt{(1-\rho^2)}}\right) \tag{5.3.4}$$

$$P(Y|x\leq 0) = \frac{\phi(y)}{1-\Phi(\mu_x)}\left(1-\Phi\left(\frac{\mu_x + \rho y}{\sqrt{(1-\rho^2)}}\right)\right) \tag{5.3.5}$$

Here, $\Phi(.)$ is the cumulative distribution of the standard normal, and $\phi(.)$ is the probability density function of the standard normal. Equation (5.3.4), represents the true distribution of the observed data. If we impute under a normal model, we approximate this by a normal distribution, with true mean $E[Y|x>0]$ and variance $V[Y|x>0]$. We hypothesise that this approximation assumption might be incorrect and thus cause the bias in the estimates. We explored this assumption (choosing arbitrary values) with $\mu_x = 0$ and $\rho = \frac{1}{\sqrt{2}}$. In Appendix (D.2) we show $E[Y|x>0] = \frac{1}{\sqrt{\pi}}$ and $V[Y|x>0] = 1 - \frac{1}{\pi}$.

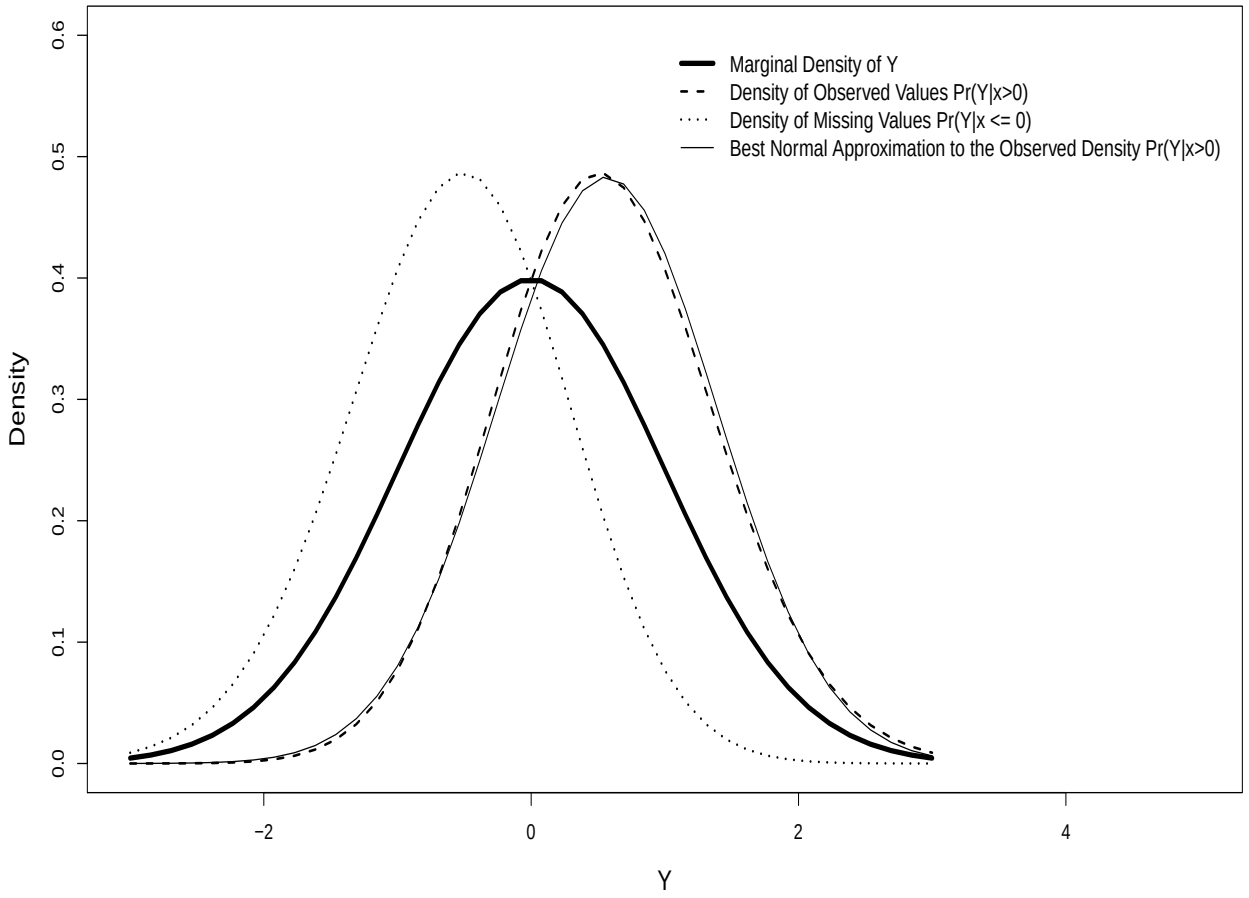

Figure 5.1: Probit selection model with $\mu_x = 0$ and $\rho = \frac{1}{\sqrt{2}}$. Plot of the marginal density of Y, the density of 'missing' values, $\Pr(Y|x \leq 0)$, the density of 'observed' values, $\Pr(Y|x > 0)$, and the normal approximation of the observed density $\Pr(Y|x > 0)$.

Figure 5.1 shows the density plots from the probit model. The densities of $\Pr(Y|x > 0)$ and $\Pr(Y|x \leq 0)$ are of equal height since the probability of a missing observation is 0.5. When Y is observed, $\Pr(Y|x > 0)$, the data is pushed to the right relative to the marginal distribution of Y. As ρ increases the proportion of negative values of Y, which are observed, tends to zero. Likewise as ρ increases the proportion of positive values of Y, which are 'missing', tends to

zero. When ρ is zero, both of the conditioned curves have the same mean and distribution as the marginal density of Y. As μ_x increases (for fixed ρ) the density of $\Pr(Y|x \leq 0)$ tends to the marginal density of Y, as it decreases the density of $\Pr(Y|x > 0)$ tends to the marginal density of Y.

Figure 5.1 also shows the normal approximation to the observed density using $E[Y|x > 0]$ and $V[Y|x > 0]$ calculated in Appendix (D.2). We see that the normal approximation to the true density is very good. We hence conclude that the bias observed in our simulation is not due to approximating the distribution of the observed data by a normal distribution. Thus, we next consider our second hypothesis for the bias.

5.4 Variability In Estimated MAR Distribution

We explored the hypothesis that within small datasets, there is significant variability between the estimated MAR distribution used to draw the imputations from and the true distribution of the observed data. The effect of this is that the distribution of the imputed data has a much heavier tail than the true distribution of the observed data. When re-weighting, this tail causes bias.

We investigated this using a probit selection model simulation, with $\mu_x = 0$ and $\rho = \frac{1}{\sqrt{2}}$. This conveniently reduces the selection model so that it is only dependent on Y (P represents the probability density function):

$$\mathrm{P}(Y|x > 0) = 2 \times \Phi(y) \times \phi(y) \tag{5.4.1}$$

$$\mathrm{P}(x > 0|Y) = \Phi(\alpha_0 + \alpha_1 y), \tag{5.4.2}$$

where in Appendix (D.3) we show:

$$\alpha_0 = \frac{\mu_x}{\sqrt{(1-\rho^2)}} = 0$$
$$\alpha_1 = \frac{\rho}{\sqrt{(1-\rho^2)}} = 1.$$

We use a probit selection model as an approximation to a logistic selection model, since it allows closed form calculation of the distribution of the observed and missing data. This means we can algebraically calculate the exact parameters of the true distribution to further understand how bias is being introduced. To link the analytic probit model with the logit selection model in the simulations, we need to choose δ in (5.4.3) to give the same level of selection as (5.4.2). The logit selection model is:

$$\mathrm{logit}(\Pr(R = 1|Y)) = \beta_0 + \delta y. \tag{5.4.3}$$

We estimated δ empirically, by drawing 1 million observations from the probit model, then setting $R = 1$ if $x > 0$ representing Y as observed and $R = 0$ otherwise to represent Y as missing. We then fitted (5.4.3) to the data, finding $\delta = 1.7$. This means in our setting to approximate the logistic selection model level with a probit section model we are required to set δ to 1.7 so that the selections are approximately equivalent.

The next step was to create the imputations under MAR. To do this the following MI algorithm was used. Suppose we intend to have n observations i but only observe n_{obs}, the standard imputation process is:

(a) Calculate $\hat{\mu}_{obs} = \sum_{i=1}^{n_{obs}} \frac{Y_i}{n_{obs}}$ and $\hat{\sigma}^2_{obs} = \sum_{i=1}^{n_{obs}} \frac{(Y_i - \hat{\mu}_{obs})^2}{n_{obs}-1}$

(b) For each imputed dataset, $m = 1, \dots M$

(i) Draw $\tilde{\sigma}^2_m \sim \frac{\hat{\sigma}^2_{obs}(n_{obs}-1)}{\chi^2_{n_{obs}-1}}$

(ii) Draw $\tilde{\mu}_m \sim N\left(\hat{\mu}_{obs}, \frac{\tilde{\sigma}^2_m}{n_{obs}}\right)$

(iii) Draw missing Y's from $N(\tilde{\mu}_m, \tilde{\sigma}^2_m)$

(c) Calculate the mean of the observed and imputed Y.

After imputing the data, we applied the Carpenter *et al.* (2007) re-weighting method with $\delta = 1.7$, to obtain the average estimate of $E[\boldsymbol{Y}]$ (true value of 0). We focused on small datasets, with the simulation replicated 1000 times with various values of M.

	Number of Imputations M				
n	5	10	50	100	1000
40	0.40	0.35	0.24	0.19	0.06
20	0.32	0.23	0.03	-0.06	-0.39

Table 5.3: Estimate of $E[\boldsymbol{Y}]$ across 1000 replications based on varying the number of imputations M and sample size $n = 20, 40$. Approximately 50% of observations were missing.

The results are shown in Table 5.3. As expected, these are similar to those in Table 5.2. The difference is that these results have been simulated under the probit model (5.3.1), so we can cross check each step of the simulation with analytic results to find the cause of the bias in the bottom right of Table 5.3.

We now investigate the variance of the imputed draws, compared to the true distribution of the observed data (5.4.1).

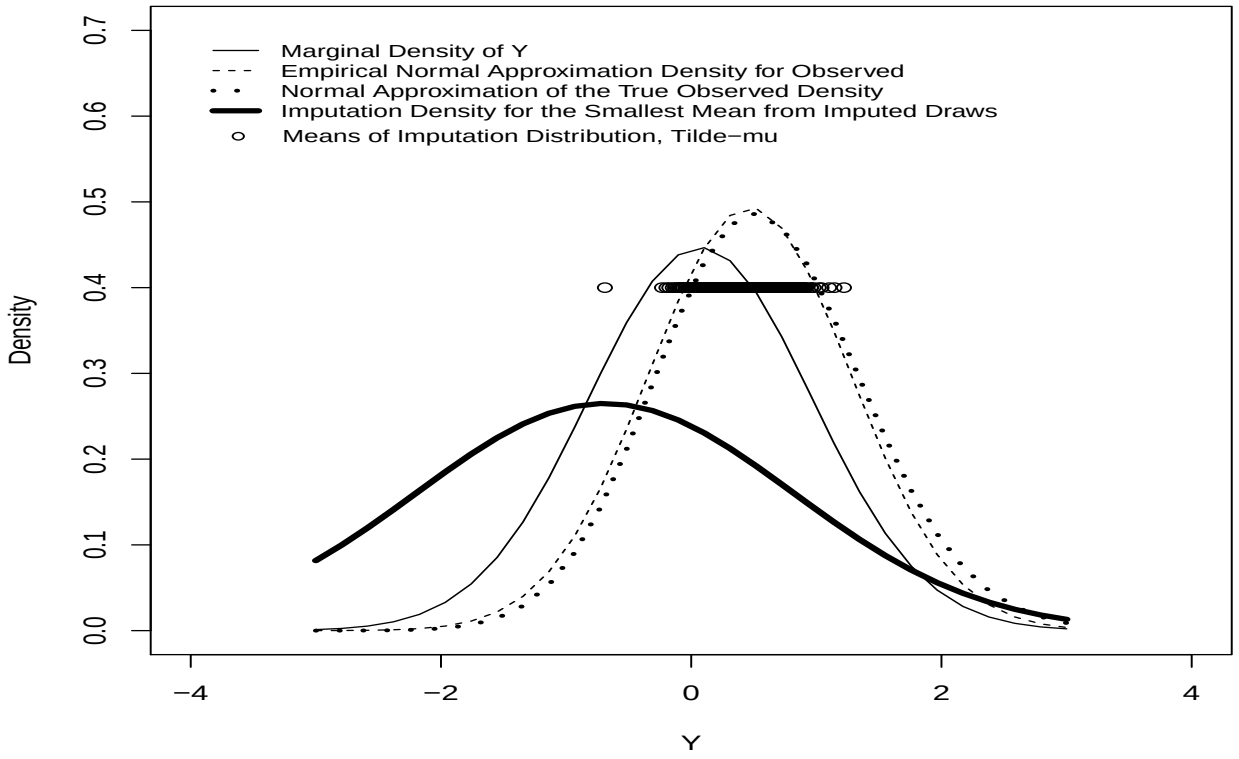

Figure 5.2: Probit simulation with $n = 20$, approximately 50% observed and $M = 1000$, showing the marginal density of Y, empirical normal approximating density for the observed data, the normal density with the true mean and variance of the observed data (from (5.4.1)), the means of the imputation distribution, $\tilde{\mu}_m$, and the imputation density for Y from the smallest imputed mean draw $\tilde{\mu}_m$.

Figure 5.2 shows the results for a simulation with $n = 20$ and $M = 1000$. It shows the marginal density of Y, the normal density with mean and variance set equal to the true mean and variance of the observed data, the empirical normal density for the observed data, the means of the imputed draws $\tilde{\mu}_m$ (created in step b)ii) and the imputation density $N(\tilde{\mu}_m, \tilde{\sigma}_m^2)$ (created in step b)iii) for the smallest value $\tilde{\mu}_m$. It is clear that the imputed density has a mean and variance very different to the empirical density and true density of the observed data. When

re-weighting the imputations, a large weight will accrue for this imputation, biasing the mean downwards.

Thus, this investigation suggests the negative values observed in Table 5.3, i.e. the over compensation of the Carpenter *et al.* (2007) re-weighting method, are caused by the variability in the distribution of the imputed data. This happens because the estimated MAR imputation distribution, which takes into account the variability in estimating the parameters of the MAR distribution, has considerably heavier tails than the true MAR distribution. This difference is especially marked in small data sets. This in turn means the method in re-weighting draws from a distribution whose tails are too heavy, and this causes the downward bias.

To verify this, and correct the imputed values so they are closer to the true distribution of the observed data, we re-weighted the imputed data (*re-weighted method*). Since the true density we wished to sample from was $N\left(\frac{1}{\sqrt{\pi}}, 1-\frac{1}{\pi}\right)$, we add a component to the importance weight $\frac{\phi\left[Y_i^m;\frac{1}{\sqrt{\pi}},1-\frac{1}{\pi}\right]}{\phi[Y_i^m;\tilde{\mu},\tilde{\sigma}^2]}$ to bring us back to the distribution we require. This is then multiplied by $e^{-\delta Y_i^m}$ to move to MNAR, as before. The new weights are:

$$w_m = \prod_{i=1}^{n_{miss}} \frac{e^{-\delta Y_i^m} \times \phi\left[Y_i^m; \frac{1}{\sqrt{\pi}}, 1-\frac{1}{\pi}\right]}{\phi[Y_i^m; \tilde{\mu}_m, \tilde{\sigma}_m^2]},$$

where $Y_i^m = Y_i$ if Y_i is observed otherwise $Y_i^m = m$th imputation under MAR. n_{miss} is the number of observations missing and $\phi[.]$ represents the normal the probability density function of the standard normal. As before we estimate the marginal mean of $\boldsymbol{Y}$ by:

$$\frac{\sum_{m=1}^{M} w_m \bar{Y}_m}{\sum_{m=1}^{M} w_m}, \tag{5.4.4}$$

where $\bar{Y}_m$ is the mean of the observed and imputed data from $m = 1, \ldots, M$ imputed datasets. The *re-weighted method* was applied with 1000 replications to the probit simulation as the number of observations/imputations varied.

	Number of Imputations M				
n	5	10	50	100	1000
40	0.40	0.36	0.27	0.24	0.16
20	0.35	0.29	0.17	0.14	0.10

Table 5.4: Probit results, using the *re-weighted method*. Results represent the average estimate of $E[\boldsymbol{Y}]$ across 1000 replications based on varying sample size n and the number of imputations M.

Comparing Table 5.4 with Table 5.3 we see that the *re-weighted* method has removed the negative estimates. As the number of imputations increase the estimates appear to converge towards zero. The estimate using the *re-weighted method* is less biased than the Carpenter *et al.* (2007) method when the number of imputations is 1000 with 20 observations. However the *re-weighting* method does appear to converge more slowly in bigger datasets. It should be noted, though, that this is a challenging example, as we discuss later.

The *re-weighted method* was further applied to 5000 imputations and $n = 20$ observations, with approximately half missing. The Carpenter *et al.* (2007) method produced an estimate of -0.71 while the *re-weighted method* produced a remarkably better estimate of 0.06. The *re-weighted method* results suggest that the new weighting is correcting for the high variability between the imputed draws, thus improving the estimation when the number of imputations is large and the dataset is small.

In this probit simulation we know the true mean and variance of the observed data distri-

bution; however in practice, of course this is unknown. We thus propose approximating the unknown distribution in the numerator of (5.4.4) with the empirical normal distribution of the observed data $\phi(\hat{\mu}_{obs}, \hat{\sigma}^2_{obs})$.

$$w_m = \prod_{i=1}^{n_{miss}} \frac{e^{-\delta Y_i^m} \times \phi\left[Y_i^m; \hat{\mu}_{obs}, \hat{\sigma}^2_{obs}\right]}{\phi[Y_i^m; \tilde{\mu}_m, \tilde{\sigma}^2_m]},$$

We applied these approximated weights (*approximated re-weighted method*) to our simulation, the results are recorded in Table 5.5.

	Number of Imputations M				
n	5	10	50	100	1000
40	0.40	0.36	0.27	0.23	0.15
20	0.35	0.28	0.17	0.13	0.09

Table 5.5: Probit results, using the empirical normal approximation of the observed data to weight the data. Results represent the average estimate of $E[\boldsymbol{Y}]$ across 1000 replications based on varying sample size n and the number of imputations M.

