The handling, analysis and reporting of missing data in patient reported outcome measures for randomised controlled trials

Ines Rombach

St Edmund Hall

University of Oxford

Thesis submitted for the degree of
Doctor of Philosophy

Michaelmas Term 2016

Acknowledgements

I would like to thank my supervisors for their guidance, help, and constructive feedback throughout this project:

Associate Professor Oliver Rivero-Arias

Professor Alastair Gray

Professor Crispin Jenkinson

I am very greatful that I have been allowed to use the data from three different randomised controlled trials for this work, and recognise the contribution of all the investigators, collaborators, those who co-ordinated these studies, and the participants of these trials. Particularly, I would like to thank the following groups for their help and support:

- The Knee Arthroplasty Trial (KAT) team, particularly Professor David Murray, Barbara Marks, Dr Suzanne Breeman, Dr Helen Dakin and Professor Graeme MacLennan
- The PD MED and PD SURG teams, particularly Professor Richard Gray, Caroline Rick and Francis Dowling

In addition, I also thank the following people

Jill Dawson
Ly-Mee Yu
Merryn Voysey
Órlaith Burke
Claire Simons
Filipa Landeiro
Seamus Kent
Iryna Schlackow
Jacqui Murphy
Larry Rope
Vandana Ayyar Gupta
Ellen Nuttall Musson
Steph Dakin
Naomi Merritt
John Broomfield
Antony Palmer
and - of course - Lottie Davies, for being the best PhD & bagel buddy imaginable

Publications, presentations and additional funding arising from this thesis

Publications

Rombach I, Rivero-Arias O, Gray AM, Jenkinson C, Burke O. The current practice of handling and reporting missing outcome data in eight widely used PROMs in RCT publications: a review of the current literature. *Quality of life research: an international journal of quality of life aspects of treatment, care and rehabilitation* 2016; **25**(7): 1613-23.

Presentations

Rombach I, Gray AM, Rivero-Arias O, Jenkinson C. Current practice in preventing and handling missing data alongside clinical trials: are we doing well? 2nd Oxford Research Network Conference; October 2014; Oxford, UK; October 2014.

Rombach I, Rivero-Arias O, Gray AM, Jenkinson C, Burke O. The current practice of handling and reporting missing outcome data in 8 widely-used PROMs in RCT publications: A review of the current literature. Society of Clinical Trials - 37th Annual Meeting; May 2016; Montreal, Canada; May 2016.

Rombach I, Gray AM, Jenkinson C, Burke O, Rivero-Arias O. Applying multiple imputation to multi-item Patient reported Outcome Measures: Advantages and disadvantages of imputing at the item, sub-scale or score level. Society of Clinical Trials - 37th Annual Meeting; May 2016; Montreal, Canada; May 2016.

Rombach I, Burke O, Jenkinson C, Gray AM, Rivero-Arias O. Applying multiple imputation to multi-item Patient reported Outcome Measures: Advantages and disadvantages of imputing at the item, sub-scale or score level. PROMs Conference; June 2016; Sheffield, UK; June 2016.

Rombach I, Burke O, Jenkinson C, Gray AM, Rivero-Arias O. Applying multiple imputation to multi-item Patient reported Outcome Measures: Advantages and disadvantages of imputing at the item, sub-scale or score level. International Society of Clinical Biostatistics - 37th Annual Conference; August 2016; Birmingham, UK; August 2016.

Additional funding

I was awarded funding from the Medical Research Council Doctoral Training Partnership to perform additional methodological and applied to further inform and validate the findings presented in thesis, as well as for the publication of academic papers arising from the DPhil project.

Abstract

The handling, analysis and reporting of missing data in patient reported outcome measures for randomised controlled trials

Ines Rombach, St Edmund Hall
D.Phil.
Michaelmas Term 2016

Missing data is a potential source of bias in the results of randomised controlled trials (RCTs), which can have a negative impact on guidance derived from them, and ultimately patient care.

This thesis aims to improve the understanding, handling, analysis and reporting of missing data in patient reported outcome measures (PROMs) for RCTs.

A review of the literature provided evidence of discrepancies between recommended methodology and current practice in the handling and reporting of missing data. Particularly, missed opportunities to minimise missing data, the use of inappropriate analytical methods and lack of sensitivity analyses were noted.

Missing data patterns were examined and found to vary between PROMs as well as across RCTs. Separate analyses illustrated difficulties in predicting missing data, resulting in uncertainty about assumed underlying missing data mechanisms.

Simulation work was used to assess the comparative performance of statistical approaches for handling missing available in standard statistical software. Multiple imputation (MI) at either the item, subscale or composite score level was considered for missing PROMs data at a single follow-up time point. The choice of an MI approach depended on a multitude of factors, with MI at the item level being more beneficial than its alternatives for high proportions of item missingness. The approaches performed similarly for high proportions of unit-nonresponse; however, convergence issues were observed for MI at the item level. Maximum likelihood (ML), MI and inverse probability weighting (IPW) were evaluated for handling missing longitudinal PROMs data. MI was less biased than ML when additional post-randomisation data were available, while IPW introduced more bias compared to both ML and MI.

A case study was used to explore approaches to sensitivity analyses to assess the impact of missing data. It was found that trial results could be susceptible to varying assumptions about missing data, and the importance of interpreting the results in this context was reiterated.

This thesis provides researchers with guidance for the handling and reporting of missing PROMs data in order to decrease bias arising from missing data in RCTs.

Table of contents

Approximate number of words in this thesis: 50,000

List of tables

List of figures

List of abbreviations

δ	Delta, here the difference between observed and unobserved data
ACA	Available Case Analysis
ASA physical status	American Society of Anaesthesiologists physical status
BMI	Body Mass Index
CC	Complete cases
CCA	Complete case analysis
CI	Confidence interval
CONSORT	Consolidated standards of reporting of trials
EMA	European Medicines Agency
EQ-5D-3L	EuroQol 5 dimension 3-level questionnaire
EQ-VAS	EuroQol visual analogue scale
FDA	U. S. Food and Drug Administration
GEE	Generalised estimating equations
HRQoL	Health related quality of life
HUI	Health utility index
IPW	Inverse probability weighting
ISPOR	International society for pharmacoeconomics and outcomes research
ITT	Intention to treat
KAT	Knee arthroplasty trial
LOCF	Last observation carried forward
MAE	Mean absolute error
MAR	Missing at random
MCAR	Missing completely at random
MCS	Mental Health Component Summary score
MI	Multiple imputation
MICE	Multiple imputation with chained equations
ML	Maximum likelihood

MLE	Maximum likelihood estimate
MNAR	Missing not at random
NHS EED	National Health Service economic evaluation database
OHS	Oxford hip score
OKS	Oxford knee score
OR	Odds ratio
PCI	Pain coping inventory
PD	Parkinson's disease
PD MED	A large, randomised assessment of the relative cost-effectiveness of different classes of drugs for Parkinson's disease
PD SURG	A large, randomised, long-term assessment of the relative effectiveness of surgery for Parkinson's disease
PDQ	Parkinson's disease questionnaire
PDQ-39-SI	39 item Parkinson's disease - single index score
PCS	Physical Health Component Summary score
PRO	Patient reported outcome
PROM	Patient reported outcome measure
QLQ-C30	European organisation for research and treatment of cancer quality of life questionnaire-core 30
QoL	Quality of life
RCT	Randomised controlled trial
RMSE	Root mean square error
REML	Restricted Maximum Likelihood
SE	Standard error
SF-12	12-item short form survey
SF-36	36-item short form survey

Chapter 1 : Introduction

The overall aim of this thesis is to contribute to improving of the handling, analysis and reporting of missing information in patient reported outcome measures (PROMs) for randomised controlled trials (RCTs). This is addressed, first, by critically appraising current practice in the handling, analysis and reporting of missing PROMs outcome data within RCTs compared to best recommended practice. Subsequently, the comparative performance of different analysis approaches for handling missing data that are validated and routinely available in statistical software are assessed in cross-sectional and longitudinal settings focussing specifically on PROMs data in RCT analyses. Finally, sensitivity analysis exploring the robustness of RCT results due to missing data is discussed. Prior to providing details of the aims, objective and structure of this thesis, a brief introduction to clinical trials, patient reported outcome measures and missing data is provided, and the scope of the thesis is clarified.

1.1 Background

1.1.1 Of clinical trials, patient reported outcome measures, and missing data

With many competing treatment options available today, and with finite resources available to spend on health care, it is more important than ever to carefully evaluate the comparative performance of various interventions in order to provide patients with the best possible health care. RCTs are considered the "best way to compare the effectiveness of different interventions" and thus can directly affect patient care[1].