Comparing Table 5.5 with Table 5.4 we see that the *approximated re-weighted method* performs very similarly to the *re-weighted method*, removing the negative estimates seen in Table 5.3. What we are doing is re-weighting our draws from the MAR imputation distribution by our best estimate of the MAR distribution. The *approximated re-weighted method* also appears to not over compensate quite as much as the *re-weighted method* in this setting. A possible reason for this is because we have a small sample the empirical normal distribution is a closer fit to the data than the normal distribution with the true mean and variance of the observed data. Again we applied this method to 5000 imputations with $n = 20$ observations and with approximately half missing. We received the same result as the *re-weighted method* (0.06), thus confirming the *approximated re-weighted method* successfully fixes the

problem with the original method. A disadvantage in both the *re-weighted method* and the *approximated re-weighted method* appears to be slow convergence to zero bias. However it must be remembered that this is a challenging problem: the mean of the observed data is $\frac{1}{\sqrt{\pi}} \approx 0.56$, and variance $1 - \frac{1}{\pi} \approx 0.68$. This means that the standard error of the mean of 10 observed Y values is $\sqrt{\frac{1-\frac{1}{\pi}}{10}} \approx 0.26$, so the observed data mean is around two standard errors from the true marginal mean of zero. For larger data sets, under the same model, the difference will be many more standard errors. Lastly, while we have not explored the variance in this investigation, we note that the key to importance sampling working is getting the weights right, particularly in the tails. Once this is done, convergence is arrived as the number of draws (here imputations) increases. Therefore, the re-weighting should lead to improved variance estimates, at least with sufficiently number of imputations.

The *approximated re-weighted method* also suggests how we may extend the approach to non-local sensitivity analyses. The idea is to impute from a distribution that has a larger than required variability under MAR, and then re-weight in the appropriate way for both the MAR and MNAR analysis.

By imputing from a distribution with a larger than required variability under MAR, we generate a range of Y_i^m that can be re-weighted to represent non-local departure from MAR. In effect we trade a loss of simplicity in the MAR analysis for the option of re-weighting to look at sensitivity for a broader range of departures from MAR.

5.5 Conclusion

In conclusion, the cause of the bias reported by Carpenter *et al.* (2007) has been traced to high variability in the distribution of the imputed data when n is small and the proportion of missing observations is large. The 'solution' is to re-weight the imputed datasets, so the imputed values better represent draws from the true distribution of the observed data. This can be achieved by using the an empirical normal approximation of the observed data distribution within the weights, as seen in (5.4.5). We cannot give definitive guidelines when to apply the *approximated re-weighted method* as it will be data dependent, however we suggest that the method should be applied and compared to the Carpenter *et al.* (2007) method when preforming a large number of imputations on small datasets. Through our development of the method we are also able to suggest a way of addressing another problem with the Carpenter *et al.* (2007) weighting method, that it is only suitable for local sensitivity analysis. Besides publishing this work, we intend to use the code written for this chapter to produce an R software package to make the method more generally available.

Future work could usefully explore the use of this method when covariates are missing, and also the usefulness of the extension to non-local sensitivity analyses described previously.

6

Discussion, Conclusions and Future Work

Multiple imputation under the assumption of missing at random is a popular tool for the analysis of partially observed data; however the MAR assumption is a strong untestable one. If the MAR assumption is incorrect inferences may be biased, hence it is extremely important to explore the robustness of inferences to the alternative assumption that missing data are MNAR. This is often not undertaken, thus this thesis has developed, explored and illustrated practical approaches for sensitivity analysis in an observational data setting, to remove barriers to their use and demonstrate their utility. We began by reviewing the practicalities of multiple imputation, highlighting its advantages over other methodologies. We then developed and applied sensitivity analysis using a pattern mixture framework for a colorectal cancer setting. This include eliciting a prior from experts in this area. Next we used the pattern mixture framework to assess the robustness of inferences in a multicentre sudden infant death syndrome case control analysis, demonstrating difficulties that will typically arise with this type of data, and possible solutions. For comparison we also performed a best/worse case sensitivity analysis, looking at extreme MNAR settings. Lastly we used an inverse probability weighting method as an approximation for selection modelling for sensitivity analysis. We demonstrated the methods' limitations and developed a methodological solution which we evaluated using simulation. We will now discuss each chapter in turn, considering our practical and methodological findings and outline possible future work.

In the motivating colorectal cancer setting we explored a model based pattern mixture sensitivity analysis approach, which involves eliciting a multivariate prior. We began by replicating the inferences in the Morris *et al.* (2011) publication using the same substantive and imputation models.

We then moved a step further, eliciting information from experts through a electronic ques-

tionnaire. We took care to frame the questionnaire in terms the expert would understand. Nevertheless, the number of experts who completed the questionnaire was small, as many experts replied saying they did not have the knowledge required. This supported the finding in White *et al.* (2007) that not all experts are comfortable giving their opinion about the unknown. Kadane and Wolfson (1998) suggest that the elicited information should only come from experts who work/have good knowledge of the colorectal cancer field, however in practice it was difficult to find/contact these experts in the UK. We used electronic communication to elicit information from experts; however a face to face meeting may have aided this process as it would have given more freedom to question and clarify concepts. The number of experts required is not currently clarified in literature but we believe the quality of the information is more important than the number of experts used. The elicitation process was time consuming as we had to wait for responses and answer questions, this we conclude may put off analysts from using the method. However an advantage of this method is that it is computationally simple once the prior has been elicited. A further advantage is that unlike selection modelling, the missing mechanism model does not have to be jointly fitted with the substantive model, which can be computationally demanding. In our method we only focused on two predictive variables (the two most predictive of Dukes' stage being missing and of the underlying Dukes' stage value), this was done to simplify the questionnaire, while retaining the focus on pratically important variables. In some situations this simplification would not be appropriate, making the elicitation of information difficult. An additional advantage of this approach is that we believe that the process of eliciting information also helps raise awareness not only of the impact of missing data and the benefits of improving data collection but also of the importance of using sensitivity analysis to investigate the robustness of inferences to the MAR assumption.

The information from the elicitation was then modelled using a Dirichlet distribution to im-

pute the data with an MNAR distribution. This was chosen as it is a conjugate prior for the multinomial distribution, however other distributions could have been chosen. Future work could look at the effect of prior distribution choice on the sensitivity analysis approach. In the chapter we also supported our pattern mixture sensitivity analysis approach with a simplified simulation study to check that Rubin's rules would give reliable variance estimates in this setting. The simulation supported our method, demonstrating close agreement between the empirical and theoretical variances. The substantive model was applied to the MNAR imputed data to investigate the robustness of inferences to departures from MAR. The inferences from the MAR and MNAR imputed data were similar, demonstrating that the substantive model estimates are robust to contextually plausible departures from MAR.

We aim to publish our work undertaken in the chapter and currently have a draft paper. We will aim the publication towards the cancer research community to demonstrate the pattern mixture sensitivity analysis method in a motivating example. This will hopefully encourage researchers to undertake the method to support their inferences after multiple imputation and hence improve research quality.

The next chapter investigated the robustness of inferences in a Sudden Infant Death Syndrome (SIDS) study using both congenial sensitivity analysis though multivariate priors and uncongenial sensitivity analysis. The chapter was motivated by research we undertook for Carpenter *et al.* (2013) paper, which when published had responses criticising the imputation approach we undertook. The letters focussed on the strong untestable MAR assumption, which they believe maybe untrue and hence they found the conclusions, particularly relating to the finding that bed sharing was a risk factor for SIDS, contestable.

We first investigated the missing data pattern and concluded that the complete records analysis would not be biased as the model of interest contained the variable *centre* which was the main reason for missingness. Hence given *centre*, the probability of missing the variables *alcohol* and *drug* use did not depend on the case-control status or other exposures. However due to the large amount of data lost if the complete records subset is used we decided to apply multiple imputation. We next described the multilevel imputation process under MAR which was very difficult to complete due to the sparse nature of the data and the need to preserve as much as possible the structure of the substantive model in the imputation model. To achieve MAR imputation through multiple imputation with a hierarchical level, we had to simplify the imputation model and constrain the level 2 (centre level) covariance matrix. Due to the complexity of the process, running the imputation was computationally time consuming. After we imputed the data we applied the model of interest and compared our inferences to the single level imputation we used in Carpenter *et al.* (2013). Post estimates were similar but inferences were now more precise.

We then carried out congenial sensitivity analysis using a multivariate prior. This required the REALCOM software to be reprogrammed to incorporate the prior. We applied the method separately for the *drug* and *alcohol* imputation. In the *alcohol* model we applied the prior to change the proportion of *bed sharing* on the Probit scale for the centres where *alcohol* was not collected. In the *drug* model we only changed the constant (as *bed shared* information was not used), again for only centres where *drug* was not collected. Values were used in a multivariate prior to change the odds of alcohol/drug behaviour among those with missing observations, relative to the odds of alcohol or drug use in the centres where the data were collected. To obtain more realistic priors we could have elicited information about the MNAR mechanism from experts. The eliciting process would be difficult in this setting due to the

complexity of the syndrome and the small prevalence level. However, our main conclusion, that after adjusting for other risk factors bed sharing is a risk for young babies, was supported in all the congenial sensitivity analyses we applied.

We then applied uncongenial sensitivity analysis, to investigate the robustness of inferences to extreme departures from the MAR assumption, for example in one setting we imputed all missing drug observations as taking drugs. While the settings were highly unrealistic it gave insight into the range of inferences which could be created. The *bed shared* inference created when changing the imputation of *drug* use, was more sensitive to the extreme settings than changes to the imputation of *alcohol* use. However in all settings bed sharing remained a significant risk to young babies and thus we concluded overall that the inferences were robust to departure from the MAR assumption and hence support our findings in Carpenter *et al.* (2013). Future work will consist of writing and publishing a rebuttal letter in response to the criticism of the imputation.

In the last chapter we built on the method proposed by Carpenter *et al.* (2007) which approximates MNAR selection models through inverse weighting and importance sampling after multiple imputation under MAR. We discussed the difficulties of applying selection modelling due to the challenge of making the selection model to represent the missingness mechanism accessible to all those involved in the analysis. We also discussed the limitations of the method proposed by Carpenter *et al.* (2007) when applied to small datasets. Through simulations we demonstrated the cause of the bias reported by Carpenter *et al.* (2007), was due to the high variability in the distribution of the imputed data when the sample size was small and the proportion of missing observations large. We proposed a possible solution to reduce the bias by re-weighting the imputed data sets, so the weights better approximate the likelihood of the

data under MAR. Our method however had a large limitation: we needed to know the true distribution of the data to calculate the improved weights. We thus recommended and tested using the empirical normal approximation to the observed data distribution for the weights. The simulation gave good results, demonstrating a comparable bias reduction to the true distribution method. With our addition, the method can be applied for non-local sensitivity analysis, which the Carpenter *et al.* (2007) version cannot be applied to, and hence making it more versatile in some situations. Currently the method focuses on the setting when data is missing in the outcome variable, however the new version of the method could be expanded in future work to missing covariate data.

We aim to publish this chapter once the method has been extended to work on missing covariates. We will aim the publication at the 'Statistics in Medicine' audience, so that researchers will be able to perform sensitivity analysis when the sample size is small, and/or the researcher wishes to look at non-local sensitivity analysis.

This thesis aimed to address the research question of why is sensitivity after multiple imputation under MAR is important, and tackle barriers to its routine use in practice. The first part of the question has been explored and illustrated in the Chapter 2, 3 and 4. The second part, relating to practical and methodological barriers is addressed in different aspects throughout.

As multiple imputation is increasingly being applied we wish to promote sensitivity analysis as part of the standard procedure. To further aid the application of sensitivity analysis after MI we suggest some possible extensions. As not all analysts are able to write problem specific code, we recommend that software for the pattern mixture approach we have explored

is made more readily available. This would include the multivariate approach we used in REALCOM, so that the multilevel imputation under MNAR with a prior would become easier. In the colorectal cancer chapter we used elicitation to find a prior to create the MNAR imputations, and we discussed the difficulties we had with the elicitation. As an extension we recommend the development of elicitation method for use in both the pattern mixture and selection modelling frameworks, together with MI software to incorporate the resulting prior information into the analysis. We also recommend future work investigating other methods for gaining the MNAR prior, we could look at estimating the MNAR distribution from a different reference group. For example, in a case/control study we could impute cases (e.g. alcohol use) from a subset of the controls, or use other observational data. These recommendations would hopefully make the process of looking at robustness of inferences to MNAR quicker, easier, more accessible and hence make more routine in practice.

Sensitivity analysis after multiple imputation is an important field which needs more applied methodology developed but also just as importantly more good examples to be published. We believe that this will broaden the awareness and understanding of the assumptions used within applying multiple imputation using standard software and how robustness of inferences to these assumptions can be explored. We hope this thesis has made a contribution to both methodology and practice and will thus facilitate more routine use of these methods in the future.

Bibliography

Aitkin, M., Anderson, D. and Hinde, J. (1981) Statistical modeling on data of teaching on styles (with disscussion). *Royal Statistical Society*, (144), 148–161.

Allison, P. D. (2001) *Missing Data*. Sage University Papers Series on Quantitative Applications in Social Sciences.

Andridge, R. R. and Little, R. (2011) A review of hot deck imputation for survey non-response. *Int Stat Rev*, (78), 40–64.

Arbuckle, J. L. (1996) *Structural Equation Modelling: Full information estimation in the presence of incomplete data*. Lawrence Erlbaum Associates. 243-277.

Azur, M. J., Stuart, E. A., Frangakis, C. and Leaf, P. J. (2011) Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*, (20), 40–49.

Barnard, J. and Rubin, D. B. (1999) Small-sample degrees of freedom with multiple imputation. *Biometrika*, (86), 949–955.

Beale, E. L. and Little, R. J. A. (1975) Missing values in multivariate analysis. *Journal of the Royal Statistical Society*, (Series B, 37), 129–145.

Bennett, N. (1976) *Teaching Styles and Pupil Progress*. Open Book.

Bousquet, A. H., Desenclos, J. C., Larsen, C., Strat, Y. L. and Carpenter, J. R. (2012) Practical considerations for sensitivity anaysis after multiple imputation applied to epidemiological studies with incomplete data. *BMC Medical Research Methodology*, (12), 73.

Brand, J. P. (1999) Development implementation and evaluation of multiple imputation strategies for the statical analysis of incomplete data sets, phd thesis. *Erasmus University*.

Buuren, S. V., Brands, J. P., Groothuis-oudshoorn, C. G. M. and Rubin, D. B. (2006) Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, (76), 1049–1064.

Carlin, J. B., Galati, J. C. and Royston, P. (2008) A new framework for managing and analyzing multiply imputed data. *Stata.Stata Journal*, (8), 49–67.

Carpenter, J. and Kenward, M. (2007a) Missing data in clinical trials- a practical guide. *UK National Health Service. National Centre for Research on Methodology*.

Carpenter, J. R. and Kenward, M. G. (2007b) Missing data in randomised controlled trials- a practical guide. *http://www.missingdata.org.uk/ (accessed July 2014)*.

Carpenter, J. R. and Kenward, M. G. (2013) *Multiple Imputation and its Applicaion*. Wiley.

Carpenter, J. R., Pocock, S. and Lamm, C. J. (2002) Coping with missing data in clinical trials: a model based approach applied to asthma trials. *Statistics in Medicine*, (21), 1043–1066.

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

Carpenter, J. R., Kenward, M. and White, I. (2007) Sensitivity analysis after multiple imputation under missing at random: A weighting approach. *Statistical Methods in Medical Research*, (16), 259–276.

Carpenter, J. R., Goldstein, H. and Kenward, M. G. (2011a) Realcom-impute software for multilevel multiple imputation with mixed response types. *Journal of Statistical Software*, (45), 1–14.

Carpenter, J. R., Rucker, G. and Schwarzer, G. (2011b) Assessing the sensitivity of meta-analysis to selection bias: A multiple imputation approach. *Biometrics*, (67), 1066–1072.

Carpenter, J. R., Goldstein, H. and Kenward, M. G. (2012a) *Modern methods for epidemiology: Statistical modelling of partially observed data using multiple imputation: principles and practice.* New York: Springer. 15-23.

Carpenter, J. R., Roger, J. H. and Kenward, M. (2012b) Analysis of longitudinal trials with protocol deviation: A framework for relevant, accessible assumptions, and inference via multiple imputation. *Journal of Biopharmaceutical Statistics*, (23), 1352–1371.

Carpenter, R., McGarvey, C., Mitchell, E. A., Tappin, D. M., Vennemann, M. M., Smuk, M. and Carpenter, J. R. (2013) Research bed sharing when parents do not smoke: is there a risk of sids? an individual level analysis of five major casecontrol studies. *BMJ Open*, (3).

Carpenter, R. G. (2015) Personal communication.

Carpenter, R. G., Irgens, L. M., Blair, P., England, P. D., Fleming, P., Huber, J., Jorch, G. and Schreuder, P. (2004) Sudden unexplained infant death in Europe: findings of the European oncerted Action on SIDS, ECAS. *Lancet*, (363), 185–191.

Clayton, D., Spiegelhalter, D., Dunn, G. and Pickles, A. (1998) Analysis of longitudinal binary data from multi-phase sampling with discussion. *Journal of the Royal Statistical Society*, (Series B), 71–87.

Cox, D. R. and Hinkley, D. V. (1974) *Theoretical Statisitcs*. Chapman and Hall.

CR (2013) Cancer Research UK: Bowel cancer statistics (Accessed 2014). *http://www.cancerresearchuk.org/cancer-info/cancerstats/keyfacts/bowel-cancer/*.

Damien, P., Dellaportas, P., Polson, N. G. and Stephens, D. A. (2013) *Bayesian Theory and Applications*. Oxford University Press.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum-likelihood from incomplete data via the EM algorithm. *Journal for Royal Statistical Society*, (Series B), 1–38.

Fessler, J. A. and Hero, A. O. (1995) Penalized Maximum-Likelihood Image Reconstruction using Space-Alternating Generalized EM Algorithms. *Image Processing*, (4), 1417–1429.

Findeise, M., Vennemann, M., Brinkmann, B., Ortmann, C., Rse, I., Kpcke, W. and Bajanowski, G. J. T. (2004) German study on sudden infant death (GeSID): design, epidemiological and pathological profile. *International journal legal medicine*, (118), 163–169.

Gelman, A. (2006) Multilevel (hierarchical) modeling: what it can and can't do. *Technometrics*, (48), 432–435.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996) *Markov Chain Monte Carlo in practice*. London: Chapman & Hall.

Goldstein, H., Carpenter, J. R., Kenward, M. G. and Levin, K. (2009) Multilevel models with multivariate mixed response types. *Statistical Modelling*, (9), 173–197.

Goldstein, H., Rasbash, J., Steele, F., Charlton, C., Browne, H. and Pollard, S. (2013) REALCOM. *Univeristy of Bristol*.

Harel, O. and Carpenter, J. R. (2014) Complete records regression with missing data: relating bias in coefficient estimates to the missingness mechanism, in preparation.

He, Y., Zaslavsky, A. M., Landrum, M. B., Harrington, D. P. and Catalano, P. (2010) Multiple imputation in a large scale complex survey: a practical guide. *Statistical Methods*, (19), 653–670.

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

Heitjan, D. F. and Rubin, D. B. (1991) Ignorability and coarse data. *Institute of Mathematical Statistics*, (19), 2244–2253.

HES (2009) Hospital episode statistics. *http://www.hesonline.nhs.uk*, (Accessed June 2010).

Heyting, A., Tolboom, J. and Essers, J. (1993) Response to letter to the editor. *Statistics in Medicin*, (12), 2248–2250.

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

Hughes, R. A., White, I. R., Seaman, S. R., Carpenter, J. R., Tilling, K. and Sterne, J. A. (2014) Joint modelling rationale for chained equations. *BMC Medical Research Methodology*, (14), 28.

Jackman, S. (2000) Estimation and Inference via Bayesian Simulation: An Introduction to Markov Chain Monte Carlo. *American Journal of Political Science*, (44), 375 – 404.

Kadane, J. B. and Wolfson, L. J. (1998) Experiences in elicitation. *Statistician*, (47), 3–19.

Kang, J. D. Y. and Schafer, J. L. (2007) Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Statistical Science*, (22), 523–580.

Kenward, M. G. and Carpenter, J. R. (2009) *Handbooks of Modern Statistical Methods: Longitudinal Data Analysis*. Chapman and Hall.

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

Li, K. H. (1988) Imputation using Markov chains. *Journal of Statistical Computation and Simulation*, (30), 57–79.

Lipsitz, S. R., Fitzmaurice, G. M., Ibrahim, J. G., Sinha, D., Parzen, M. and Lipshultz, S. (2009) Joint generalized estimating equations for multivariate longitudinal binary outcomes with missing data: an application to acquired immune deficiency syndrome data. *Journal of the Royal Statistical Society*, (172), 3–20.

Little, R. and Rubin, D. (2002) *Statistical analysis with missing data*. Hoboken, NJ : Wiley.

Little, R. and Yau, L. (1996) Intent to treat analysis for longitudinal studies with drop-outs. *Biometrics*, (52), 1324–1333.

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

Liu, C. H. and Rubin, D. B. (1994) The ECME algorithm: a simple extension of EM and ECM with fast montone convergence. *Biometrika*, (81), 633–648.

Liu, C. H., Rubin, D. B. and Wu, Y. (1998) Parameter expansion to accelerate EM: the PX-EM algorithm. *Biometrika*, (85), 755–770.

Louis, T. A. (1982) Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society*, (44), 226–233.

Lullaby (2014) The Lullaby Trust (Last accessed April 2014). *http://www.lullabytrust.org.uk/?gclid=CNyz-4fu5L0CFaQfwwodkhMASg*.

Mackinnon, A. (2010) The use and reporting of multiple imputation in medical research - a review. *Journal of International Medicine*, (268), 586–593.

McGarvey, C., McDonnell, M., Hamilton, K., Oregan, M. and Matthew, T. (2006) An eight-year study of risk factors for SIDS: Bed-sharing vs. non bed-sharing. *Archives of disease in childhood*, (91), 185–191.

Mckendrick, A. G. (1926) Applications of mathematical problems. *Proceedings Edinburgh Mathematical Society*, (44), 98–130.

Meng, X. L. (1994) Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science*, (9), 538–573.

Meng, X. L. and van Dyk, D. A. (1997) The EM algorithm-an old folk song sung to a fast new tune (with discussion). *Royal Statistical Society*, (59), 511–567.

Meng, X. L. and Pedlow, S. (1992) EM: A bibliographic review with missing articles. *American Statistical Association*, (Computing Section), 24–27.

Meng, X. L. and Rubin, D. B. (1993) Maximum Likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, (80), 267–278.

Mitchell, E. A., Taylor, B. J. and Ford, R. P. (1992) Four modifiable and other major risk factors for cot death: The new zealand study. *Journal of Pediatric Child Health*, (28), 3–8.

Molenberghs, G., Thijs, H., Jansen, I., Beunkens, C., Kenward, M. G., Mallinkrodt, C. and Carroll, R. J. (2004) Analyzing incomplete longitudinal clinical trial data. *Biostatistics*, (5), 445–464.

Morris, E. J. A., Taylor, E. F., Thomas, J. D., Quirke, P., Finan, P. J., Coleman, M. P., Rachet, B. and Forman, D. (2011) Thirty-day postoperative mortality after colorectal cancer surgery in england. *GUT*, (60), 806–813.

Myers (2000) Handling missing data in clinical trials: An overview. *Drug Information Journal*, (34), 523–533.

Nakai, M. and Ke, W. (2011) Review of the methods for handling missing data in longitudinal data analysis. *International Journal of Mathematics*, (5), 1–13.

NCDR (2009) National Cancer Intelligence Network. *http://www.ncin.org.uk*, (Accessed June 2010).

NCIN (2014) National Cancer Intelligence Network. *http://www.ncin.org.uk/*, (Accessed March 2014).

O'Hagan, A. (1998) Eliciting expert beliefs in substantial practical applications. *The Statistician*, (47), 55–68.

Orchard, T. and Woodbury, M. A. (1972) A missing information principle: theory and applications. *Symposium on Mathematical Statistics and Probability*, (1), 697–715.

Pawitan, Y. (2001) *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press.

Pocock, S. J. (1983) *Clinical trials: A practical approach*. Wiley & Sons.

Pocock, S. J. (1996) *Clincial trials: A statistician's perspective*. 405-421. Wiley.

Ripley, B. D. (1987) *Stochastic simulation*. Wiley.

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

Robins, J. M. and Wang, N. (2000) Inference in imputation estimators. *Biometrika*, (87), 113–124.

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

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, (90), 106–129.

Rubin, D. B. (1976) Inference and missing data. *Biometrika*, (63), 581–592.

Rubin, D. B. (1987) *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Rubin, D. B. (1996) Multiple imputation after 18+ years. *Journal of the American Statistical Association*, (91), 473.

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

Saltelli, A. and Annoni, P. (2010) How to avoid a perfunctory sensitivity analysis. *Environmental Modelling & Software*, (25), 1508–1517.

Schafer, J. L. (1999) *Analysis of Incomplete Multivariate Data*. Chapman and Hall/CRC.

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

Schafer, J. L. and Graham, J. W. (2002) Missing data: our view of the state of the art. *Psychological Methods*, (7), 147–177.

Scharfstein, D. O., Rotnizky, A. and Robins, J. M. (1999) Adjusting for nonignorable drop-out using semi-parametric nonresponse models with comments. *Journal of the American Statistical Association*, (94), 1096–1146.

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

StataCorp (2011a) *Stata 12 Multiple-Imputation Reference Manual*. Stata Press.

StataCorp (2011b) *Stata 12 Lrtest*. Stata Press.

Streiner, D. L. (2008) Missing data and the trouble with LOCF. *Evidence Based Mental Health*, (11), 3–5.

Tappin, D., Ecob, R. and Brooke, H. (2005) Bedsharing, roomsharing and sudden infant death syndrome in scotland: a casecontrol study. *Journal of Pediatrics*, (147), 32–37.

Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M. (2006) Hierarchical dirichlet processes. *Journal of the American Statistical Association*, (476), 1566–1581.

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

Tsiatis, A. A. (2006) *Semiparametric Theory and Missing Data*. Springer, NewYork.

Vansteelandt, S., Carpenter, J. and 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), 37–48.

White, I. R. and Carlin, J. B. (2010) Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics In Medicine*, (29), 2920–2931.

White, I. R., Carpenter, J., Stephen, E. and Schroter, S. (2007) Eliciting and using expert opinions about dropout bias in randomised controlled trails. *Clinical Trials*, (4), 125–139.

Woolson, R. F. and Clarke, W. R. (1984) Analysis of categorical incomplete longitudinal data. *Royal Statistical Society*, (Series A, 147), 87–99.

A

Chapter 1 Appendix: Introduction

Data bases used for literature search:

Academic Search Complete, BASE, Clinical Trials, Current Control Trials, EMBASE, Global Health, HEER, HISA, IBSS, MEDLINE, PubMed, Scirus, Scopus, Trip Database, Web Science, Missing Data.org

B

Chapter 3 Appendix: Estimating Cancer Survival from Registry Data: A practical framework for understanding the impact of missing data on conclusions

B.1 Chapter 3 Appendix: Colorectal Cancer Data Variables

The table below lists the variables used in the colorectal cancer dataset.

Variable	Information
Postoperative mortality in 30 days (MORT)	Binary yes or no.
Age at diagnosis	In years.
Year of diagnosis (YOD)	Actual year date.
Year of operation	Actual year date.
Sex	
Medium annual workload of a trust (MAWT)	Hours
Operation Type (OT)	Elective or emergency operation, emergency is defined as being operated on within 2 days of hospital admission.
Duke's stage at diagnosis	Four stages: A, B, C and D (A is the least severe, D is the most severe).
Index for Multiple Deprivation (IMD) income categories	Scale 1 to 5: 1 is the most affluent, 5 is the most deprived.
Cancer site (CS)	Colon, Rectosigmoid or Rectum.
Charlson comorbidity score (CCS)	Created a scale from 0 to 3+ from diagnostic codes (excluding cancer) for hospital admissions a year prior to being diagnosed with colorectal cancer. 0 is the lowest risk and 3+ is the greatest risk.
Admission type (AMISS)	Elective or emergency admission.
Hospital trust	
Cancer registry	

Table B.1: Information on variables used from the colorectal cancer dataset.

B.2 Chapter 3 Appendix: Cancer Fully Conditional Specification Algorithm

Assume we have r variables, for each partially observed variable Y_k we create the imputation model $f(Y_k|Y_{-k}, Z, \theta_k)$ where $Y_1, \ldots Y_{k-1}, Y_{k+1}, \ldots, Y_r$, θ_k is a parameter and Z contains fully

observed variables. The imputation model is chosen to reflect the type of Y_k, for example if Y_k is binary then $f()$ will be a logistic regression model. Let Y_k^o and Y_k^m denote vectors of the observed and missing values respectively in Y_k. We also require a non-informative prior distribution $p(\theta_k)$ for θ_k. The fully conditional specification algorithm is a followed for each iteration t:

- Replace all missing values in each Y_k with a random selection of observed values from Y_k.
- Impute each Y_k^m at each stage from the conditional distribution of Y_k^o on the most recent imputation of the other variables in the preceding step. Let $Y_k^{m(t)}$ denote the imputation of the missing values of Y_k^m and $Y_k^{(t)}$ represents the vector of observed and imputed value of Y_k at iteration t. Thus for the t^{th} iteration:

$$\begin{aligned}
\theta_1^{(t)} &\sim p(\theta_1) f(Y_1^o | Y_2^{(t-1)}, \ldots, Y_r^{(t-1)}, Z, \theta_1) \\
Y_1^{m(t)} &\sim f(Y_1^m | Y_2^{(t-1)}, \ldots, Y_r^{(t-1)}, Z, \theta_1^{(t)}) \\
&\vdots \qquad\quad \vdots \qquad\quad \vdots \qquad\quad \vdots \\
\theta_r^{(t)} &\sim p(\theta_r) f(Y_r^o | Y_1^{(t)}, \ldots, Y_{r-1}^{(t)}, Z, \theta_r) \\
Y_r^{m(t)} &\sim f(Y_r^m | Y_1^{(t)}, \ldots, Y_{r-1}^{(t)}, Z, \theta_r^{(t)})
\end{aligned} \tag{B.2.1}$$

- The algorithm will cycle around n times which will hopefully find the stationary distribution (convergence can be checked graphically). The final draw is used as a single imputed dataset.
- The process is repeated to create the specified number of imputed dataset required.

B.3 Chapter 3 Appendix: Questionnaire

Questionnaire Information for Sensitivity Analysis with Multiple Imputation.

The Study

The study aim is to investigate factors affecting thirty day postoperative mortality after colorectal cancer surgery in England. The data for this analysis consists of pooled data from eight population based cancer registries that cover England. This was matched with data from the hospital episode statistics data base extracting patients who "presented in any NHS hospital with a diagnostic code for cancer between April 1997 and June 2007. Information was extracted from this resource on all individuals who underwent a major resection for a primary colorectal cancer diagnosed between 1 January 1998 and 31 December 2006" Morris *et al.* (2011). There are 160,920 patients in the dataset of whom 85% have complete records. We aim to assess the sensitivity of the inferences to the assumptions about the reason for the missing data.

Missing Data

Three variables within the dataset contained missing observations. The variables are Duke's stage (15% missing), index of multiple deprivation (0.25% missing) and emergency admission indicator (0.05% missing). The focus is the sensitivity of inferences to different assumptions about the missing Dukes stage category (4 ordered categories, with A the least severe and D the most severe).

The default assumption is that the distribution of Duke's stage, (i.e. the probability that a patient has a particular stage) given other covariates and postoperative mortality, is the same whether or not Duke's stage is actually observed. This is known as the missing at random (MAR) assumption.

The alternative is that, the probability that a patient has a particular stage, given other covariates and postoperative mortality differs between patients with Duke's stage missing and observed.

What we would like you to do

We would like you to indicate your views about the probability a patient having a particular Duke's stage among those where Dukes stage is missing. We ask for probabilities separately for patients who did and did not survive 30 days after surgery, and for patients aged less than or equal to 70 and over 70.

In the attached spreadsheet, for each combination of age and mortality, we would like you to enter what you think the probability of Dukes stage being A-D. Your probabilities must sum to 1 (check cell will turn green). To help you, the spreadsheet will automatically graph your probabilities alongside those predicted from the data.

As you fill in the probabilities, the peach colour cells tell you the maximum / minimum probabilities you can submit in for the particular cell. The cells will turn yellow if your entry is less than the data prediction or blue if it is greater than the data prediction.

Helpful example

For a patient with missing Duke's stage who is under 70 and died within 30 days of surgery, the middle row of the Table below shows the probability, derived from the observed data, that they would have been diagnosed with Dukes' stage A ,B, C or D. The third row of the table shows the corresponding probabilities, which an idiosyncratic clinician expected them to be.

	Probability that Dukes stage is:				
	A	B	C	D	Sum A, B, C, D
Derived from the data	0.075	0.311	0.411	0.203	1
Idiosyncratic clinician's prediction	0.01	0.09	0.550	0.350	1

Table B.2: Probability associated with Duke's stage derived from the data and an idiosyncratic clinicians point of view.

Thus, while our prediction from the observed data is that 8% have Dukes stage A, 31% stage B, 41% stage C and 20% stage D, the idiosyncratic clinician differs from this, giving 1% having stage A, 9% stage B, 55% stage C and 35% stage D.

If anything is not clear, or you have any comments then please contact Mel by email on melanie.smuk@lshtm.ac.uk. Thank you for your help.

What are your probability predictions for patients with missing stage, given age and mortality?

"Your Prediction" is an **increase** from the "Data prediction"

"Your Prediction" is a **decrease** from the "Data prediction"

Maximum = The maximum percentage that can be applied for that stage, dependent on other stages

Minimum = The minimum percentage that can be applied for that stage, dependent on other stages

Alive After 30 Days and Less Than or Equal to 70 Years Old

	Duke's Stage A	B	C	D	Check Add To One
Data prediction	0.132	0.354	0.393	0.121	
Your prediction					← 0.000
Maximum	1.000	1.000	1.000	1.000	GREEN OK
Minimum	0.000	0.000	0.000	0.000	

Dead After 30 Days and Less Than or Equal to 70 Years Old

	Duke's Stage A	B	C	D	Check Add To One
Data prediction	0.075	0.311	0.411	0.203	
Your prediction					← 0.000
Maximum	1.000	1.000	1.000	1.000	GREEN OK
Minimum	0.000	0.000	0.000	0.000	

Alive After 30 Days and Greater Than 70 Years Old

	Duke's Stage A	B	C	D	Check Add To One
Data prediction	0.126	0.432	0.359	0.082	
Your prediction					← 0.000
Maximum	1.000	1.000	1.000	1.000	GREEN OK
Minimum	0.000	0.000	0.000	0.000	

Dead After 30 Days and Greater Than 70 Years Old

	Duke's Stage A	B	C	D	Check Add To One
Data prediction	0.074	0.393	0.389	0.144	
Your prediction					← 0.000
Maximum	1.000	1.000	1.000	1.000	GREEN OK
Minimum	0.000	0.000	0.000	0.000	

Figure B.1: Screen shot of electronic questionnaire.

B.4 Chapter 3 Appendix: Elicitation Results

	Dukes stage A	Dukes stage B	Dukes stage C	Dukes stage D
r=1				
Data Prediction	0.132	0.354	0.393	0.121
Expert Prediction				
Responder 1	0.100	0.150	0.200	0.550
Responder 2	0.160	0.330	0.330	0.180
Responder 3	0.250	0.400	0.200	0.150
Responder 4	0.100	0.400	0.300	0.200
Responder 5	0.330	0.330	0.330	0.010
Responder 6	0.210	0.400	0.320	0.070
r=2				
Data Prediction	0.075	0.311	0.411	0.203
Expert Prediction				
Responder 1	0.100	0.200	0.300	0.400
Responder 2	0.050	0.100	0.400	0.450
Responder 3	0.010	0.200	0.400	0.390
Responder 4	0.080	0.300	0.400	0.220
Responder 5	0.050	0.150	0.400	0.400
Responder 6	0.050	0.200	0.400	0.350
r=3				
Data Prediction	0.126	0.432	0.359	0.082
Expert Prediction				
Responder 1	0.100	0.150	0.200	0.550
Responder 2	0.280	0.380	0.280	0.060
Responder 3	0.250	0.450	0.200	0.100
Responder 4	0.100	0.390	0.300	0.210
Responder 5	0.400	0.400	0.100	0.100
Responder 6	0.160	0.450	0.350	0.040
r=4				
Data Prediction	0.074	0.393	0.389	0.144
Expert Prediction				
Responder 1	0.100	0.200	0.300	0.400
Responder 2	0.100	0.180	0.360	0.360
Responder 3	0.150	0.400	0.300	0.150
Responder 4	0.050	0.350	0.380	0.220
Responder 5	0.050	0.200	0.300	0.450
Responder 6	0.020	0.330	0.550	0.100

Table B.3: Results from the data and elicitation on missing data probabilities for each Dukes stage A to D.