An RCT is an experiment during which the study population is randomly allocated to receiving one of two or more clearly defined interventions that are to be compared to each other. Randomising participants to their treatments, ideally with patients and those treating them blinded to the allocation of each patient, removes selection bias[2], i.e. the practice whereby the intervention is decided upon by either the patient or the treating clinician. Such non-randomised allocation is likely to lead to patients who systematically differ in terms of factors including socio-demographic characteristics, disease history and severity, as well as prognostic factors, being allocated to different intervention groups. Instead, randomisation creates balanced groups for a fair comparison, and is therefore a crucial contribution to the fact that RCTs, where ethically permissible[3], have long been accepted as an essential component in the reliable assessment of new treatments and interventions, i.e. in evaluating the presence of a causal effect between interventions and outcomes[4].

Traditionally, 'objective' measures such as death, recurrence of cancer, or blood pressure were the primary outcome measures used in many clinical trials. However, over the last 20 years, clinicians, researchers and policy makers have increasingly become aware of the importance of taking into account the patient perspective to inform patient care and policy

decisions[5-7]. As a consequence, a number of instruments have been developed to collect information on patients' perceived health states or their perceived health-related quality of life (HRQOL). Often referred to as patient reported outcomes (PROs) or PROMs, these measures, usually in questionnaire form, include 'any report coming directly from patients, without interpretation by physicians or others, about how they [the patients] function or feel in relation to a health condition and its therapy'[8]. PROMs may be either generic, i.e. assessing patients' overall HRQOL without referring to a specific illness or health condition, or specific, i.e. focussing on a specific patient group, a certain disease or particular areas of function[9, 10]. Before being utilised in clinical research, PROMs should undergo rigorous assessment to test their validity, reliability and robustness, to ensure that even small relevant changes in HRQOL can be measured, and to verify that differences in measurements are not due to error or noise[9, 11-13]. Due to these characteristics, PROMs are an important addition to traditional measures, which may not fully capture the patient experience of a specific treatment or disease burden[14]. Therefore, PROMs are increasingly utilised in randomised controlled trials (RCTs), where their use as primary or secondary outcome measures is considered of great importance[5, 6].

In reality, some outcome data is likely to be missing in all RCTs[15], but the use of PROMs data may exaggerate this problem, as this data is often thought to be more susceptible to being missing than more objective, and possibly routinely collected, outcome data. One reason is that RCTs utilising PROMs rely on their participants being able and willing to complete the relevant outcome measures throughout the period of their follow-up, and it is therefore often impossible to obtain complete follow-up data for all randomised participants, for example if their health is too poor to complete the measures[15, 16]. It is just

this feature of RCTs utilising PROMs that can call into question their ability to provide reliable estimates for the clinical and cost-effectiveness of interventions.

Missing data, i.e. data that was intended to be collected and considered relevant for the statistical analysis and interpretation of the results, but is unavailable for the analysis[17], affects RCT results in two ways: It increases the uncertainty around the trial results, and can lead to bias being introduced into the trial results[18]. Whether or not RCT results are biased due to the occurrence of missing data, and how much bias is introduced as a result depends on a multitude of factors, including the extent of missing data within the study, the assumptions made about the underlying missing data mechanism and the handling of the missing data in the analysis, which are discussed further in section 1.1.2.

1.1.2 On the implications of missing data

An integral part of the planning of any RCT is a sample size calculation, which aims to ensure that a sufficient number of participants are recruited to the study to answer the primary objective with the "required statistical precision and certainty"[19]. The sample size calculation is also supposed to guard against too many participants being recruited to a trial, who may be exposed unnecessarily to an inferior or even harmful treatment. Trials with both too many or too few participants compared to an appropriately calculated sample size are described by Cook et al as "[arguably] unethical, wasteful, and potentially misleading" [19].

Therefore, if participants with missing data are ignored in the analysis, and less observations than required are available for the analysis, this can result in the trial having reduced power[20], i.e. a decreased ability to detect a true treatment difference between the trial arms should such a different exist. Again, this can have real-life impact on patient

care, for example if a potentially beneficial treatment is not recognised as such in an RCT due to reduced power because of missing data. In these circumstances, beneficial treatments may not be prescribed to patients due to missing data.

Another much more disconcerting potential consequence of missing data is the possible introduction of bias, as referred to above. Bias can be introduced in the estimated trial results with respect to between as well as within group effects, i.e. in the estimation of treatment effects and effects over time, respectively[20]. This potential for bias is related to the underlying missing data mechanism, i.e. the relationship between the probability of a specific observation being missing and the values of the data, observed and unobserved[21]. More details on the definition of the missing data mechanism are provided in Chapter 2.

Most methods of handling missing data implicitly assume that it follows a specific mechanism. As an example, consider a complete cases analysis (CCA), whereby participants with incomplete observations are excluded from the analysis, which is still the most commonly used analysis method for RCTs[22-25]. For an unadjusted analysis, this approach requires the data to be missing completely at random (MCAR), i.e. assumes that the participants with incomplete outcome data are exactly representative of the participant with complete follow-up data[25], in terms of observed and unobserved variables. These assumptions are impossible to verify with the data available[15], but unbiased results are obtained only if the underlying assumptions are correct; otherwise unquantifiable bias may be introduced[15, 21, 26]. This can lead to inaccurate results which may have real-life, and potentially dangerous impacts on regulatory frameworks, guidelines and ultimately patient care and welfare. For example, in a hypothetical trial comparing two different drugs, participants in one trial arm may be more likely to be lost to follow-up, and hence missing

data may be more prevalent in this trial arm. If the reason for the loss to follow-up is that participants in that arm have worse outcomes than those with complete follow-up data, then the data are missing not at random (MNAR). Any CCA that does not take this into account appropriately leads to trial results that wrongly overestimate the benefit of this intervention. This, in turn, may lead to patients being prescribed this treatment instead of one that offers more health benefits. Similarly, if participants have missing data due to dropping out of a trial because of adverse events, this could lead to the trial wrongly underestimating the toxicity of a treatment, which again could results in treatments with unacceptably high toxicity being prescribed to patients, based on trial results biased due to missing data. For these reasons, further research into the handling and reporting of missing data in RCTs is warranted to ensure that the bias introduced into the results of clinical trials due to missing data is minimised. The reduction of bias in RCT results is crucial in ensuring that health care policies based on RCTs maximise patient welfare.

1.2 Scope of this thesis

1.2.1 Methodological approaches of handling missing data used in this project

Over time, a large number of methods have been developed to address missing data, with varying degrees of sophistication. It is not within the scope of this introduction to compare and contrast all these methods, but additional detail is provided in Chapter 2. General consensus amongst methodologists is that multiple imputation (MI) is one of the most appropriate methods[17, 20, 27, 28]. In simple terms (more detail on MI is provided in later chapters), MI replaces missing observations with plausible values based on the observed values in the dataset. By repeating the process a pre-specified number of times, as opposed to many simple or single imputation methods, and pooling the results using the relevant rules, this method takes into account the uncertainty around these estimates and thus produces appropriate standard errors and unbiased results, provided that the assumptions about the underlying missing data mechanism are appropriate [20, 29]. For this reason, the majority of the work within this thesis is based on applications of MI.

Other methodologically robust approaches to handling missing data exist, and continue to be further extended and developed on an ongoing basis[30-37]. Such development is crucial in continuing to improve the reporting and analysis of RCTs when confronted with some missing outcome data. However, often it takes considerable time for these methods to move from their theoretical premise to being sufficiently tested and validated, widely accepted by the statistical community and readily available for implementation via standard statistical software. Even once more widely available, there is often a time lag before researchers shift from using established or commonly used methodology to more recently developed analysis methods[22, 25, 38, 39]. Similarly, the requirements to pre-specify analyses methods to be used during the grant application and design stages of RCTs and

other clinical studies, when the primary statistical analysis may not take place for some years, contribute to the fact that there is a delay in the development and implementation of new methodology generally in the analysis of RCTs.

Further research into the application of existing, well established methods in order to understand how and when these should be applied to minimise the bias introduced into study results is also vital. This is why the simulation work within this thesis focusses on the most appropriate application of readily available MI methodology to realistic missing data problems, aiming to provide guidance to researchers to obtain the most robust results given the missing data scenarios they are most frequently confronted with. The advantage of MI is that it is intuitive and easily understood by both statisticians and other researchers involved in the conduct of RCTs. Also, MI is widely available in standard statistical software, such as Stata, SAS, R and others. However, there are still outstanding questions about the level of application within multi-item PROMs, and its application to longitudinal PROMs data, compared to alternative, equally established analysis approaches.

1.2.2 Focus of this work

In some areas of clinical research, and specifically in epidemiological studies where researchers may depend on data from a variety of sources, and may aim to link outcomes to potential risk factors and exposures, data may be missing for the study outcome, as well as exposure variables and patient characteristic, i.e. the explanatory variables. However, this work focusses on RCT data, the design features of which mean that many fewer explanatory variables need to be collected to investigate the causal relationship between an intervention and the study outcome. Only selected explanatory variables, which are considered to be informative of the outcome, are usually collected prior to randomisation,

and included in the analysis models. This data collection usually occurs in a controlled setting, thus making it more likely for this data to be complete. It is for this reason that this thesis focusses on missing data in PROMs collected as study outcomes in RCTs, and assumes that all other covariate data are complete.

1.2.3 A cautious note on missing data

It has been emphasised previously that all methods for handling missing data rely on numerous assumptions. Therefore, analysis based on missing data can never replace the information or produce the robust and unbiased results that would have been available from an RCT with complete follow-up[17, 40].

This project aims in no way to undermine the seriousness of the implications that can stem from the occurrence of missing data within an RCT or other research, which can call into question "the scientific credibility of causal conclusions from [any such studies]"[17].

The simulation studies presented aim to identify analysis approaches that minimise bias, but acknowledge that such bias cannot be eliminated entirely. Therefore, appropriate steps should be taken and acknowledged in the appropriate documents to facilitate the collection of a data set as complete as possible throughout the follow-up of a trial to minimise the occurrence of missing data.[18, 40]. Only where all possible steps to minimise the amount of missing data have been taken, should the analytical approaches of handling missing data, as described in this thesis, be employed.

For those RCTs that have made appropriate attempts to minimise missing data, but are still facing some unobserved outcomes, this thesis aims to provide guidance on the choice of analysis approaches for specific missing data scenarios. Any assumptions underlying these approaches are clearly stated in the relevant chapters. Any subsequent interpretation of

those results should be made in the light of the assumptions used in the analysis, and should appropriately acknowledge any limitations. In addition, each primary analysis of data including missing outcomes should always be accompanied by a range of sensitivity analyses to assess the robustness of the results when varying the assumptions made about the underlying missing data mechanism. Examples of appropriate sensitivity analysis are therefore also discussed within this thesis.

1.3 Overview of the Thesis

1.3.1 Objectives

The main objective for this thesis is to develop a greater understanding of how missing data in PROMs should be handled and reported when presenting the results of RCTs in order to increase the value and information contained in these reports, as well as their robustness.

This work aims to achieve this goal by considering a number of aspects related to missing PROMs data in RCTs:

1. To create a clear picture of how missing data in PROMs is handled and reported in the current literature, and to compare this to recommended best practice.
2. To examine in detail missing data patterns within selected PROMs in three RCTs.
3. To investigate the benefits and disadvantages of multiply imputing missing PROMs outcome data at the composite score, item or subscale level, where these are validated, and to draw up a set of recommendations.
4. To investigate the advantages and disadvantages of applying MI to PROMs data in a longitudinal setting, compared with alternative recommended methods, namely maximum likelihood and inverse probability weighting.
5. To explore the recommended components of appropriate sensitivity analysis that should be performed in the context of missing PROMs data within an RCT as a way of assessing how robust the primary trial results are to varying the underlying assumptions made about the missing data mechanisms.

The outlined work creates a greater understanding of how missing data within PROMs should be handled, analysed and reported within the analysis of RCTs, and helps to

generate important guidance with regards to the handling, analysis and reporting of missing PROMs data in RCTs.

1.3.2 Chapter outline

Following on from this introduction, Chapter 2 reviews and presents current proposed best practice with regards to handling and reporting missing data in RCTs and discusses the relevant methodology and underlying assumptions. Subsequently, the chapter reviews reports from clinical trials utilising one of eight pre-specified PROMs to investigate how missing data was reported and handled in these articles. This systematic assessment of the literature establishes current practice with regards to missing data in RCT reports, which is compared to proposed best practice. The reported rates of missing data are also extracted and summarised, and feed into subsequent simulation work.

Chapter 3 considers three RCT datasets to establish common rates and patterns of missing data within widely used PROMs. The chapter focusses on the Oxford Knee Score (OKS)[13, 41, 42], the 12-item short form survey (SF-12)[43, 44], EuroQol 5 Dimension 3-Level Questionnaire (EQ-5D-3L)[45, 46], and the 39-item Parkinson's Disease Questionnaire (PDQ-39)[47, 48]. In addition, potential predictors of missing data are considered using appropriate statistical modelling. This empirical study is important in its own right, as it contributes to the understanding of missing data within these PROMs, but also provides estimates to be used in the subsequent simulation work.

Chapter 4 compares and contrasts different approaches to implementing MI when considering a single follow-up time point. Specifically, this chapter addresses the question of whether MI should be applied to the composite scores of the questionnaires, to the

individual items or to the subscales where applicable. Simulation models are used to compare the performance of the different approaches for a number of different scenarios, such as varying amounts of missing data, different missing data patterns, and different sample sizes.

Chapter 5 contributes to the discussion on the most appropriate analysis method for missing PROMs outcome data for longitudinal follow-up. The theoretical advantages and drawbacks of different approaches to handling missing outcome data in a longitudinal setting are considered, namely maximum likelihood models, multiple imputation and inverse probability weighting. The different approaches are compared within a simulation study, again considering a number of different scenarios, including variations in proportions of missing data and sample sizes, as well as the impact of different auxiliary variables being available.

Chapter 6 explores sensitivity analyses that should be presented in conjunction with any analysis incorporating missing data. Sensitivity analyses are a crucial contribution in the interpretation of any analysis involving missing data which relies on assumptions that cannot be tested using the available data. Sensitivity analyses can be employed in these circumstances to consider how robust the trial results are if the assumptions made about the underlying missing data mechanism are varied. This chapter reviews the literature in order to develop proposed components of a comprehensive sensitivity analysis in the context of missing outcome data observed within an RCT, whilst allowing for the possibility that data is missing not at random, i.e. the probability of data being missing is related to the missing data values themselves. A case study is utilised to illustrate two approaches to sensitivity analysis when PROMS outcome data may be missing not at random.

Interpretation of these sensitivity analyses, as well as their implication on the study conclusions are discussed.

A discussion of the research presented in this thesis, as well as relevant conclusions are provided in Chapter 7. This discussion also reiterates the strength and limitations of the research presented, and identifies future areas of research.

1.4 Contribution to the literature

The aim of this thesis is to contribute to three areas of the literature. Firstly, current practice in the handling, analysis and reporting of missing PROMs outcome data in RCTs is established. The review focusses on eight commonly used PROMs, and in each case identifies a larger numbers of studies reporting results including these PROMs than previous reviews. This review also contrasts current practice with contemporary guidance, which emphasise the importance of minimising the occurrence of missing data prospectively, as well as the performance of appropriate sensitivity analysis. Therefore, this review adds new aspects to the existing literature, while also identifying future areas for research.

Chapters 4 and 5 contribute to the methodological literature on analytical approaches for missing PROMs outcome data. Firstly, the limited evidence base regarding the level at which MI should be applied, i.e. the composite score, subscale or item level, is validated and extended to additional PROMs and data sets. Furthermore, extensive simulation studies are performed to establish best practice for the analysis of longitudinal data comprising some missing PROMs. Again, the current literature lacks direct comparative assessments of established statistical methods focussing specifically on PROMs data and RCT contexts, and therefore this research is considered to be an important contribution to the methodological literature in this area.

Finally, the use of appropriate sensitivity analysis related to the way missing PROMs data were handled in the reports of RCTs is very limited. Therefore, this thesis contributes to the literature by summarising advice and guidance by methodologists and regulatory bodies on the performance and interpretation of sensitivity analysis. In addition, two case

studies are used to demonstrate the application and interpretation of sensitivity analyses in realistic RCT settings. Examples provided are easily implementable using standard statistical software and publication of this reasearch may contribute to an increased uptake and reporting of such essential analyses within RCT publications.

Chapter 2 : The current practice of handling and reporting missing patient reported outcomes data in the publication of randomised controlled trials – a review of the literature

2.1 Introduction

The presence of missing data within statistical analysis, including medical research, has long been recognised as a potential source of bias in the reporting of the results from research, including RCTs[1, 21, 49-51], as outlined in Chapter 1. This chapter investigates more closely the handling and reporting of missing data for PROMs in a review of the current literature, in order to generate an overview of current practice, which is contrasted with existing methodology and guidance.

In order to adequately assess any potential shortcomings in the handling and reporting of missing data, an understanding of the possible origins of and mechanism underlying missing data, as well as different approaches of handling missing data is required, which is discussed in section 2.3. In section 2.4, an overview of the existing literature on this topic is presented, and the need for an updated literature review is justified. Section 2.6 introduces the PROMs included in this literature review, and provides the rationale for focussing on these specific measures; the search strategy is summarised in section 2.7. The results of this literature review are provided in section 2.8, and are discussed and critically evaluated in section 2.9; conclusions of this research are provided in section 2.10.