B.5 Chapter 3 Appendix: Variability in MAR and MNAR Imputations

	Multiple Imputation MAR			Multiple Imputation MNAR		
Characteristic	Within	Between	Total	Within	Between	Total
Age at diagnosis (per 10 years)	1.71E-06	1.25E-08	1.72E-06	1.69E-06	3.75E-09	1.69E-06
Year of diagnosis (per advancing year)	1.55E-05	1.03E-08	1.55E-05	1.55E-05	1.02E-08	1.55E-05
Sex						
Female						
Male	6.70E-04	2.52E-06	6.73E-04	6.60E-04	3.08E-07	6.67E-04
Operation						
Elective						
Emergency	5.35E-03	1.33E-05	5.37E-03	5.69E-03	1.94E-05	5.72E-03
Dukes' stage at diagnosis						
A						
B	2.49E-03	3.53E-04	2.91E-03	1.92E-03	2.79E-04	2.26E-03
C	3.76E-03	1.18E-04	3.90E-03	2.39E-03	4.33E-04	2.91E-03
D	1.30E-02	4.20E-04	1.35E-02	5.96E-03	7.63E-04	6.88E-03
IMD income category						
Most affluent						
2	1.28E-03	1.05E-06	1.28E-03	1.28E-03	3.27E-07	1.28E-03
3	1.48E-03	2.72E-06	1.48E-03	1.48E-03	6.42E-07	1.48E-03
4	1.80E-03	1.81E-06	1.80E-03	1.79E-03	6.43E-07	1.79E-03
Most deprived	2.41E-03	1.85E-06	2.41E-03	2.41E-03	2.35E-07	2.41E-03
Cancer site						
Colon						
Rectosigmoid	1.32E-03	1.13E-06	1.32E-03	1.32E-03	5.81E-07	1.31E-03
Rectum	6.60E-04	1.35E-06	6.62E-04	6.22E-04	1.88E-06	6.24E-04
Charlson comorbidity score						
0						
1	3.73E-03	3.79E-06	3.73E-03	3.75E-03	4.08E-06	3.76E-03
2	8.95E-03	5.87E-06	8.96E-03	8.87E-03	1.90E-05	8.89E-03
≥ 3	4.55E-02	9.22E-05	4.56E-02	4.45E-02	5.42E-05	4.46E-02

Table B.4: Within, between and total variance from the multivariate analyses on the MOI for the 'Missing At Random' missing data assumption and 'Missing Not At Random' missing data assumption, based on 10 imputations.

C

Chapter 4 Appendix: Sudden Infant Death Syndrome

Degrees of Freedom	Level 1 Covariance Matrix		Level 1 Covariance Standard Deviations		Level 1 Correlation Matrix	
S No Prior	1.000	-0.002	0.000	0.043	1.000	-0.002
	-0.002	1.000	0.043	0.000	-0.002	1.000
4	1.000	-0.002	0.000	0.043	1.000	-0.002
	-0.002	1.000	0.043	0.000	-0.002	1.000
5	1.000	-0.002	0.000	0.043	1.000	-0.002
	-0.002	1.000	0.043	0.000	-0.002	1.000
10	1.000	-0.002	0.000	0.043	1.000	-0.002
	-0.002	1.000	0.043	0.000	-0.002	1.000
15	1.000	-0.001	0.000	0.043	1.000	-0.001
	-0.001	1.000	0.043	0.000	-0.001	1.000
20	1.000	-0.002	0.000	0.043	1.000	-0.002
	-0.002	1.000	0.043	0.000	-0.002	1.000
25	1.000	-0.002	0.000	0.043	1.000	-0.002
	-0.002	1.000	0.043	0.000	-0.002	1.000
30	1.000	-0.001	0.000	0.043	1.000	-0.001
	-0.001	1.000	0.043	0.000	-0.001	1.000
100	1.000	-0.001	0.000	0.043	1.000	-0.001
	-0.001	1.000	0.043	0.000	-0.001	1.000

Table C.1: Table showing the effect on the SIDS cases level 1 covariance matrix when varying the degrees of freedom in the inverse Wishart prior.

DF	Level 2 Covariance Matrix				Level 2 Covariance Standard Deviations				Level 2 Correlation Matrix			
NP	4.036	-3.446	7.232	8.852	7.555	32.919	34.225	74.574	1.000	-0.076	0.155	0.084
	-3.446	505.720	158.550	-210.040	32.919	748.130	521.480	1178.900	-0.076	1.000	0.304	-0.179
	7.232	158.550	538.870	-604.260	34.225	521.480	707.380	1101.800	0.155	0.304	1.000	-0.498
	8.852	-210.040	-604.260	2728.700	74.574	1178.900	1101.800	4014.300	0.084	-0.179	-0.498	1.000
4	0.135	0.027	0.120	0.032	0.118	0.073	0.121	0.078	1.000	0.262	0.665	0.320
	0.027	0.076	0.047	0.014	0.073	0.106	0.119	0.065	0.262	1.000	0.348	0.184
	0.120	0.047	0.241	0.032	0.121	0.119	0.344	0.111	0.665	0.348	1.000	0.237
	0.032	0.014	0.032	0.074	0.078	0.065	0.111	0.105	0.320	0.184	0.237	1.000
5	0.109	0.026	0.090	0.028	0.085	0.056	0.083	0.059	1.000	0.324	0.685	0.355
	0.026	0.061	0.040	0.011	0.056	0.083	0.083	0.056	0.324	1.000	0.403	0.181
	0.090	0.040	0.158	0.023	0.083	0.083	0.164	0.079	0.685	0.403	1.000	0.244
	0.028	0.011	0.023	0.057	0.059	0.056	0.079	0.083	0.355	0.181	0.244	1.000
10	0.051	0.010	0.031	0.011	0.039	0.020	0.031	0.021	1.000	0.273	0.598	0.316
	0.010	0.025	0.011	0.004	0.020	0.021	0.023	0.014	0.273	1.000	0.296	0.159
	0.031	0.011	0.051	0.009	0.031	0.023	0.048	0.021	0.598	0.296	1.000	0.260
	0.011	0.004	0.009	0.025	0.021	0.014	0.021	0.021	0.316	0.159	0.260	1.000
100	0.011	0.000	0.000	0.000	0.002	0.001	0.001	0.001	1.000	0.015	0.028	0.014
	0.000	0.011	0.000	0.000	0.001	0.002	0.001	0.001	0.015	1.000	0.012	0.006
	0.000	0.000	0.011	0.000	0.001	0.001	0.002	0.001	0.028	0.012	1.000	0.010
	0.000	0.000	0.000	0.011	0.001	0.001	0.001	0.002	0.014	0.006	0.010	1.000

Table C.2: Table showing the effect on the SIDS cases level 2 covariance matrix when varying the degrees of freedom (DF) in the inverse Wishart prior against no prior (NP).

Covariate	Degrees of Freedom								
Alcohol Model	No Prior	4	5	10	15	20	25	30	100
Constant	3.24	2.92	2.92	2.98	3.09	3.14	3.17	3.19	3.24
Mothers Age	-0.21	-0.24	-0.25	-0.27	-0.30	-0.31	-0.33	-0.33	-0.34
Bed Share	8.07	-0.49	-0.47	-0.53	-0.59	-0.64	-0.65	-0.66	-0.69
Bottle Fed	0.13	-0.01	-0.01	-0.08	-0.14	-0.17	-0.20	-0.22	-0.24
Interaction A (2)	0.55	0.56	0.57	0.57	0.58	0.58	0.58	0.59	0.59
Interaction A (3)	1.62	1.66	1.63	1.68	1.66	1.69	1.71	1.70	1.72
Interaction A (4)	0.43	0.45	0.46	0.48	0.51	0.53	0.53	0.55	0.55
Interaction A (5)	0.47	0.53	0.54	0.55	0.57	0.58	0.59	0.59	0.59
Interaction A (6)	-0.27	-0.26	-0.26	-0.23	-0.20	-0.18	-0.17	-0.17	-0.15
Interaction A (7)	0.50	0.09	0.08	0.13	0.18	0.21	0.22	0.26	0.28
Age 6 Months	0.16	0.24	0.23	0.26	0.29	0.31	0.31	0.32	0.33
Interaction B	-0.62	-0.60	-0.59	-0.60	-0.63	-0.65	-0.64	-0.65	-0.66
Other Room	0.09	0.16	0.17	0.16	0.17	0.17	0.18	0.18	0.18
Mother Smokes	-0.09	-0.12	-0.12	-0.16	-0.20	-0.23	-0.25	-0.26	-0.28
Partner Smokes	-0.87	-0.77	-0.76	-0.73	-0.70	-0.67	-0.66	-0.65	-0.65
Birth Weight	-0.18	-0.16	-0.16	-0.15	-0.15	-0.15	-0.14	-0.14	-0.14
Married or Cohabiting	0.22	0.17	0.16	0.13	0.08	0.06	0.04	0.04	0.02
Live Births	0.81	0.81	0.85	0.86	0.88	0.88	0.89	0.91	0.91
Sex	-0.55	-0.49	-0.48	-0.47	-0.47	-0.46	-0.46	-0.46	-0.45
Drug Model									
Constant	16.07	2.40	2.37	2.22	2.14	2.10	2.08	2.06	2.04

Interaction A= Position Left In, Bed Shared, Baby Age Interaction
Interaction B= Mother Drank Alcohol, Bed Share Interaction

Table C.3: Table showing the effect on the SIDS cases covariates in the imputation model when varying the degrees of freedom in the inverse Wishart prior.

Table C.4: Controls

Degrees of Freedom	Level 1 Covariance Matrix		Level 1 Covariance Standard Deviations		Level 1 Correlation Matrix	
No Prior	1.000	0.002	0.000	0.024	1.000	0.002
	0.002	1.000	0.024	0.000	0.002	1.000
4	1.000	0.001	0.000	0.024	1.000	0.001
	0.001	1.000	0.024	0.000	0.001	1.000
5	1.000	0.001	0.000	0.024	1.000	0.001
	0.001	1.000	0.024	0.000	0.001	1.000
10	1.000	0.001	0.000	0.024	1.000	0.001
	0.001	1.000	0.024	0.000	0.001	1.000
15	1.000	0.001	0.000	0.024	1.000	0.001
	0.001	1.000	0.024	0.000	0.001	1.000
20	1.000	0.002	0.000	0.024	1.000	0.002
	0.002	1.000	0.024	0.000	0.002	1.000
25	1.000	0.001	0.000	0.024	1.000	0.001
	0.001	1.000	0.024	0.000	0.001	1.000
30	1.000	0.002	0.000	0.024	1.000	0.002
	0.002	1.000	0.024	0.000	0.002	1.000
100	1.000	0.001	0.000	0.024	1.000	0.001
	0.001	1.000	0.024	0.000	0.001	1.000

Table C.5: Table showing the effect on the SIDS controls level 1 covariance matrix when varying the degrees of freedom in the inverse Wishart prior.

DF	Level 2 Covariance Matrix				Level 2 Covariance Standard Deviations				Level 2 Correlation Matrix			
NP	135.370	155.350	157.710	1210.900	187.000	319.750	191.590	1643.200	1.000	0.307	0.665	0.614
	155.350	1894.700	337.000	2392.900	319.750	2586.700	614.570	6798.600	0.307	1.000	0.380	0.325
	157.710	337.000	415.030	2084.000	191.590	614.570	547.700	2994.000	0.665	0.380	1.000	0.604
	1210.900	2392.900	2084.000	28706.000	1643.200	6798.600	2994.000	38112.000	0.614	0.325	0.604	1.000
4	0.411	0.156	0.186	-0.095	0.699	0.308	0.458	0.328	1.000	0.493	0.643	-0.305
	0.156	0.244	0.107	-0.042	0.308	0.419	0.285	0.263	0.493	1.000	0.480	-0.173
	0.186	0.107	0.203	-0.046	0.458	0.285	0.446	0.292	0.643	0.480	1.000	-0.211
	-0.095	-0.042	-0.046	0.237	0.328	0.263	0.292	0.628	-0.305	-0.173	-0.211	1.000
5	0.356	0.145	0.173	0.045	0.884	0.371	0.555	0.937	1.000	0.567	0.717	0.157
	0.145	0.184	0.089	0.016	0.371	0.363	0.269	0.397	0.567	1.000	0.512	0.078
	0.173	0.089	0.163	0.023	0.555	0.269	0.403	0.575	0.717	0.512	1.000	0.120
	0.045	0.016	0.023	0.229	0.937	0.397	0.575	0.995	0.157	0.078	0.120	1.000
10	0.089	0.020	0.021	-0.003	0.075	0.051	0.057	0.052	1.000	0.354	0.392	-0.054
	0.020	0.037	0.009	0.001	0.051	0.054	0.041	0.038	0.354	1.000	0.264	0.016
	0.021	0.009	0.033	0.000	0.057	0.041	0.055	0.043	0.392	0.264	1.000	0.009
	-0.003	0.001	0.000	0.032	0.052	0.038	0.043	0.048	-0.054	0.016	0.009	1.000
100	0.011	0.000	0.000	0.000	0.002	0.001	0.001	0.001	1.000	0.014	0.006	-0.001
	.000	0.011	0.000	0.000	0.001	0.002	0.001	0.001	0.014	1.000	0.002	0.000
	0.000	0.000	0.011	0.000	0.001	0.001	0.002	0.001	0.006	0.002	1.000	0.000
	0.000	0.000	0.000	0.011	0.001	0.001	0.001	0.002	-0.001	0.000	0.000	1.000

Table C.6: Table showing the effect on the SIDS controls level 2 covariance matrix when varying the degrees of freedom (DF) in the inverse Wishart prior against no prior (NP).

Covariate	Degrees of Freedom								
Alcohol Model	No Prior	4	5	10	15	20	25	30	100
Constant	2.85	2.70	2.66	2.61	2.60	2.59	2.62	2.61	2.64
Mothers Age	-0.23	-0.25	-0.25	-0.27	-0.28	-0.28	-0.29	-0.29	-0.30
Bed Share	11.71	0.00	0.00	-0.15	-0.19	-0.21	-0.23	-0.22	-0.24
Bottle Fed	-0.16	-0.17	-0.17	-0.17	-0.18	-0.19	-0.19	-0.20	-0.20
Interaction A (2)	-0.20	-0.14	-0.13	-0.15	-0.16	-0.16	-0.17	-0.17	-0.17
Interaction A (3)	0.15	0.18	0.20	0.24	0.31	0.31	0.31	0.32	0.35
Interaction A (4)	0.03	0.06	0.07	0.05	0.04	0.02	0.02	0.02	0.01
Interaction A (5)	0.00	0.08	0.09	0.13	0.17	0.19	0.19	0.21	0.22
Interaction A (6)	1.06	0.96	0.93	0.86	0.86	0.87	0.85	0.85	0.85
Age 6 Months	-0.20	-0.17	-0.17	-0.16	-0.14	-0.13	-0.12	-0.12	-0.11
Interaction B	0.37	0.36	0.36	0.35	0.33	0.32	0.31	0.31	0.30
Other Room	-0.18	-0.16	-0.15	-0.13	-0.12	-0.11	-0.11	-0.11	-0.10
Mother Smokes	-0.24	-0.29	-0.30	-0.33	-0.36	-0.37	-0.38	-0.38	-0.40
Partner Smokes	-0.14	-0.13	-0.12	-0.12	-0.11	-0.10	-0.10	-0.10	-0.10
Birth Weight	0.06	0.05	0.05	0.03	0.03	0.03	0.02	0.03	0.02
Married or Cohabiting	0.13	0.10	0.10	0.11	0.08	0.09	0.07	0.07	0.07
Live Births	1.02	1.00	1.03	1.11	1.16	1.19	1.19	1.19	1.27
Sex	0.15	0.13	0.13	0.12	0.11	0.11	0.11	0.11	0.11
Drug Model									
Constant	4.71	3.12	3.13	3.00	2.97	2.97	2.97	2.95	2.97

Interaction A= Position Left In, Bed Shared, Baby Age Interaction
Interaction B= Mother Drank Alcohol, Bed Share Interaction

Table C.7: Table showing the effect on the SIDS controls covariates in the imputation model when varying the degrees of freedom in the inverse Wishart prior.