2.2 Objectives for this chapter

This research aims to investigate the current literature to:

- Create a comprehensive overview of the current practice of the handling, analysis and reporting of missing PROMs outcome data, thus updating and adding to previous reviews
- To compare the currently used methods to handle, analyse and report missing PROMs outcome data in RCTs against recommended best practice

2.3 Background

Missing data are defined as values that were planned to be collected within the remit of a study and which are considered relevant for the analysis and interpretation of a study, but which are unavailable at the time of the analysis[18]. Missing data in PROMs occur in clinical trials when the PROMs data is not received by the trial team for data entry. The contemporary literature distinguishes between two types of missing data, namely missing items[52], also referred to as 'item non-response'[53], where no responses have been indicated for individual items, i.e. questions of a PROMs questionnaire, and 'unit non-response'[54], whereby no data is received for an entire PROMs questionnaire.

Reasons for the former type of missing data, i.e. missing items include cases where trial participants only completed their questionnaires partially, i.e. because they may choose not to answer some questions, not understand some questions/consider them irrelevant, or omit some questions by mistake. Sometimes the omission of one or more questions can make the calculation of the final score or relevant subscale impossible (more detail provided in section 2.6). On the other hand, unit non-response may occur where the research team fails to send or hand out questionnaires to participants at the appropriate time points, questionnaires are left uncompleted, questionnaires are lost, participants are no longer contactable or no longer willing/able to take part in the research[54]. Both item-non-response and unit-non-response can prevent the composite PROMs score, derived from the individual items, from being calculated.

Missing data can lead to decreased precision[20] and the introduction of bias in the treatment effects if missing data are related to the study outcomes, relevant prognostic factors or the study intervention[18]. It is crucial to understand the mechanism by which data is missing in order to assess whether or not study results may be biased.

2.3.1 A note on missing data mechanisms

The potential for bias in an analysis of data including missing values in the outcomes variable is related to the underlying mechanism of the missing data. Rubin identified the three mechanisms of missing data defined below[21]. For their algebraic representation, the following notation is used, based on the book by Carpenter and Kenward[50]:

- $\boldsymbol{Y}_i = (Y_{i,1}, Y_{i,2}, \dots, Y_{i,p})^T$ denotes the p variables intended to be collected from the i^{th} unit (e.g. participant), where $i = 1, \dots, n$

  Whereby:

  - $Y_{i,O}$ denotes the subset of p variables that are observed
  - $Y_{i,M}$ denotes the subset of p variables that are missing

- $R_{i,j} = 1$ if $Y_{i,j}$ observed and $R_{i,j} = 0$ if $Y_{i,j}$ missing for all individuals $i = 1, \dots, n$ and all variables $j = 1, \dots, n$

  Consequently, the missing data mechanism can then be written as $\Pr(\boldsymbol{R_i}|\boldsymbol{Y_i})$, i.e. the probability of observing unit $i's$ data given their potentially unseen values $\boldsymbol{Y_i}$.

Missing completely at random (MCAR) describes the scenario whereby the probability of data being missing is not related to the values of the observed or missing data. In effect, this means that all outcomes have the same probability of being missing, and the characteristics of the participants with observed data are the same as the outcomes of individuals with missing data. Therefore, the subset of participants with available outcome data is as representative of the target population as the total number of randomised participants, and unbiased results can be obtained even in an unadjusted analysis (CCA). Algebraically, the MCAR mechanism can be represented by the following formula:

$$\Pr(\boldsymbol{R_i}|\boldsymbol{Y_i}) = \Pr(\boldsymbol{R_i})$$

Missing at random (MAR) is used to define data where the probability of data being missing depends on the values of other observed data, but is independent of the values of the missing data, given the observed data.

Algebraically, this can be expressed as:

$$\Pr(\boldsymbol{R_i}|\boldsymbol{Y_i}) = \Pr(\boldsymbol{R_i}|\boldsymbol{Y}_{i,O})$$

In their book, Carpenter and Kenward[50] point out that this relationship does not dictate that 'the probability of observing a variable is independent of the value of that variable. […] Crucially though, given the observed data, this dependence is broken'.

The expression **missing not at random (MNAR)** is used when the probability that data are missing is related to the underlying value of that data, and this dependence remains to some extent even when the observed data is taken into account.

Mathematically, this missing data mechanism can be written as:

$$\Pr(\boldsymbol{R_i}|\boldsymbol{Y_i}) \neq \Pr(\boldsymbol{R_i}|\boldsymbol{Y}_{i,O})$$

The difference between these missing data mechanisms can be illustrated by be following example: Data on PROMs may be collected within an RCT. Before the analysis, it is noted that PROMs data is missing for a considerable proportion of participants. If certain questionnaires were not sent out due to a computer error or staff changes, the probability of a missing outcome is entirely independent from any observed and unobserved variables and participant characteristics, and therefore the data is said to be MCAR. On the other hand, if data is more likely to be missing for participants with worse health outcomes, possibly because their worsening health status does not permit them to complete the questionnaires, and no other data is collected to capture this relationship, then the data is missing according to the MNAR mechanism. However, if additional data on health status

had been collected either through clinical examination or the use of predictive baseline variables, or the reason for missing follow-up assessments had not been recorded, then the data is MAR.

2.3.2 Overview of commonly used approaches to analysing missing data

The literature describes many different approaches for handling missing data in a statistical analysis; the most common methods are detailed in Table 2-1[20, 29, 55-59]. This list is not exhaustive, but focusses on the analysis approaches covered in the subsequent literature review. All of these methods make assumptions about the underlying missing data mechanism, which cannot be verified using the available data[18, 50]. If the assumptions are not valid, the results may be biased.

Table 2-1: Overview of common approaches of handling missing outcome data

Imputation method	Complete case analysis (CCA)	Last observation carried forward (LOCF)	Mean imputation	Regression imputation	Maximum likelihood	Multiple imputation (MI)
Category	No imputations	Single imputation	Single imputation	Single imputation	No imputations	Multiple imputation
Description of imputation method	Only participants with non-missing outcome data are utilised, and therefore potentially a high proportion of participants is excluded from the analysis.	For missing data in repeatedly measured variables, the last observed value (even if measured at baseline) is used to replace missing data.	Missing data is replaced by the mean value of all observed values for the relevant variable and time point.	A regression model is built, which is used to predict the missing values based on other data observed for the subject.	Usually used for longitudinal outcomes, parameter estimates are obtained through an ite-rative process so as to maximise the likelihood of producing the sample data[60]. Information from the available data points is utilised to make inferences about the missing data.	Missing observations are replaced by plausible values based on the observed values in the dataset. This process is repeated a pre-specified number of times and results are pooled using relevant rules.
Assumptions made about missing data mechanism	MCAR (or MAR in an appropriately adjusted analysis)	Data does not change after the last observation was made.	Participants with missing data, on average, have the same outcomes as those with observed data.	MAR – based on variables in imputation model	MAR – based on variables in analysis model	MAR – based on variables in imputation model
Method specific features:	**Disadvantages:** The sample size is reduced, reducing, in turn, the power of the statistical tests.	**Disadvantages:** The above described methods do not allow for uncertainty around these estimates. This leads to spuriously low standard errors, which in turn produce confidence intervals more narrow than expected, and lower p-values, which can lead to new interventions being rejected or approved inappropriately.			**Advantages:** Can handle MAR data when the analysis considers longitudinal follow-up.	**Advantages:** By taking into account the uncertainty around the imputed values, appropriate standard errors are produced.

2.4 Previous missing data reviews

As outlined in Chapter 1, missing data has long been recognised as a potential source of bias. Although the methodology on this topic is well developed, uptake of the proposed and recommended methods appears to lag behind their theoretical foundation[23, 25, 61-63]. Over the past ten years, a number of academic papers have assessed how missing data has been handled in peer-reviewed publications; more details on six such studies are presented in Table 2-2.

In general, the reports agree in their findings that missing data is often handled inadequately compared to current methodology, the description of the missing data is frequently lacking, and assumptions about the missing data mechanism are often not clarified. There is also general consensus that the CC analysis may be used inappropriately, or that other unsuitable methods such as LOCF are utilised. Moreover, the use of sensitivity analysis to assess the robustness of the results with regards to the assumptions made about missing data are often not performed or inappropriately implemented. All reviews agree that the handling and reporting of missing data in medical research is suboptimal and ought to be improved.

Similar reviews have also been performed to assess the handling of missing data in trial based cost-effectiveness analyses. This includes a paper by Noble et al[64], which showed shortcomings in the reporting, methodological approaches and performance of sensitivity analysis.