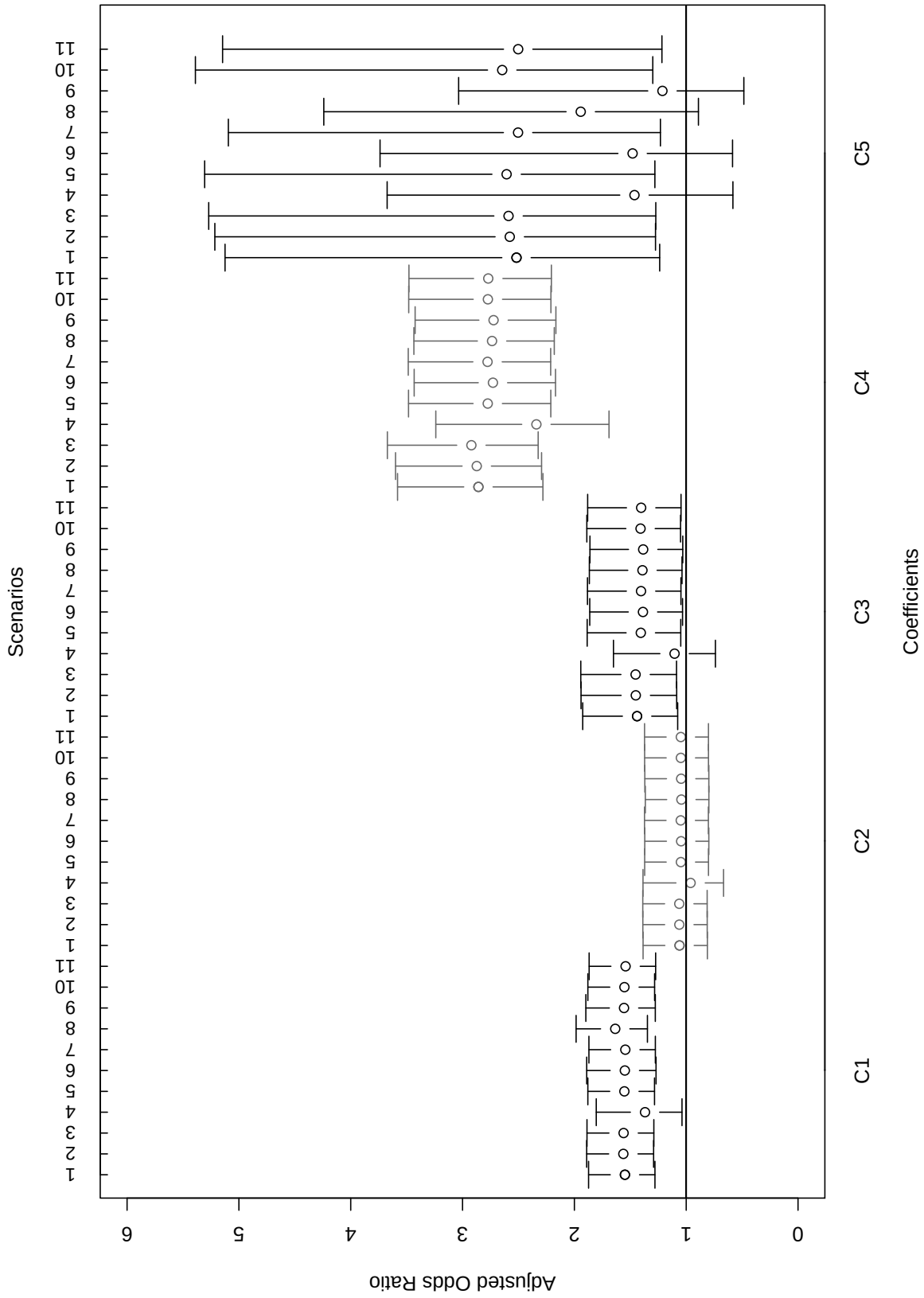

Figure C.1: Coefficient results from the adjusted odds ratio model of interest (Interaction B included) for the 11 scenarios changing Alcohol imputation. Scenarios are represented numerically, corresponding to the rows in Table 4.9; scenario 11 represents MAR. (Coefficients: C1=Bottle Fed, C2=Interaction A.2, C3=Interaction A.3, C4=Interaction A.4, C5=Interaction A.5).

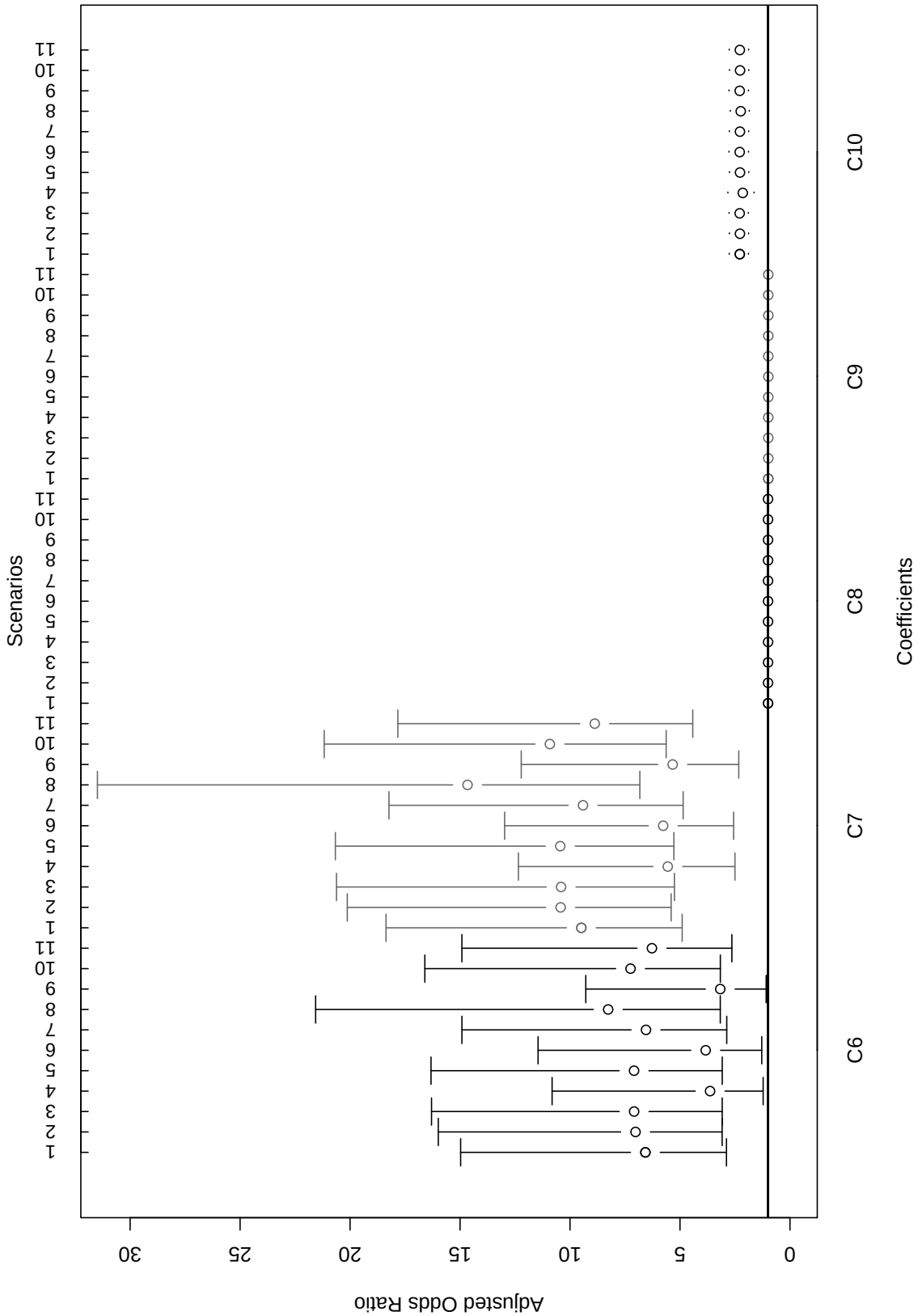

Figure C.2: Coefficient results from the adjusted odds ratio model of interest (Interaction B included) for the 11 scenarios changing Alcohol imputation. Scenarios are represented numerically, corresponding to the rows in Table 4.9; scenario 11 represents MAR. (Coefficients: C6=Interaction A.6, C7=Interaction A.7, C8=Centred Age, C9=Interaction B, C10=Other Room).

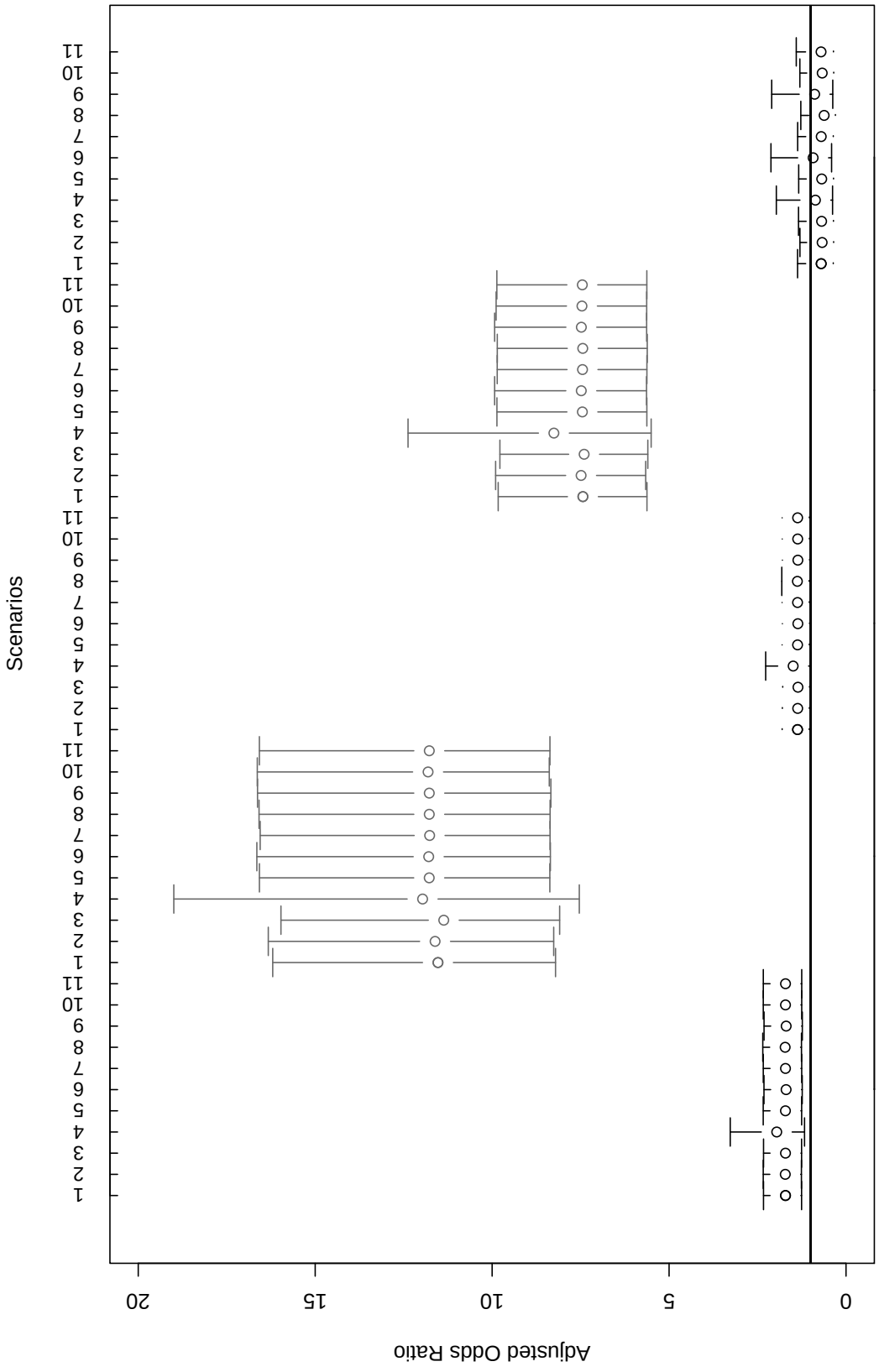

Figure C.3: Coefficient results from the adjusted odds ratio model of interest (Interaction B included) for the 11 scenarios changing Alcohol imputation. Scenarios are represented numerically, corresponding to the rows in Table 4.9; scenario 11 represents MAR. (Coefficients: C11=Interaction C.2, C12=Interaction C.3, C13=Interaction C.4, C14=Interaction C.5, C15=Interaction C.6).

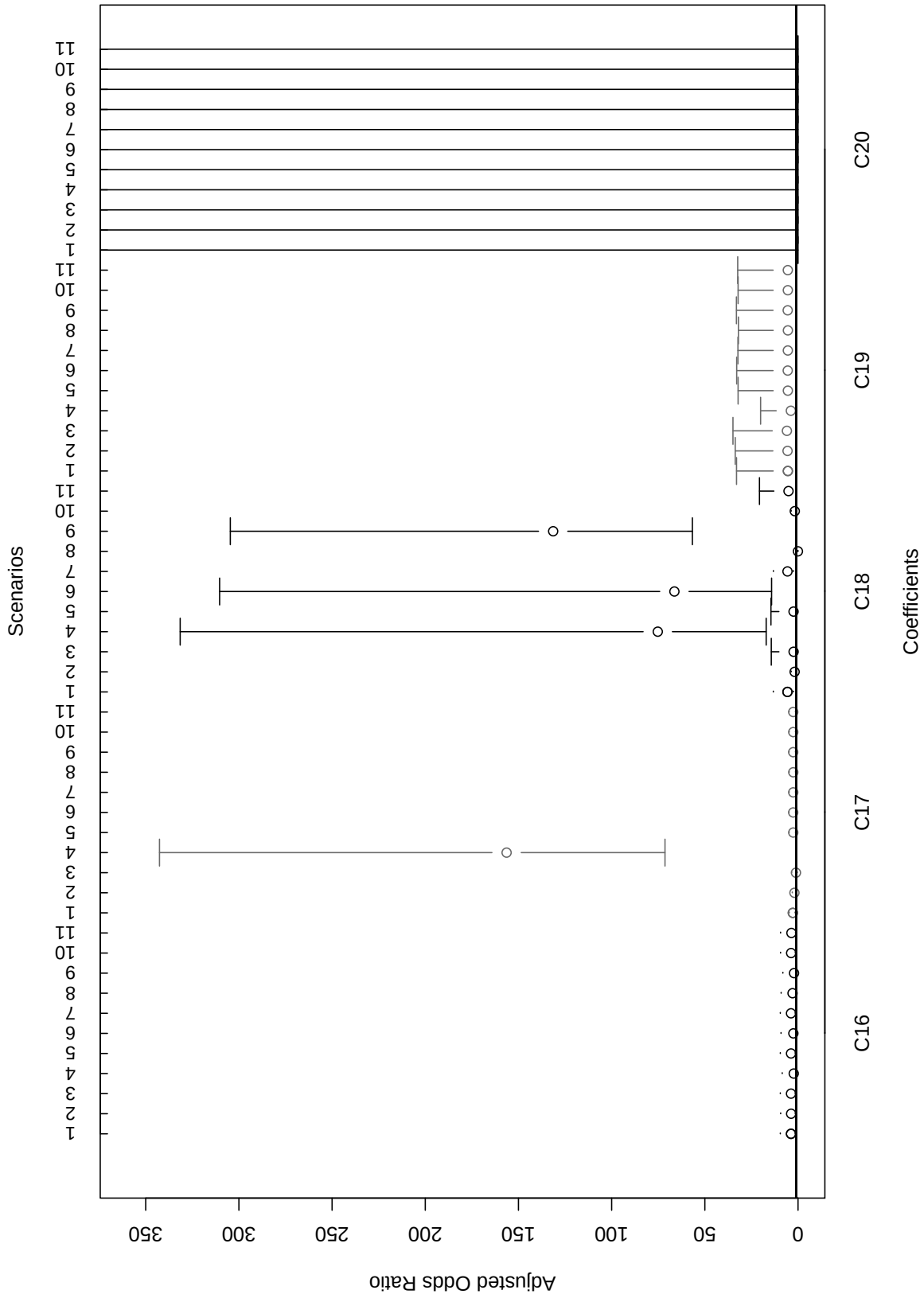

Figure C.4: Coefficient results from the adjusted odds ratio model of interest (Interaction B included) for the 11 scenarios changing Alcohol imputation. Scenarios are represented numerically, corresponding to the rows in Table 4.9; scenario 11 represents MAR. (Coefficients: C16=Interaction C.7, C17=Interaction D.2, C18=Interaction D.3, C19=Interaction E, C20=Interaction F).

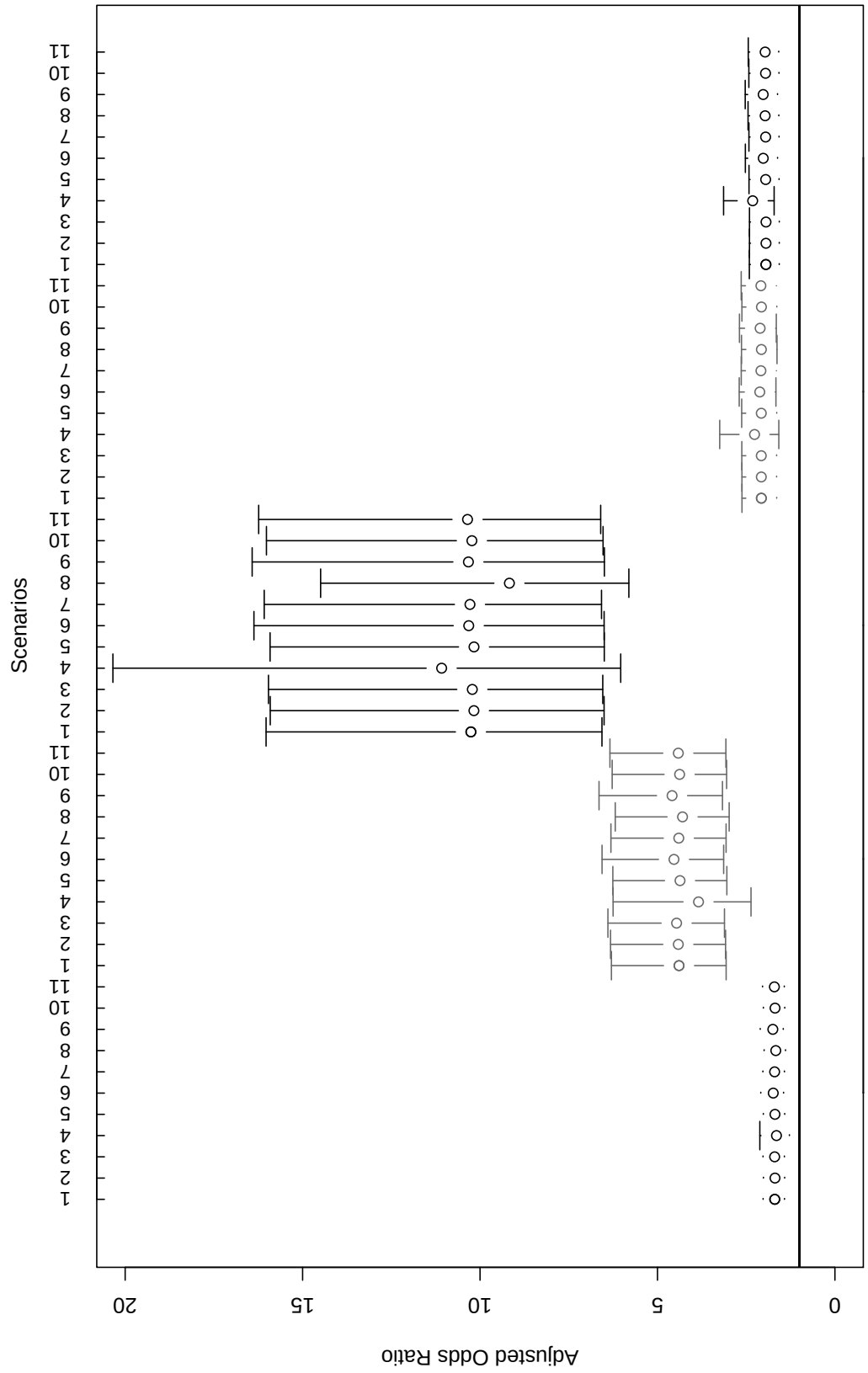

Figure C.5: Coefficient results from the adjusted odds ratio model of interest (Interaction B included) for the 11 scenarios changing Alcohol imputation. Scenarios are represented numerically, corresponding to the rows in Table 4.9; scenario 11 represents MAR. (Coefficients: C21=Birth Weight Grouped 2, C22=Weight Group 3, C23=Weight Group 4, C24=Married or Cohabiting, C25=Mother Age Grouped 2).

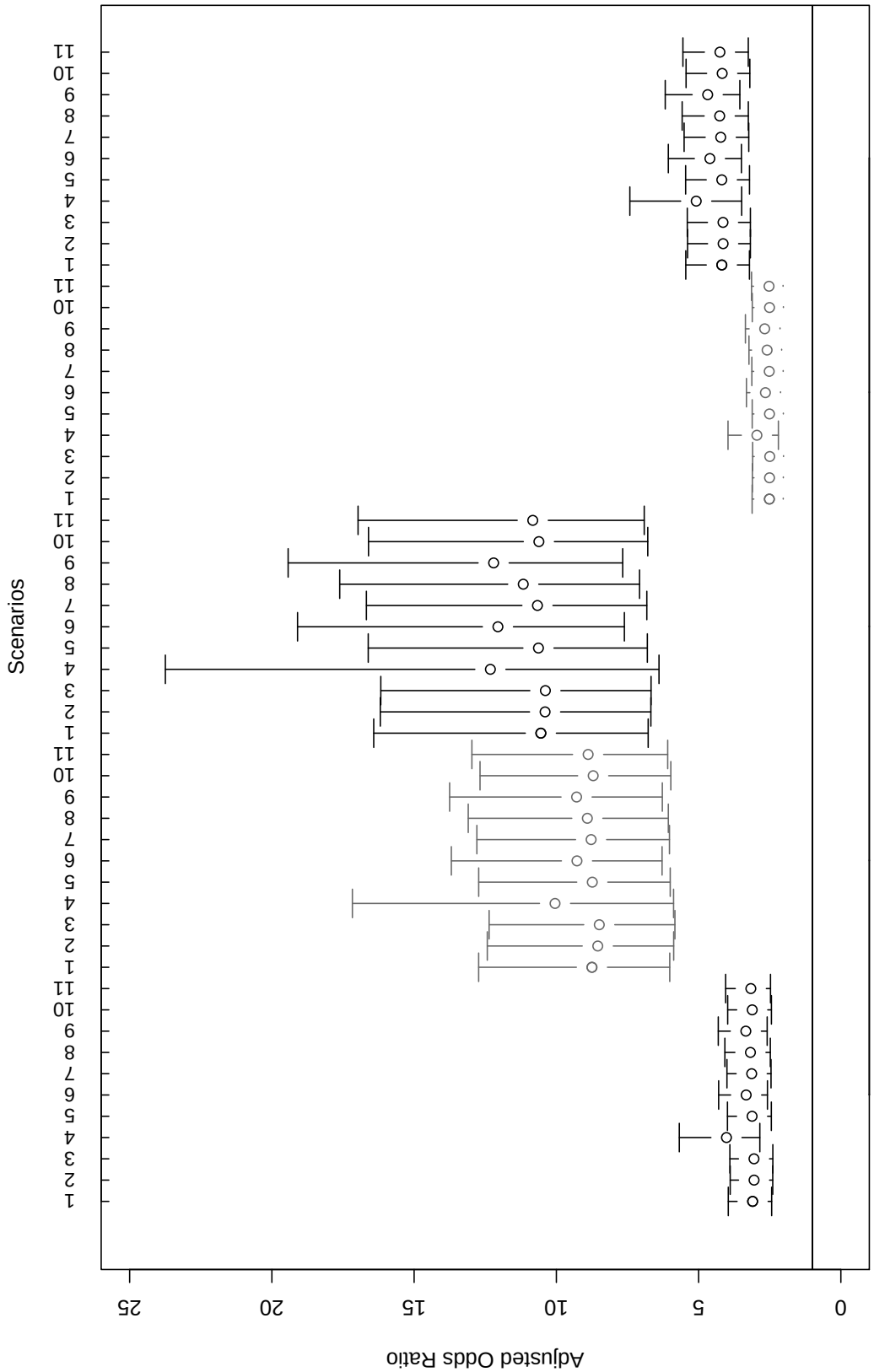

Figure C.6: Coefficient results from the adjusted odds ratio model of interest (Interaction B included) for the 11 scenarios changing Alcohol imputation. Scenarios are represented numerically, corresponding to the rows in Table 4.9; scenario 11 represents MAR. (Coefficients: C26=Mother Age Grouped 3, C27=Mother Age Grouped 4, C28=Mother Age Grouped 5, C29=Number Of Live Births Grouped 2, C30=Number Of Live Births Grouped 3).

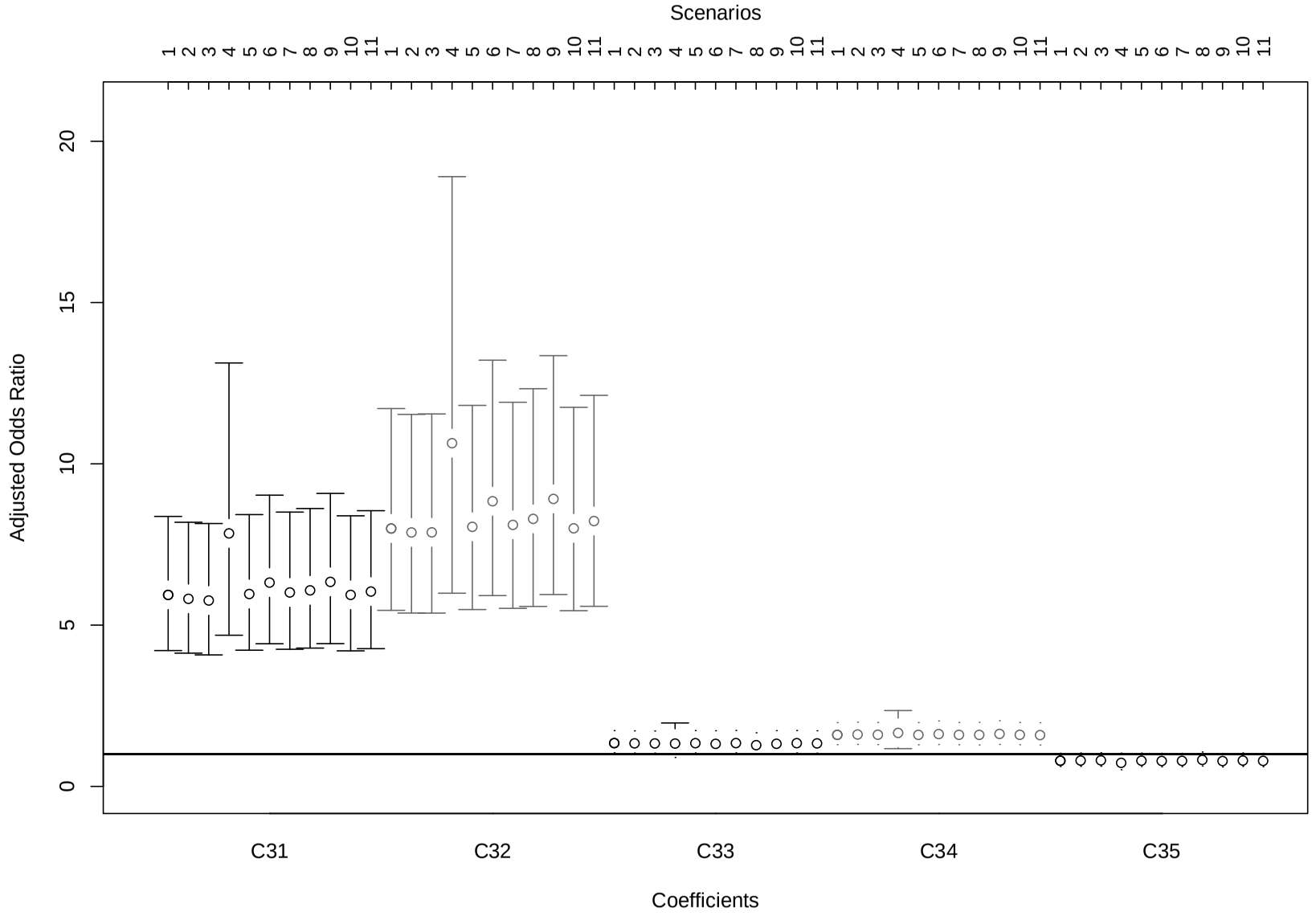

Figure C.7: Coefficient results from the adjusted odds ratio model of interest (Interaction B included) for the 11 scenarios changing Alcohol imputation. Scenarios are represented numerically, corresponding to the rows in Table 4.9; scenario 11 represents MAR. (Coefficients: C31=Number Of Live Births Grouped 4, C32=Number Of Live Births Grouped 5, C33=Race, C34=Matched by Sex 2, C35=Matched by Sex 3).

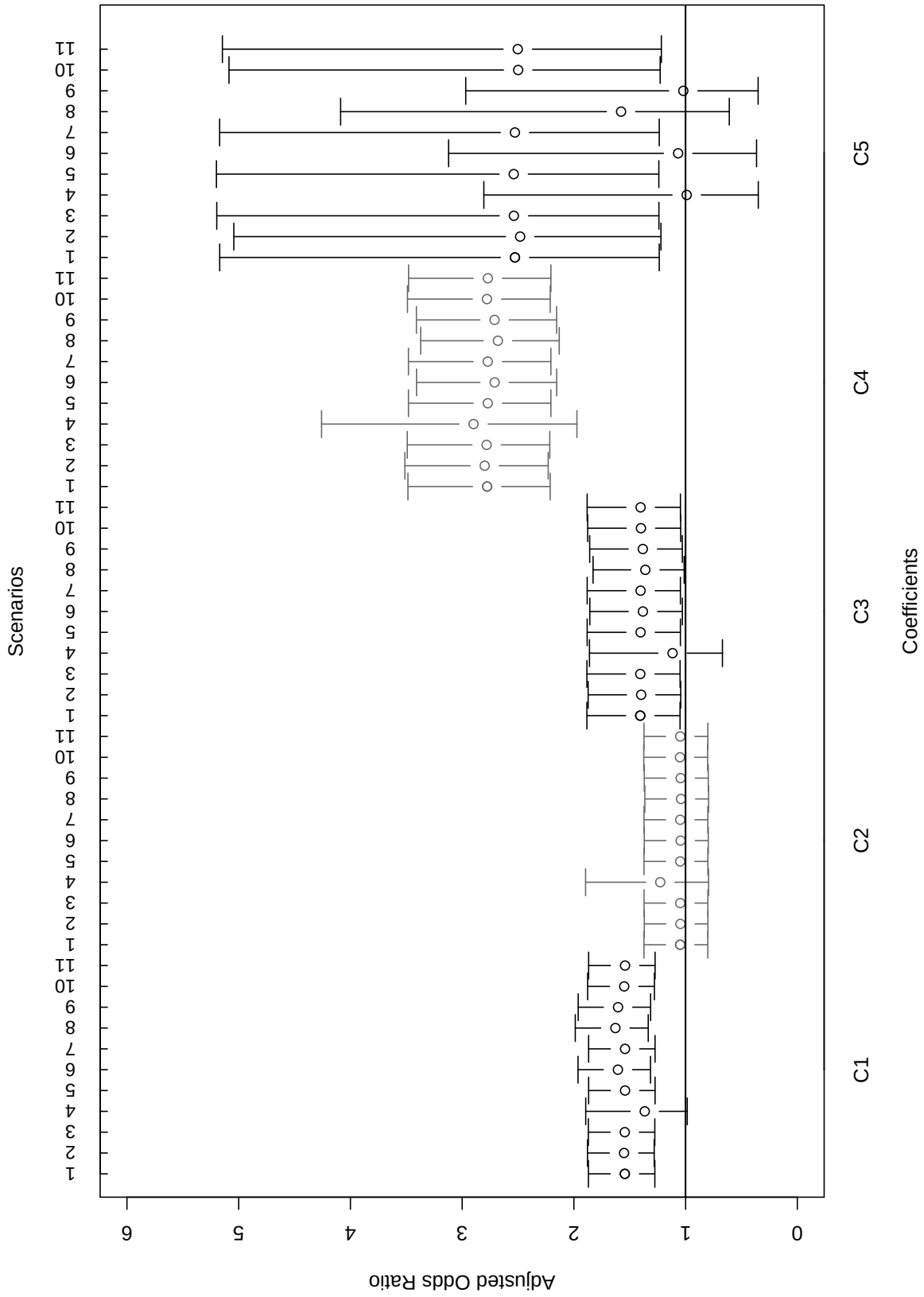

Figure C.8: Coefficient results from the model of interest (4.3.1, with Interaction B removed) for the 11 scenarios changing Drug imputation. Scenarios are represented numerically, corresponding to the rows in Table 4.13. (Coefficients: C1=Bottle Fed, C2=Interaction A.2, C3=Interaction A.3, C4=Interaction A.4, C5=Interaction A.5)

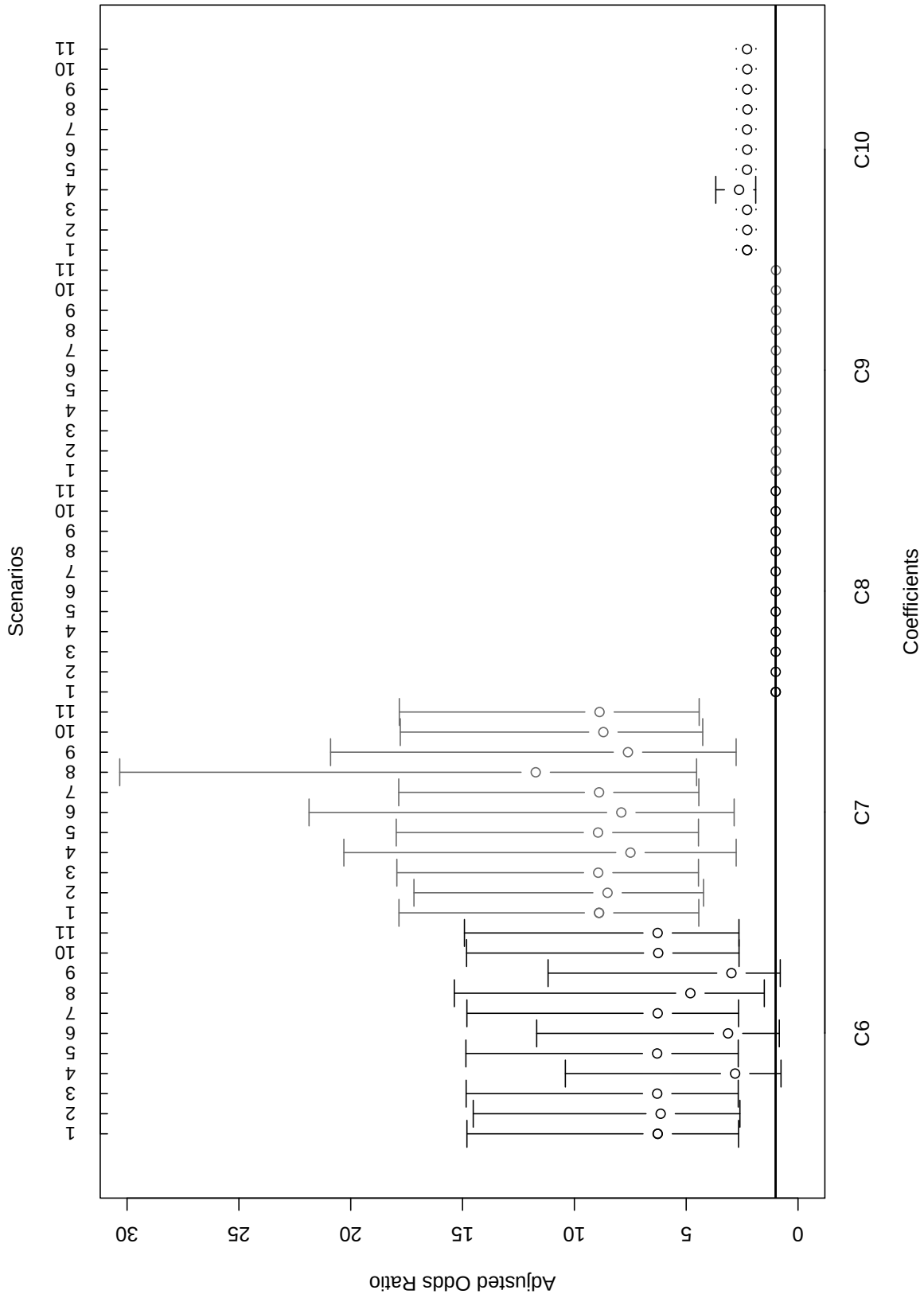

Figure C.9: Coefficient results from the model of interest (4.3.1, with Interaction B removed) for the 11 scenarios changing Drug imputation. Scenarios are represented numerically, corresponding to the rows in Table 4.13. (Coefficients: C1=Interaction A.6, C2=Interaction A.7, C3=Centred Age, C4=Other Room, C5=Interaction C.2)