Table 2-2: Overview of studies looking at how missing data was handled in the literature

Authors	Wood et al[63]	Fielding et al[22]	Deo et al[65]	Eekhout et al[23]	Powney et al[62]	Bell et al[25]
Published	2004	2008	2011	2012	2014	2014
Type of studies included in the review	RCTs published between July and December 2001 in 'major medical journals' (BMJ, JAMA, Lancet and New England Journal) – concentrating on primary analyses.	RCTs using QoL outcomes (primary or secondary) published between 2005/06 in four the leading journals (BMJ, JAMA, Lancet, NEMJ).	RCTs in adults with chronic kidney disease, published in 2007/08 excluding time to event analysis, looking at reporting of the primary outcome.	Epidemiological studies using questionnaires published in 2010 in three leading journals – impact factor > 5.0 (American Journal of Epidemiology, Epidemiology, International Journal of Epidemiology).	RCTs with longitudinal follow-up of non-binary outcomes; published between 2005-2012, with 100 papers randomly selected. No restrictions on journals.	RCTs published in top medical journals (BMJ, JAMA, Lancet, New England Journal of Medicine) published between July and December 2013.
Outcomes considered	All types of outcomes except time to event	QoL measures	Outcomes related to chronic kidney disease	Outcomes using questionnaires	All non-binary outcomes	All outcomes excluding survival data
Studies included	71	61	110	262	100	77
Comments	The authors state that as the journals with highest impact and a presumably thorough statistical review processes were reviewed, 'it is hard to imagine that the situation would be better in [other] journals'.	The paper included nine different generic QoL measures – however, numbers for individual questionnaires are low, and it is unclear how generalisable the findings are.	Not restricted to specific journals, but focussing on RCTs in a specific disease area.	The authors state that the papers included represent research at least as good as actual research in the field as a whole; implying that actual practice may be overestimated.	The authors aimed to provide representative assessment and the review was not restricted to specific journals.	The authors comment that review focussed on top four journals and conclude that this may have under-estimated the extent of missing data and overestimated the use of sensitivity analysis

Authors	Wood et al[63]	Fielding et al[22]	Deo et al[65]	Eekhout et al[23]	Powney et al[62]	Bell et al[25]
Main findings	Overall: **"missing outcome data are a common problem in RCTs, and are often inadequately handled in the statistical analysis in the top tier medical journals"** • Insufficient description of missing data • Repeated measures are often ignored in CCA when intermediate measures could be used • Use of CCA despite lager percentage of missing data • Widespread use of inappropriate methods (e.g. LOCF) • Wrong/ inconsistent use of ITT • Low use of sensitivity analysis; if performed of varying quality and described in varying detail	Overall: **Clearer reporting needed on methods used and the amount of missing data** • Often insufficient description of missing data • Simple or no imputation methods commonly used although they may be inappropriate in many circumstances • Assumptions underlying each method of handling missing data usually not discussed • Rationale of imputation methods not described • Lack of discussion of the potential impact of missing data on results (possible introduction of bias)	Overall: "**Major deficiencies in transparency and completeness of reporting of data lost in primary outcome analysis"** • Exclusions from analysis (due to missing data) not always clearly described • Reasons for missing data not clearly described • ITT analysis does not always include all randomised participants • Lack of clarity about methods of imputation used • Inappropriate methods of imputation used • Differential drop-out rates observed (by treatment arm)	Overall: **"The reporting of missing data in epidemiological studies is highly variable and mostly poor"** • Criticise lack of distinction between missing items and complete missingness in multi-item instruments • Lack of clarity about the extent of missing data within the analysis • Authors unclear about assumed missing data mechanism • Too many studies use CCA, or inappropriate single imputation methods • Sensitivity analysis mostly not performed	Overall: "[..] **a large proportion of papers failed to recognise the issue of missing data"** • Reasons for the choice of methods for handling missing data are not clarified • Reasons for missing data are important to assess the appropriateness of the assumptions made about the missing data • Insufficient information is provided on missing data	Overall: **"Applied researchers and statisticians need to improve their handling of missing data in RCTs."** • Statistical methods research is not applied sufficiently • Appropriate sensitivity analysis (i.e. changing assumptions about the missing data mechanism) are rare • Methods known to produce bias are applied to many primary analyses • The uptake of MI remains low • The definitions of ITT and modified ITT were inconsistent across the trials • Attempts to reduce missing data were reported in ~34% of studies

Authors	Wood et al[63]	Fielding et al[22]	Deo et al[65]	Eekhout et al[23]	Powney et al[62]	Bell et al[25]
On prevention of missing data	Not assessed. The authors refer to the literature for emphasis on avoiding missing data, and the use of secondary data sources.	Not assessed. Authors mention that ideally missing data should be reduced/ avoided from the outset, partly through appropriate data collection methods.	Not mentioned	Not mentioned	Not mentioned	Assessment of planning for and prevention of missing outcome data was one of the authors' secondary aims. "Prevention is the best way to handle missing data, so more effort needs to be put into missing data at the design and conduct stage."

2.5 Current guidance and rationale for this review

The literature on how missing data should be handled and reported in clinical research, comparisons of different methods of handling missing data, is manifold. Li et al[18] reviewed the literature for recommendations on methods to prevent, handle and report missing data. They presented a set of ten possible guidelines or "minimum standards", which were agreed on using a Delphi consensus, i.e. a consensus among a multidisciplinary team of experts, encompassing areas of study design, conduct, analysis and reporting.

The review conducted in the context of this thesis utilises some of the recommendations by Li et al as the basis for assessing whether missing data was handled and reported appropriately in the current literature. However, it should be noted that other researchers and regulatory agencies such as the Food and Drug Administration (FDA), the regulatory body in the US and the European Medicines Agency (EMA) provide similar guidance methodologies[1, 17, 40, 51, 66-69].

This review specifically focussed on the following recommendations:

- **Standards of study design/conduct**
    - Studies should be designed and conducted in order to minimise the amount of missing data occurring (Standard 2), and the reporting should include details of such steps taken
- **Standards of analysis**
    - The analytical methods used should be able to account for the uncertainty associated with the missing data (Standard 6)
    - The use of single imputation methods should be discouraged (Standard 7)

- Appropriate sensitivity analysis with regards to the assumed missing data mechanism should be performed (Standard 8)

- **Standards of reporting**
  - All randomised participants should be accounted for when reporting on the study results (Standard 9)
  - Any reports should include details on the data completeness and the methods used to handle missing data, as well as 'the potential influence of missing data on the study results [..]'[18]

Other recommendations by Li et al that were not addressed in this chapter include:

i) Other aspects of the study design, such as, a clear description of the research question and relevant outcomes and the importance of pre-specifying statistical methods for handling missing data.

ii) Standards on the study conduct, such as, the importance of continued data collection on key outcomes even after a participant decides to withdraw from the intervention or is withdrawn from aspects of the study, as well as details of and reasons for the withdrawal, and the continuous monitoring of missing data.

While the above are also considered important in the design and conduct of clinical research, it was felt that they are less likely to be assessable via academic papers reporting the results RCTs, and are therefore not included in the review.

Although published in 2014, the above mentioned recommendations form the basis of the review of the handling and reporting of missing data in current publications ranging from 2009 to 2013. This is because the proposed minimum standards consolidate recommendations for researchers already available from other sources, including the

CONSORT Statement, which raised the importance of clearly describing the patient flow through RCTs as early as 1996[70], and more recently the CONSORT PRO in 2013, which specifically addresses the reporting of PROs in RCTs[71]. In addition, a paper by Curran et al[54] in 1998, highlighted the problems associated with using the LOCF or single imputation methods such as mean and regression imputations as methods for handling missing data. Most of the recommendations provided are also mentioned in at least some of the previously discussed reviews of how missing data is handled and reported in the current literature[22, 23, 62, 63, 65], as discussed in section 2.4.

In the context of this thesis, a new review of how missing data is handled and reported in the current literature was conducted to add to the findings of the above mentioned previous studies, with the inclusion of more recently published RCT results, and a broader range of disease areas and journals to create a very generalisable account of current practice. The current review focusses specifically on PROMs endpoints, which are now increasingly used in RCTs as primary or important secondary outcome measures[71]. Although Fielding et at[22] were also focussing specifically on PROMs, large numbers of studies were not identified for any particular outcome measure, which were felt to be needed to establish a generalisable overview of the methods used to handle and report missing data occurring within these PROMs.

In addition to the above, the review undertaken in the context of this thesis adds to the existing literature by specifically investigating if the potential impact of missing data on the study results has been discussed, and if steps were taken at the design and planning stage to minimise the amount of missing data within the study, and if these were described sufficiently in the publication. The previous reviews have not considered these features,

the latter of which is considered as an important, and possibly even the most important step in addressing missing data[40].