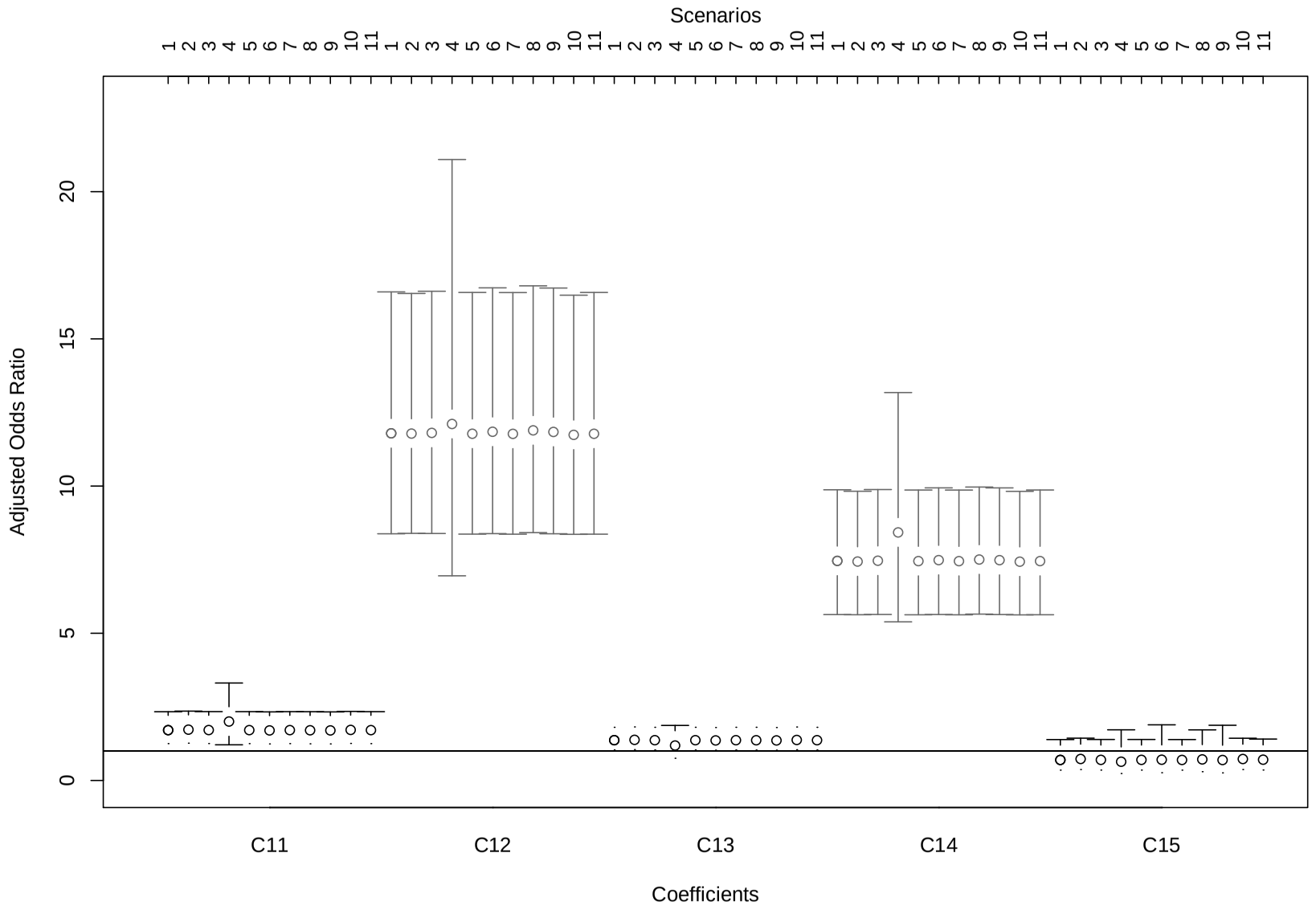

Figure C.10: Coefficient results from the model of interest (4.3.1, with Interaction B removed) for the 11 scenarios changing Drug imputation. Scenarios are represented numerically, corresponding to the rows in Table 4.13. (Coefficients: C1=Interaction C.3, C2=Interaction C.4, C3=Interaction C.5, C4=Interaction C.6, C5=Interaction C.7)

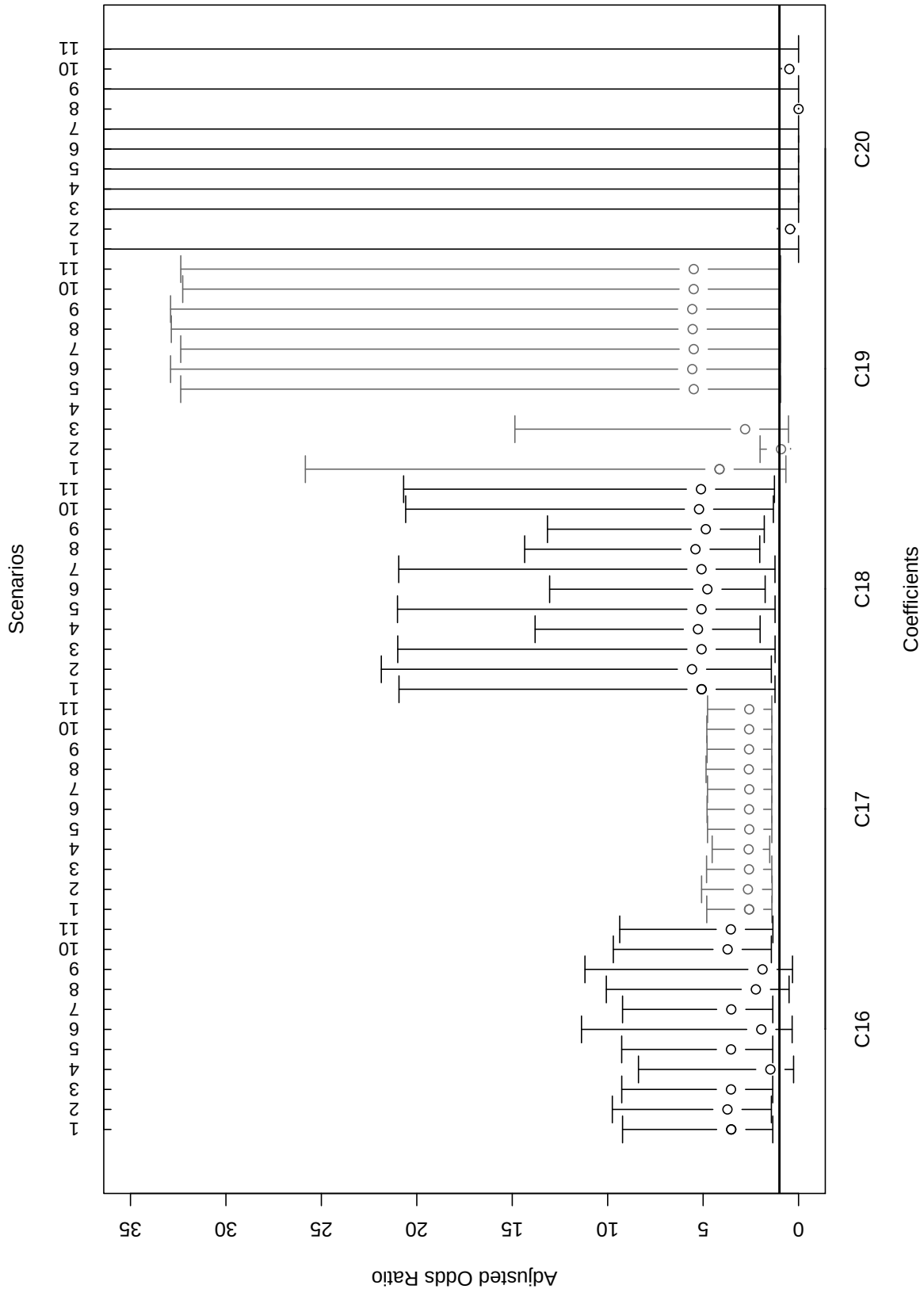

Figure C.11: Coefficient results from the model of interest (4.3.1, with Interaction B removed) for the 11 scenarios changing Drug imputation. Scenarios are represented numerically, corresponding to the rows in Table 4.13. (Coefficients: C1=Interaction D.2, C2=Interaction D.3, C3=Interaction E, C4=Interaction F, C5=Birth Weight Grouped 2)

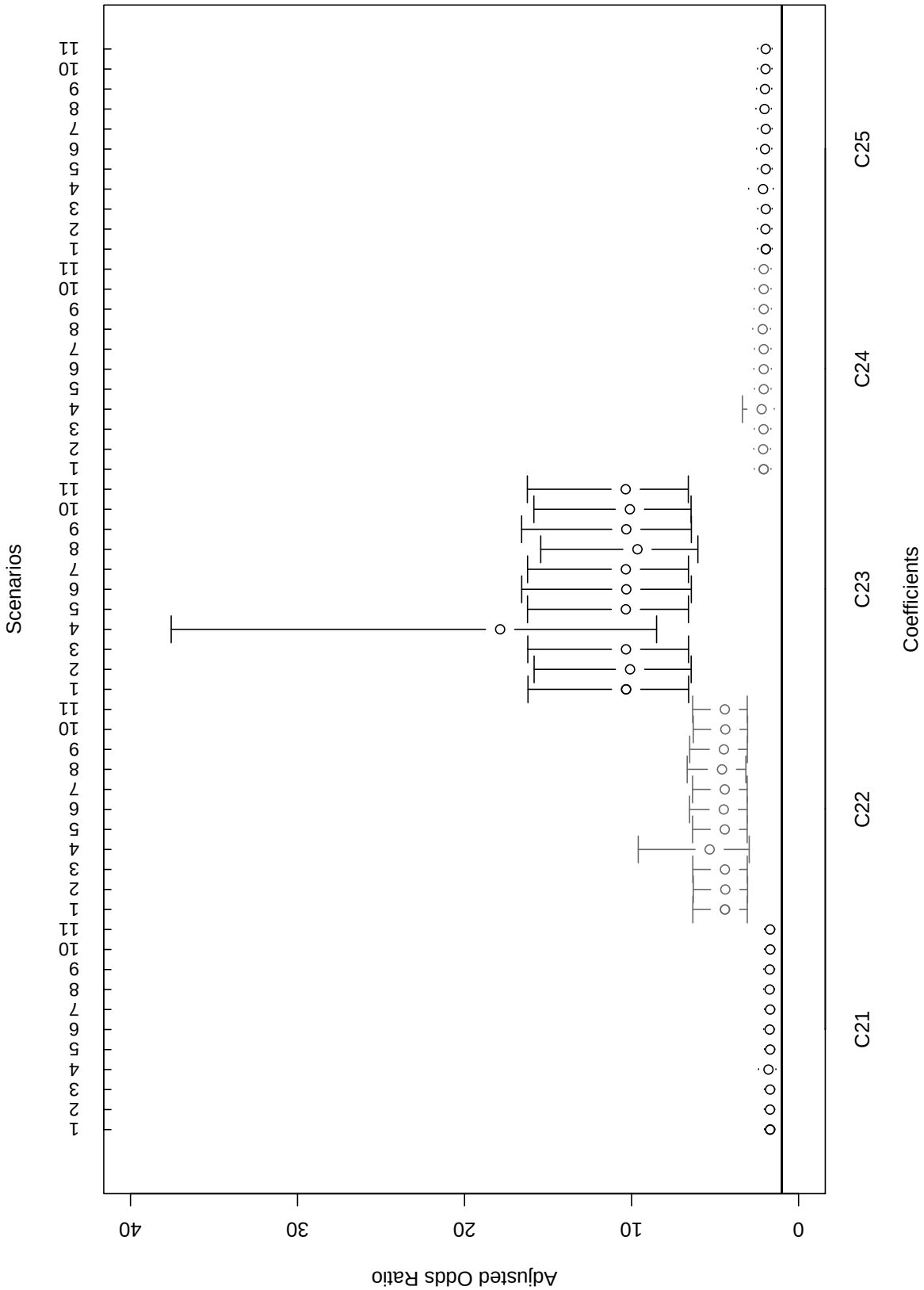

Figure C.12: Coefficient results from the model of interest (4.3.1, with Interaction B removed) for the 11 scenarios changing Drug imputation. Scenarios are represented numerically, corresponding to the rows in Table 4.13. (Coefficients: C1=Weight Group 3, C2=Weight Group 4, C3=Married or Cohabiting, C4=Mother Age Grouped 2, C5=Mother Age Grouped 3)

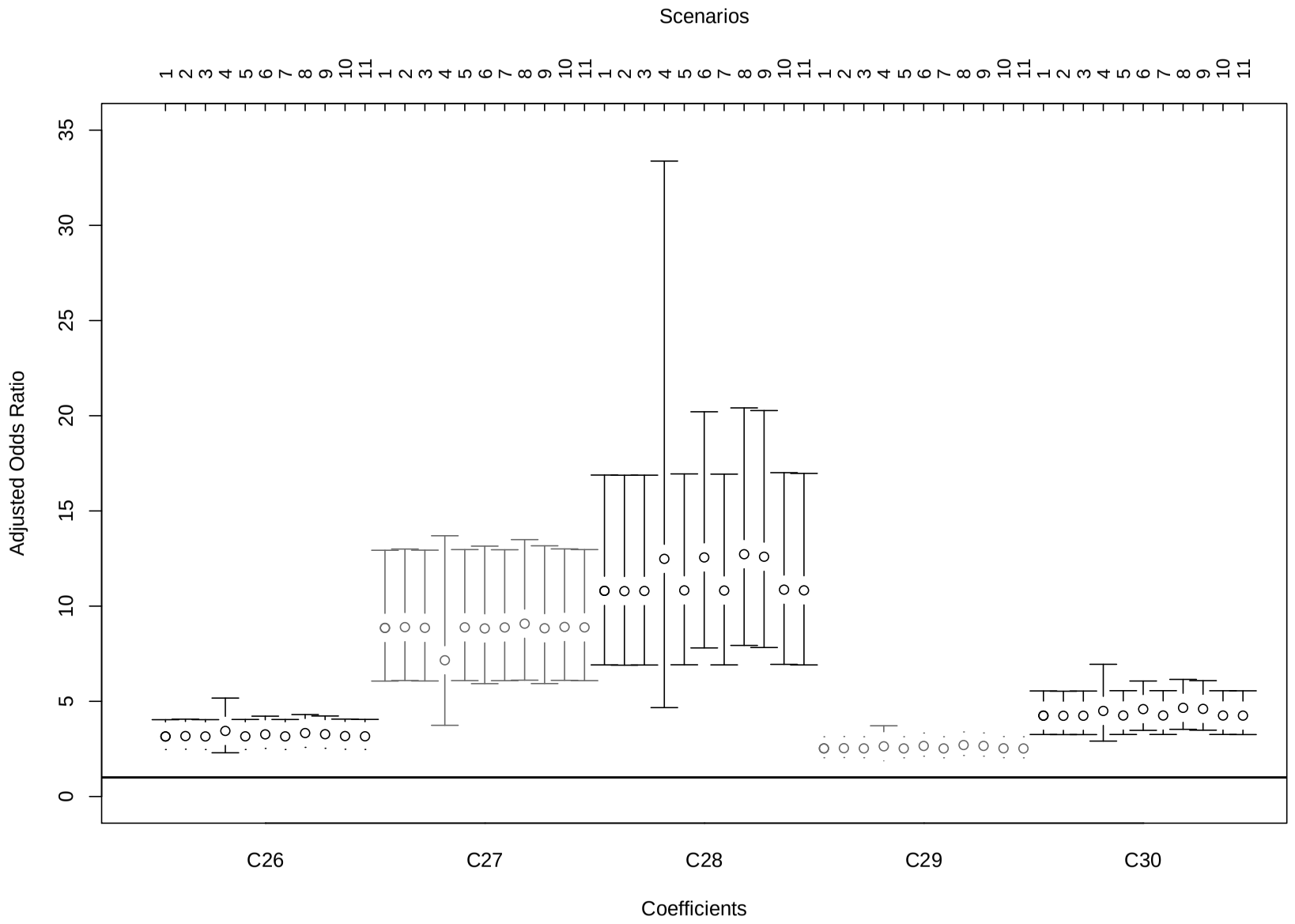

Figure C.13: Coefficient results from the model of interest (4.3.1, with Interaction B removed) for the 11 scenarios changing Drug imputation. Scenarios are represented numerically, corresponding to the rows in Table 4.13. (Coefficients: C1=Mother Age Grouped 4, C2=Mother Age Grouped 5, C3=Number Of Live Births Grouped 2, C4=Number Of Live Births Grouped 3, C5=Number Of Live Births Grouped 4)