2.6 Questionnaires to be used in this review

Due to the large number of PROMs available and their extensive use in RCTs, it was felt necessary to select a sample of representative instruments and search the literature specifically for RCT publications utilising these questionnaires. This approach has the advantage that large numbers of studies using a specific PROM can be identified (section 2.5), thus increasing the generalisability of the conclusions made with regards to these PROMs.

HRQL measures can be categorised into two groups: generic and specific instruments[5]. Generic instruments apply to a variety of populations and include generic "health profiles", which aim to capture all important aspects of HRQL, and "utility measures", which are often used in health economic analysis as they are able to "reflect both the health status and the value of that health status to the patient"[72]. Disease specific questionnaires have been designed to be used within specific health conditions, while site or location specific questionnaires focus on assessing complaints or improvements related to the specific anatomical sites.

Two questionnaires from each of these four categories are used in this review; details are provided in Table 2-3. All of these questionnaires are have been widely adopted, assessed for validity and reliability, translated into a number of languages and align with the author's research interests and experience.

Of note, multi-item PROMs, as considered here, consist of multiple items, i.e. questions, which are combined in an overall score. Here, the focus is on missing data in the composite scores.

Table 2-3: Questionnaires to be included in the literature search

Questionnaire	Questionnaire category	Brief description of the questionnaire
EQ-5D-3L	Utility questionnaire	The EQ-5D-3L[45, 46] questionnaire consists of five questions assessing participants' mobility, self-care, usual activities, pain/ discomfort and anxiety/ depression. The answers can be converted into a composite score using a country specific algorithm, whereby a score of 1 indicates full health and lower scores indicate worse health states. Scores below zero indicate health states worth than death. The full EQ-5D-3L questionnaire also contains a visual analogue scale ranging from 0 to 100, which is not the focus of this review. The scoring manual does not provide any provision to impute values if one or more items are missing, and the EQ-5D-3L cannot be calculated in the presence of missing data.
HUI	Utility questionnaire	The currently used HUI2 and HUI3 collect a measurement of ability/ disability with regards to different health-state attributes, which can be used complementary to each other[73]. The HUI2 comprises of seven attributes, namely sensation, mobility, emotion, cognition, self-care, pain and fertility, each consisting of three to five levels. The HUI3 consists of eight attributes, vision, hearing, speech, ambulation, dexterity, emotion, cognition and pain on a five or six level Likert scale. The attributes are combined into a score ranging from 0 (equivalent to death) to 1 (full health). The scoring manual does not provide any provision to calculate the scores if one or more items are missing; thus any item-nonresponse leads to missing composite scores.

Questionnaire	Questionnaire category	Brief description of the questionnaire
OHS	Site/ location specific questionnaire	A questionnaire designed to assess the outcome of total hip replacements from a patient perspective[74] in terms of pain and function. The 12 questions are combined in a single score ranging from 0 to 48, whereby higher scores indicate better outcomes[42]. The OHS scoring manual stipulates that the OHS can be calculated when up to two items are missing by substituting the missing item scores by the average of all available item scores.
OKS	Site/ location specific questionnaire	A questionnaire designed to assess the outcome of total knee replacements[41] in terms of pain and function. As for the OHS, the 12 questions are combined to a single score ranging from 0 to 48, whereby higher scores indicate better outcomes in terms of pain and function[42]. The OKS scoring manual stipulates that the OKS can be calculated when up to two items are missing by substituting the missing item scores by the average of all available item scores.
PDQ	Disease specific questionnaire	The PDQ (Parkinson's disease questionnaire) has been developed to assess QoL in patients with Parkinson's disease[75]. The PDQ-39 consists of 39 questions addressing aspects of mobility, activities of daily living, emotions, stigma, social support cognitions, communication and bodily discomfort. Answers are combined to a standardised score ranging from 0 to 100 with lower scores indicating lower functioning and wellbeing for the subscales and a single index score. The PDQ-8 is a shortened version of the PDQ-39, consisting of only eight questions, which are scored in line with the PDQ-39[48]. In the presence of any missing data, relevant subscales (PDQ-39) and the single index scores (PDQ-8 and PDQ-39) cannot be calculated. The user manual suggests handling missing data via an expectation maximisation algorithm. However, this is not part of the calculation of the PDQ scores, and not further considered within this thesis.

Questionnaire	Questionnaire category	Brief description of the questionnaire
EORT QLQ-C30	Disease specific questionnaire	EORTC QLQ-C30 (European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire-Core 30) consists of a total of 30 questions including five functional scales (physical, role, cognitive, emotional and social), three symptom scales (fatigue, pain, and nausea and vomiting), one global health, one QOL scale, and six symptom items (dyspnoea, insomnia, appetite loss, constipation, diarrhoea and financial difficulties) [76, 77]. As a simple method for handling missing items, the scoring manual suggests that where at least half the items for a subscale are available missing data can be ignored in the calculation of the scale scores. This is equivalent to assuming that the responses to the missing items are equal to the average of all completed items for that subscale. The scoring manual, however, indicates that this assumption may not always be appropriate and should be applied with caution.
SF-12	Generic health profile	The SF-12 (Short Form 12 health survey questionnaire) is a reduced version of the SF-36, covering general health, a physical component score and a mental health component score[44]. The scoring manual includes no imputation rule for missing data, implying that subscales cannot be calculated in the presence of even a single missing item[78].
SF-36	Generic health profile	The SF-36[79] (Short Form 36 health survey questionnaire) consists of 36 items measuring health in the following eight dimensions: physical functioning, social functioning and role limitations to measure functional status, mental health, vitality and pain to measure wellbeing and finally general health perception. The scoring manual includes no imputation rule for missing data, implying that subscales cannot be calculated in the presence of even a single missing item[78].

2.7 Search strategy

A systematic search of EMBASE, PubMed and Web of Science, as well as the NHS Economic Evaluation Database (NHS EED) for the two utility questionnaires only, was performed to identify articles reporting the finding of RCTs using PROMs as either a primary or secondary outcome. Broad search terms were used to avoid missing relevant research papers. Details on all search terms used, and the number of results obtained for each iteration are shown in Appendix 1. Due to large numbers of articles identified, searches were restricted to 2013 for the EQ-5D-3L, QLQ-C30, SF-12 and SF-36, while data extraction was extended to include years 2009–2013 for the HUI, OHS, OKS and PDQ.

Fewer results were expected from the NHS EED database. This is because its content concentrates on assessments of costs and outcomes of competing interventions, and relies on studies having been reviewed and critically appraised by staff at the NIHR Centre for Reviews and Dissemination at the University of York. Therefore, there may be a considerable time lag between a study being published and it having been reviewed and added to this database.

Identified references were exported from each database into the reference manager (EndNote X7). The references for each questionnaire were combined and screened for eligibility after duplicates had been removed. Initial screening involved identifying references that were not research articles reporting the results of definitive RCTs with the help of titles, abstracts and full texts as appropriate. Full texts were obtained where the information in title and abstract was insufficient to group the article into one of the screening categories. See Figure 2-1 for details on the characteristics of studies that were excluded. Further screening of the identified RCTs (see Figure 2-2) aimed to exclude trials that were not using a parallel group design or were considered to be of insufficient size

(<50 randomisations per arm) for a robust evaluation. The generalisability from smaller studies is likely to be unreliable. Therefore, a requirement of at least 50 randomised participants per trial arm was chosen to include studies of sufficient size permitting the use of potentially complex methods of handling missing data and quantitative assessments between treatment arms.

RCTs with more than two arms were included into this review. However, where extraction of information by trial arm was required, data was extracted for two arms only, namely for the randomisation group considered the control treatment or usual care, as well as for the arm using the combination of most drugs or most frequent intervention appointments, to enable comparisons across all trials.

Details on the reporting of missing data, including the attrition rates, information provided in the analysis section and reporting of missing data within each article was extracted using a pre-specified set of criteria. A full list of the data extracted can be seen in Appendix 2.