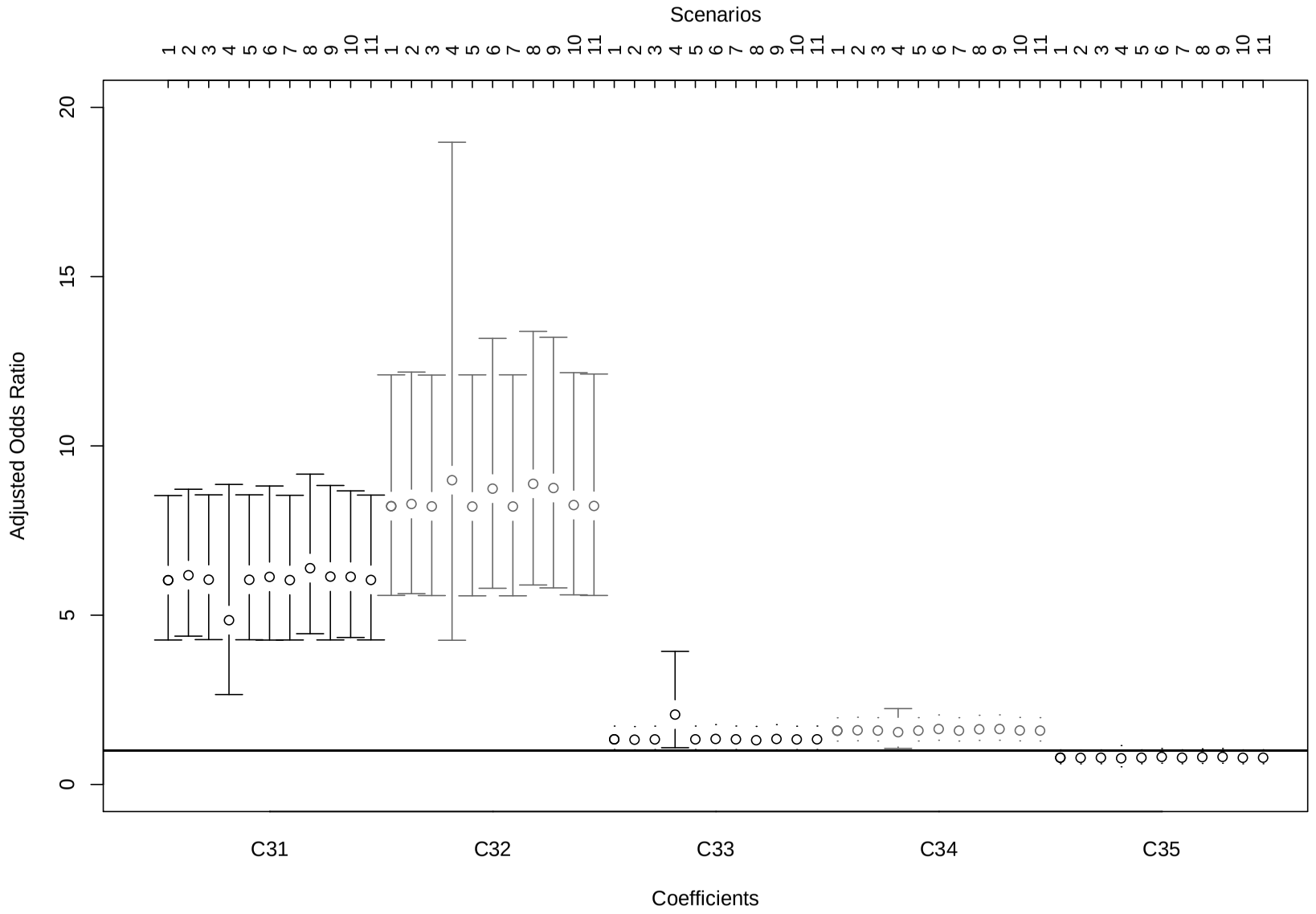

Figure C.14: Coefficient results from the model of interest (4.3.1, with Interaction B removed) for the 11 scenarios changing Drug imputation. Scenarios are represented numerically, corresponding to the rows in Table 4.13. (Coefficients: C1=Number Of Live Births Grouped 5, C2=Race, C3=Matched by Sex 2, C4=Matched by Sex 3, C5=Constant)

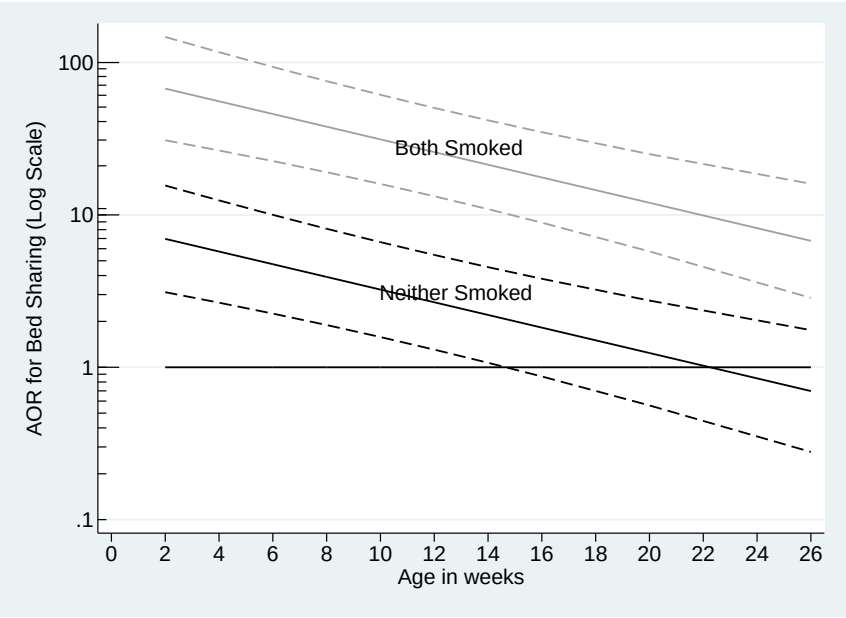

Figure C.15: Adjusted ORs (AORs; log scale) with 95% confidence interval (dotted lines) for Sudden Infant Death Syndrome by age for bed sharing breast-fed infants, when neither parent smokes and both smoke. Imputed with an MNAR bed share, centre indicator prior within the *alcohol* model with a mean of 0.5 and variance of 0

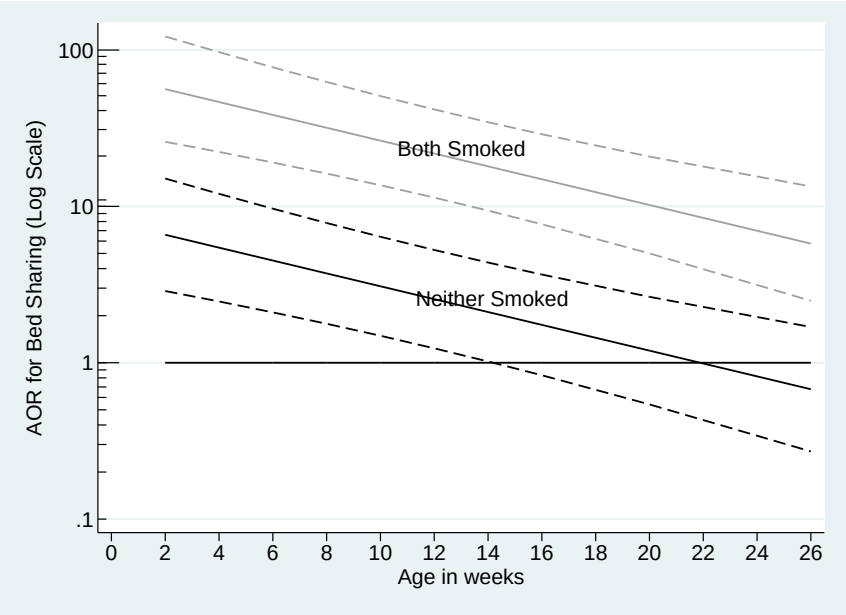

Figure C.16: Adjusted ORs (AORs; log scale) with 95% confidence interval (dotted lines) for Sudden Infant Death Syndrome by age for bed sharing breast-fed infants, when neither parent smokes and both smoke. Imputed with an MNAR bed share, centre indicator prior within the *alcohol* model with a mean of -0.5 and variance of 0

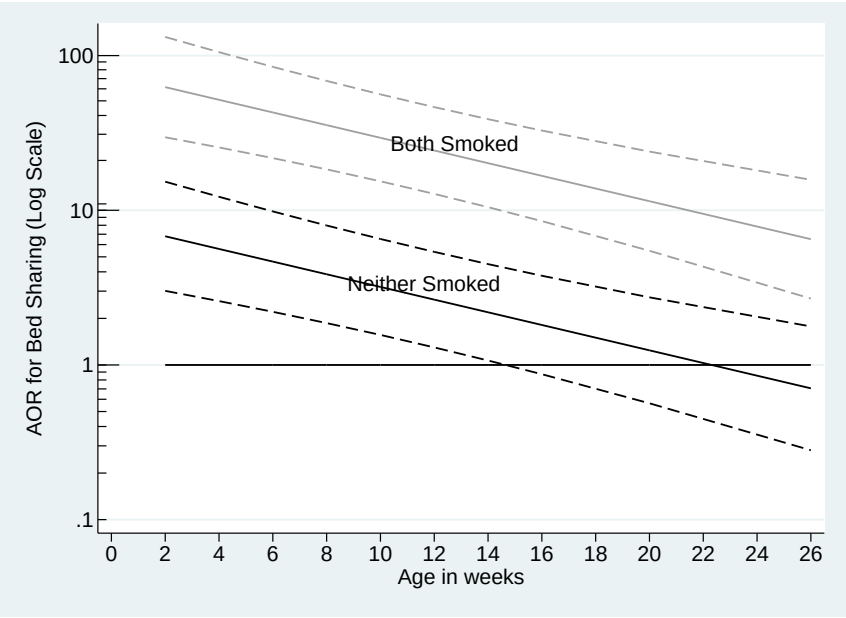

Figure C.17: Adjusted ORs (AORs; log scale) with 95% confidence interval (dotted lines) for Sudden Infant Death Syndrome by age for bed sharing breast-fed infants, when neither parent smokes and both smoke. Imputed with an MNAR indicator prior within the *drug* model with a *Probit prior mean* of 0.5 and variance of 0

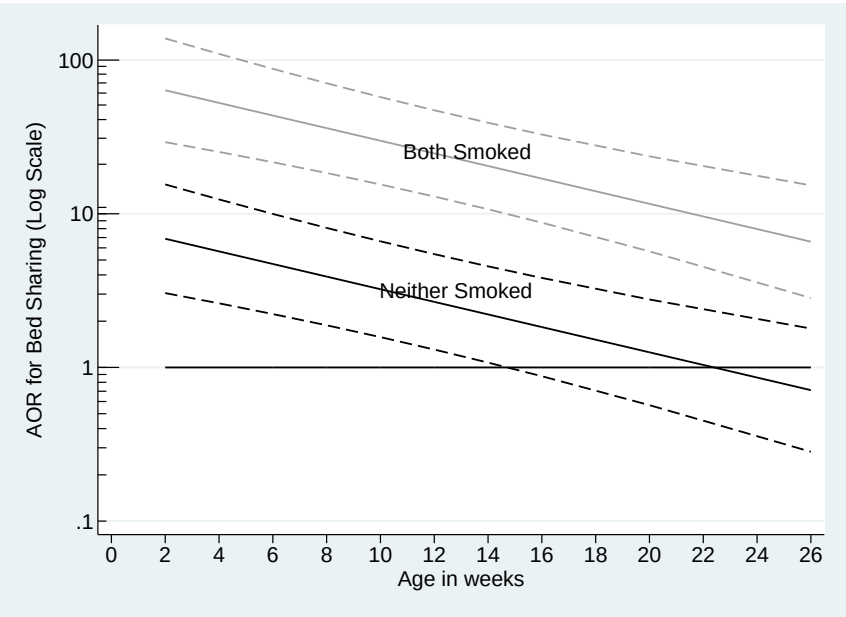

Figure C.18: Adjusted ORs (AORs; log scale) with 95% confidence interval (dotted lines) for Sudden Infant Death Syndrome by age for bed sharing breast-fed infants, when neither parent smokes and both smoke. Imputed with an MNAR indicator prior within the *drug* model with a *Probit prior mean* of -0.5 and variance of 0

D

Chapter 5 Appendix: Sensitivity Analysis via Re-Weighting after Multiple Imputation Assuming Missing At Random: Issues with Small Datasets

D.1 Chapter 5 Appendix: Calculation of $P(Y|x > 0)$ and $P(Y|x \leq 0)$

Under model 5.3.1 we have:

$$P(Y|x > 0) = \frac{P(Y) \times P(x > 0|Y)}{P(x > 0)}$$

Calculating each part of the right hand side (RHS) separately:

$$\begin{aligned} P(x > 0|Y) &= P(x - \mu > -\mu) \\ &= P\left(\frac{x - \mu}{\sigma} > \frac{-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{\mu}{\sigma}\right) \\ &= \Phi\left(\frac{\mu_x + \rho y}{\sqrt{(1 - \rho^2)}}\right) \end{aligned}$$

As σ equals $\sqrt{(1 - \rho^2)}$ and μ equals $\mu_x + \rho y$ from equation (5.3.2).

$$\begin{aligned} P(x > 0) &= P(x - \mu > -\mu) \\ &= P\left(\frac{x - \mu}{\sigma} > \frac{-\mu}{\sigma}\right) \\ &= \Phi(\mu_x) \end{aligned}$$

As σ equals one from equation (5.3.1).

$$\begin{aligned} P(Y) &= \frac{1}{\sqrt{(2\pi\sigma^2)}} e^{\frac{-(y-\mu)^2}{2\sigma^2}} \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} \\ &= \phi(y) \end{aligned}$$

As σ equals one and μ equals 0 from equation (5.3.1). Combining the results we can calculate $Y|x > 0$:

$$P(Y|x > 0) = \frac{\phi(y)}{\Phi(\mu_x)} \Phi\left(\frac{\mu_x + \rho y}{\sqrt{(1-\rho^2)}}\right)$$

Similarly for $Y|x \leq 0$:

$$P(Y|x \leq 0) = \frac{\phi(y)}{1 - \Phi(\mu_x)} \left(1 - \Phi\left(\frac{\mu_x + \rho y}{\sqrt{(1-\rho^2)}}\right)\right)$$

D.2 Chapter 5 Appendix: Calculation of mean and variance for the simulation study

Under model (5.3.4), the probability density of Y given $x > 0$ is

$$P(Y|x>0) = \frac{\phi(y)}{\Phi(0)}\Phi\left(\frac{0+\frac{1}{\sqrt{2}}y}{\sqrt{1-\left(\frac{1}{\sqrt{2}}\right)^2}}\right) = \frac{\phi(y)}{\Phi(0)}\Phi(y)$$

As $Y \sim N(0,1)$ then $\Phi(0) = P(Y \leq 0) = 0.5$ thus:

$$P(Y|x>0) = 2\Phi(Y)\phi(Y).$$

The mean is:

$$E[Y|x>0] = \int_{-\infty}^{\infty} 2\Phi(y)\phi(y)y\, dy.$$

Applying integration by parts:

$$E[Y|x>0] = 2\left\{[-\Phi(y)\phi(y)]_{-\infty}^{\infty} + \int_{-\infty}^{\infty}(\phi(y))^2\, dy\right\}.$$

We next find $\int_{-\infty}^{\infty} (\phi(y))^2 \ dy$:

$$\begin{aligned}\int_{-\infty}^{\infty} (\phi(y))^2 \ dy &= \int_{-\infty}^{\infty} \frac{e^{-y^2}}{2\pi} \ dy \\ &\qquad \text{(appying the substitution } y = \frac{z}{\sqrt{2}}) \\ &= \int_{-\infty}^{\infty} \frac{e^{\frac{-z^2}{2}}}{2\pi\sqrt{2}} \ dz \\ &= \frac{\sqrt{(2\pi)}}{2\pi\sqrt{2}} \\ &= \frac{1}{2\sqrt{\pi}}.\end{aligned}$$

And thus we get:

$$E[Y|x > 0] = 2\left\{[-\Phi(y)\phi(y)]_{-\infty}^{\infty} + \frac{1}{2\sqrt{\pi}}\right\} = \frac{1}{\sqrt{\pi}}.$$

To find the variance:

$$Var[Y|x > 0] = E[y^2] - (E[y])^2$$

We require $E[y^2]$:

$$E[Y^2|x > 0] = \int_{-\infty}^{\infty} 2\Phi(y)\phi(y)y^2 \ dy.$$

Integrate by parts setting:

$$u = \Phi(y)y,$$

$$dv = \phi(y)y,$$

so that

$$v = -\phi(y). \tag{D.2.1}$$

Using the product rule to find du:

$$\begin{aligned}\frac{du}{dy} &= y\frac{d}{dy}\phi(y) + \Phi(y)\frac{d}{dy}y \\ &= \phi(y)y + \Phi(y)\end{aligned}$$

Thus:

$$\begin{aligned}E[Y^2|x>0] &= 2\left\{[-\Phi(y)\phi(y)y]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} -\phi(y)(\phi(y)y + \Phi(y))\ dy\right\} \\ E[Y^2|x>0] &= 2\left\{[-\Phi(y)\phi(y)y]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} (\phi(y))^2 y\ dy + \int_{-\infty}^{\infty} \Phi(y)\phi(y)\ dy\right\} \\ E[Y^2|x>0] &= 2\left\{[-\Phi(y)\phi(y)y]_{-\infty}^{\infty}\right\} + 2\left\{\int_{-\infty}^{\infty} (\phi(y))^2 y\ dy\right\} \\ &\quad + 2\left\{\int_{-\infty}^{\infty} \Phi(y)\phi(y)\ dy\right\}.\end{aligned}$$

Now, $\lim_{y\to\infty} \Phi(y)\phi(y)y = 0$ because the exponential tail of $\phi(y)$ dominates the expression; likewise $\lim_{y\to-\infty} \Phi(y)\phi(y)y = 0$ as $\Phi(y) \to 0$ and $\phi(y) \to 0$ as $y \to -\infty$. Also, $\int_{-\infty}^{\infty} \phi(y)^2 y\ dy = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{-y^2}{2}} y\ dy = 0$ because of the symmetry of $e^{\frac{-y^2}{2}}$ about 0.

Therefore,

$$E[Y^2|x>0] = 2\left\{\int_{-\infty}^{\infty} \Phi(y)\phi(y)\ dy\right\}.$$

To calculate the integral let:

$$
\begin{aligned}
I &= \int_{-\infty}^{\infty} \Phi(y)\phi(y)dy \\
I &= \int_{-\infty}^{\infty} \Phi(y)(\Phi(y))' \, dy = [\Phi(y)\Phi(y)]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \Phi(y)(\Phi(y))' \, dy \\
I &= \left[(\Phi(y))^2\right]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \Phi(y)\phi(y) \, dy \\
I &= 1 - I \\
2I &= 1 \\
I &= \frac{1}{2}.
\end{aligned}
$$

Putting this altogether,

$$
\begin{aligned}
E[Y^2|x > 0] &= 2\left\{\frac{1}{2}\right\} = 1. \\
(E[Y|x > 0])^2 &= \left(\frac{1}{\sqrt{\pi}}\right)^2 = \frac{1}{\pi}.
\end{aligned}
$$

Therefore,

$$
Var[Y|x > 0] = 1 - \frac{1}{\pi}.
$$

D.3 Chapter 5 Appendix: Calculations for α's

Using a generalised linear model (GLM) with a Probit link based on the binomial family the estimated parameters for α_0 and α_1 are created. By comparing $P(\boldsymbol{R} = 1|\boldsymbol{Y}) = \phi(\alpha_0 + \alpha_1 y)$ and (5.3.4), ρ^2 and μ_x can be calculated:

ρ^2 :

$$\alpha_1 = \frac{\rho}{\sqrt{(1-\rho^2)}} \Rightarrow \rho^2 = \frac{\alpha_1^2}{(\alpha_1+1)}$$

μ_x :

$$\alpha_0 = \frac{\mu_x}{\sqrt{(1-\rho^2)}} \Rightarrow \mu_x = \alpha_0 \left(\sqrt{\frac{(1-\alpha_1^2)}{(\alpha_1+1)}} \right).$$