Studies with the following characteristics were excluded:

- Cross-sectional or cohort studies, as well as case series, or articles only using baseline data of an RCT, or one arm of an RCT
- Conference abstracts, book chapters or comments and editorials
- Work validating or assessing measurement properties of questionnaires
- Pilot and feasibility studies
- Trials using questionnaires other than those specified, but which use the same abbreviations
- Methodological work
- Publications of protocol or statistical analysis plans
- Reviews of previously reported trial results and analyses combining data from more than one trial (including meta-analyses and systematic reviews)
- Simulation work based on data from RCTs
- Articles not written in English

Figure 2-1: Exclusion criteria employed during the screening process

RCTs with the following characteristics were excluded from the review:

- RCTs analysed as factorial designs
- Cross-over studies
- RCT reports not focussing on the relevant questionnaire as an endpoint (i.e. the article may have been picked up in the search as a specific questionnaire has been mentioned in the paper, but was not an outcome measure in the trial)
- Monographs
- Trials with less than 50 participants randomised to each trial arm

Figure 2-2: Additional exclusion criteria applied to the identified RCTs

Data as reported by the authors in the main text or tables of the publication were transferred onto the data collection spreadsheet with the following exceptions. By definition, the Intention To Treat (ITT) population should include all randomised participants, regardless of their compliance with the allocated intervention, the protocol, or, in fact, the absence of relevant outcome data, as any other form of analysis is likely to be subject to bias[80]. Where authors described their analysis as based on the ITT population, but not all randomised participants were included into the analysis, the analysis population recorded in the data extraction spreadsheet was described as 'modified ITT'.

Endpoints were classed as primary or secondary in the context of the paper. Endpoints were considered as secondary endpoints unless they were clearly labelled as primary outcome measures, used in the sample size calculation, or were secondary endpoint of the study, but used as the primary focus of a particular paper. Also, where the primary endpoint encompassed a health economic evaluation which was based on the EQ-5D-3L or the HUI, these utility measures were considered primary endpoints.

Attrition was defined as the difference in the number of participants randomised to the relevant trials arms and the amount of PROMs outcome data available at the specified primary follow-up time point. Information on attrition for the relevant questionnaire was extracted only where it was reported specifically for the outcome score in question. Where

overall attrition in the trial was reported, these figures were considered inaccurate, as it was thought likely that some PROMs data may be missing even for participants who attended all follow-up assessments.

As a minimum, abstracts, methods and statistical analysis sections were read in full for each article, but a keyword search was employed to identify other relevant information to be extracted. A full list of the keywords and search strategies used to identify relevant data can be seen in Appendix 3. Web-tables and online material were only reviewed when specifically referenced in the text of the article.

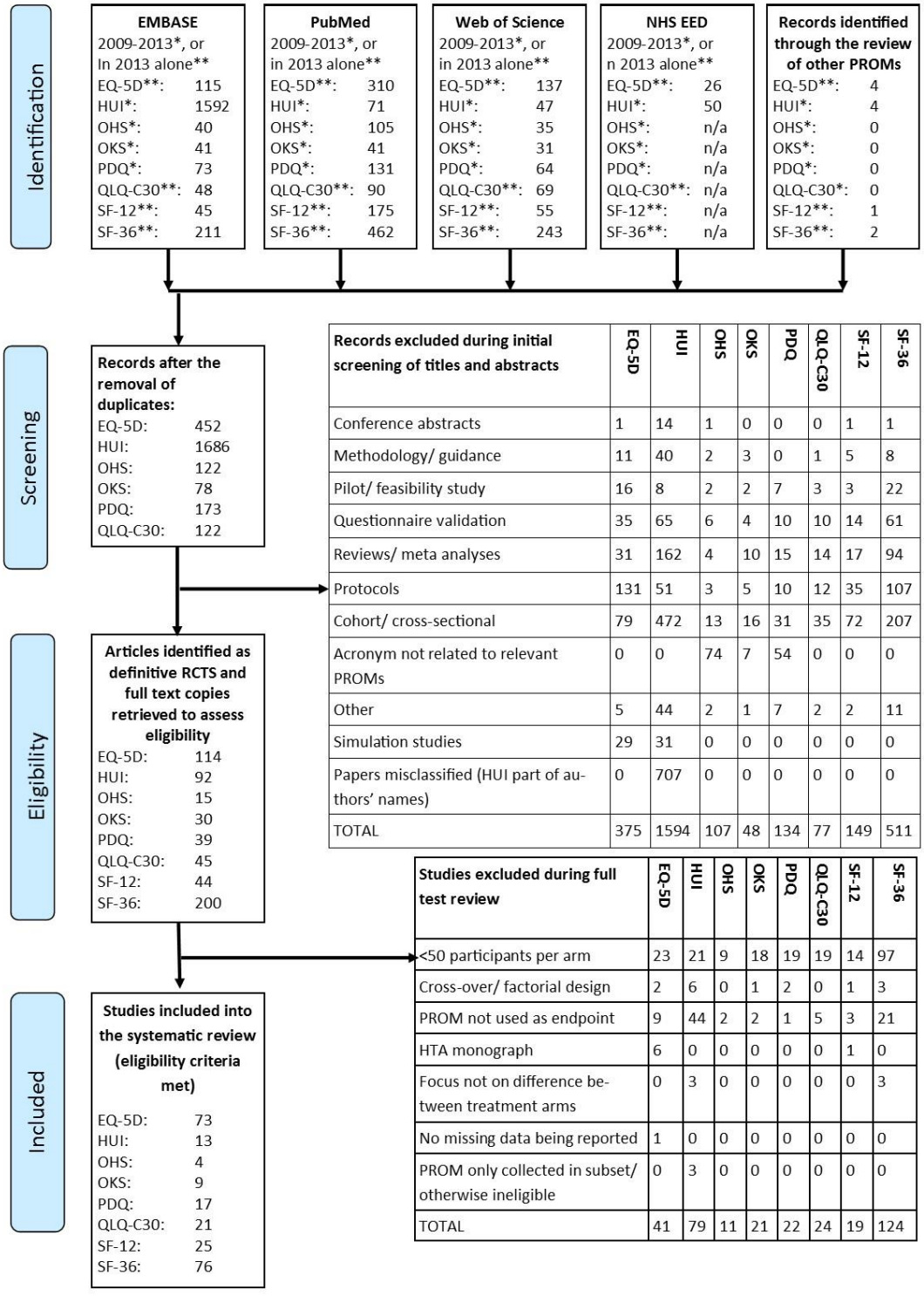
2.8 Results

2.8.1 Screening and identification process

The results show that the number of identified eligible studies varies widely between PROMs, from over 70 studies using the EQ-5D-3L and SF-36 identified in 2013 alone, to less than ten studies utilising the OKS and OHS identified between 2009 and 2013.

Summaries of the database searches for all PRO measures are summarised in a PRISMA diagram in Figure 2-3.

Where an eligible publication reports on several of the pre-specified outcome measures, this study was included within the summaries for all relevant PROMs.



Records excluded during initial screening of titles and abstracts	EQ-5D	HUI	OHS	OKS	PDQ	QLQ-C30	SF-12	SF-36
Conference abstracts	1	14	1	0	0	0	1	1
Methodology/ guidance	11	40	2	3	0	1	5	8
Pilot/ feasibility study	16	8	2	2	7	3	3	22
Questionnaire validation	35	65	6	4	10	10	14	61
Reviews/ meta analyses	31	162	4	10	15	14	17	94
Protocols	131	51	3	5	10	12	35	107
Cohort/ cross-sectional	79	472	13	16	31	35	72	207
Acronym not related to relevant PROMs	0	0	74	7	54	0	0	0
Other	5	44	2	1	7	2	2	11
Simulation studies	29	31	0	0	0	0	0	0
Papers misclassified (HUI part of authors' names)	0	707	0	0	0	0	0	0
TOTAL	375	1594	107	48	134	77	149	511

Studies excluded during full test review	EQ-5D	HUI	OHS	OKS	PDQ	QLQ-C30	SF-12	SF-36
<50 participants per arm	23	21	9	18	19	19	14	97
Cross-over/ factorial design	2	6	0	1	2	0	1	3
PROM not used as endpoint	9	44	2	2	1	5	3	21
HTA monograph	6	0	0	0	0	0	1	0
Focus not on difference between treatment arms	0	3	0	0	0	0	0	3
No missing data being reported	1	0	0	0	0	0	0	0
PROM only collected in subset/ otherwise ineligible	0	3	0	0	0	0	0	0
TOTAL	41	79	11	21	22	24	19	124

Figure 2-3: PRISMA flow chart detailing the identification process of articles to be included into the review

2.8.2 Study characteristics and missing data observed

The sample size of the RCTs included in this review varied greatly, from a sample size of 100, the cut-off for eligibility to be included into the review, to over 18,000 participants from 43 countries[81].

The percentage of studies using the relevant PROMs as a primary outcome measure was highest for those utilising the HUI with almost 70%, and lowest for the QLQ-C30, SF-12 and SF-36. In the case of the QLQ-C30, the RCTs often favoured primary endpoints focussing on survival or progression-free survival, while RCTs utilising the latter two PROMs used endpoints that were either less objective or more disease specific.

Additional details on study characteristics of the identified RCTs are provided in Table 2-4. The median percentages of PROMs data available at the primary follow-up time point were between approximately 50% and 83% within the different PROM categories, and there was evidence of some degree of differential drop-out between trials arms in most studies, as described in Table 2-5.

Table 2-4: Overview of the characteristics of the identified RCTs by PROM category

Questionnaires	EQ-5D-3L index	HUI	OHS	OKS	PDQ	QLQ-C30	SF-12	SF-36	Overall
Number of studies	72	13	4	9	17	21	25	76	237
Years searched	2013	2009-2013	2009-2013	2009-2013	2009-2013	2013	2013	2013	2009-2013
Studies using PROM as a primary outcome	38.9%	69.2%	25.0%	44.4%	41.2%	23.8%	24.0%	26.3%	33.8%
Size of studies*									
Median	329	337	155	165	294	309	241	202	251
IQR	190, 600	139,622	n/a	120, 200	184, 359	178, 420	195, 392	138, 304	159, 416
Range	100, 18,624	104, 762	126, 161	100, 1715	109, 586	108, 1528	116, 1573	100, 3.057	100, 18,624
Follow-up data was measured repeatedly (opposed to once only)	87.5%	92.3%	50.0%	88.9%	76.47%	81.0%	84.0%	77.6%	82.3%
Length of follow-up to primary assessment time point (months)									
Median	12	12	18	12	6	12	9	15	12
IQR	6, 17	9, 24	7.5, 36	12, 24	4, 10.5	6, 12	6, 15	4, 12	6, 15
Range	1, 60	6, 36	3, 48	3, 60	3, 36	0.25, 78	1.5, 24	0.75, 60	0.25, 78

*the size of the studies described here refers to the number of participants randomised to the two relevant treatment arms considered in this review.

Table 2-5: Overview of the amount of missing data within the identified RCTs by PROM category

Questionnaires	EQ-5D-3L index	HUI	OHS	OKS	PDQ	QLQ-C30	SF-12	SF-36	Overall
Number of studies	72	13	4	9	17	21	25	76	237
% of data available at primary analysis time point (overall)**	(n=37, 51.4%)	(n=4, 30.8%)	(n=2, 50.0%)	(n=4, 44.4%)	(n=4, 23.5%)	(n=10, 47.6%)	(n=10, 40.0%)	(n=24, 31.6%)	(n=95, 40.1%)
Median	74.8%	76.2%	63.3%	83.7%	83.2%	50.7%	68.6%	84.2%	75.0%
IQR	59.7%, 85.7%					47.6%, 74.6%	61.9%, 80.8%	69.7%, 94.7%	57.1%, 86.2%
Range	34.1%, 91.6%	50.7%, 86.2%	55.9%, 70.7%	62.4%, 98.8%	51.8%, 94.5%	35.1%, 85.4%	37.1%, 90.5%	26.0%, 99.2%	26.0%, 99.2%
% difference in follow-up data (%) available (active – control)**	(n=35, 48.6%)	(n=3, 23.1%)	(n=2, 50.0%)	(n=4, 44.4%)	(n=3, 17.6%)	(n=7, 33.3%)	(n=9, 36.0%)	(n=24, 31.6%)	(n=87, 36.7%)
Median	0.3%	3.7%	-2.0%	-2.2%	4.91%	6.6%	5.1%	-0.5%	0.3
IQR	-4.0%, 4.0%					2.4%, 12.3%	-5.2%, 7.7%	-3.6%, 2.0%	-3.2, 5.1
Range	-15.7%, 10.9%	-1.8%, 6.37%		-3.0%, 9.4%	-3.2, 9.6%	-13.1%, 13.9%	-12.9%, 11.5%	-13.4%, 13.9%	-15.7, 13.9

**The first lines of the summaries specify the number (and percentage) of studies for which this information is available.

2.8.3 Adherence with proposed reporting standards

Full details on the approaches to handling missing data are presented in Table 2-6.

With the exception of the RCTs using the OHS and SF-12, the proportion of publications referring to the use of strategies employed to minimise the occurrence of missing data within the study were around 25% or less. Methods described included the provision of pre-paid envelopes to increase the return of postal questionnaires, the substitution of clinic visits that cannot be attended by the use of postal questionnaires, telephone interviews, or home visits, as well as reminder emails, phone calls or letters where follow-up data had not been received. Other approaches included the provision of payments or rewards for questionnaire completion, the reiteration to participants and staff that data collection was encouraged even after participants withdrew from their allocated intervention, and the exclusion of potential participants that are unlikely to (be able to) comply with follow-up visits, including those with terminal diagnosis or hospice care.

Table 2-5 demonstrated that information on the number of participants with the relevant PROMs data at the main follow-up could only be obtained for approximately 55% or less of the studies. This information could not always be derived from the CONSORT flow chart, as it is likely that not all participants who are still in the study at a certain follow-up time point will have completed and/ or returned their questionnaires, or will have responded to a sufficient number of questions to enable the calculation of the PROMs scores. Not all studies split this information by trial arm, this resulted in less information relating to randomisation allocation and the difference in attrition between the trial arms being available.

Despite observing differences in the attrition rates by treatment arm, the relationship of missing data to baseline characteristics was rarely investigated. The vast majority of

publications (more than 80%) did not state the assumed missing data mechanism, and in many cases the analysis population was not clearly described.

Complete cases were a widely used analysis approach, with many authors being unclear about the primary method of handling missing data in the analysis. Multiple imputation and repeated measures models were less frequently used, in up to 16% and 25% of publications respectively.

Very few authors justified the primary method of handling missing data (between 0% and 25% by PROMs used), undertook sensitivity analysis to assess the robustness of their results with regards to the missing data in the studies (0% to 32%), or commented on the potential influence of missing data on the study results (0% to 25%). With regards to sensitivity analysis, the alternative analyses often did not include variations of the assumptions made about the underlying data mechanism. For example, where the primary analysis may have consisted of a CCA, the sensitivity analysis may have included LOCF, MI, simple imputation methods or vice versa, thus all analyses assumed a MAR mechanism, and did not investigate the possibility of MNAR.

Full details of the quantitative results from this literature review on the handling and reporting of missing data in PROMs in articles reporting the comparative results from RCTs in the current literature are presented in Table 2-6.

Table 2-6: Results from the literature review

Questionnaires	EQ-5D-3L index	HUI	OHS	OKS	PDQ	QLQ-C30	SF-12	SF-36	Overall
Number of studies	72	13	4	9	17	21	25	76	237
Methods to limit missing data described	25.0%	15.4%	50.0%	22.2%	11.8%	14.3%	36.0%	21.1%	22.8%
Differential missingness assessed* **	25.0%	23.1%	0%	11.1%	11.8%	14.3%	28.0%	18.4%	20.3%
Assumed missing data mechanism									
Not described	91.7%	92.3%	100%	100%	82.4%	100%	88.0%	96.0%	93.3%
Missing At Random	6.9%	7.7%	-	-	17.6%	-	12.0%	4.0%	6.3%
Missing Completely At Random	1.4%	-	-	-	-	-	-	-	0.42%
Missing data mentioned in methods/analysis section	62.5%	61.5%	25.0%	11.1%	75.0%	42.9%	52.0%	52.6%	54.2%
Analysis population									
Intention To Treat	27.8%	7.7%	-	11.1%	29.4%	9.5%	24.0%	19.7%	21.1%
Modified Intention To Treat	54.2%	46.2%	50.0%	66.7%	47.1%	59.1%	48.0%	46.1%	50.6%
Per Protocol	1.4%	-	-	-	5.9%	-	-	1.3%	1.3%
Unclear	16.7%	46.2%	50.0%	22.2%	17.7%	33.3%	28.0%	32.9%	27.0%
Primary method of handling with missing data									
Complete Cases	38.9%	30.8%	50.0%	22.2%	5.9%	14.3%	32.0%	39.5%	32.9%
Last Observation Carried Forward	11.1%	7.7%	-	11.1%	41.2%	9.5%	4.0%	10.5%	11.8%
Mean imputation	5.6%	-	-	-	-	-	4.0%	2.7%	3.0%
Regression imputation	-	-	-	-	-	-	4.0%	-	0.4%
Direct likelihood analysis	-	-	-	-	5.9%	-	-	-	0.4%
Repeated measures model	8.3%	23.1%	-	11.1%	17.7%	14.3%	20.0%	25.0%	16.9%
Multiple Imputation	15.3%	15.4%	-	-	-	-	16.0%	5.3%	8.9%
unclear	20.8%	23.1%	50.0%	55.6%	29.4%	61.9%	20.0%	17.1%	28.7%
Justification provided for primary method of handling missing data	13.9%	23.1%	25.0%	0%	11.8%	0%	8.0%	5.3%	9.3%
Sensitivity analysis was performed	25.0%	23.1%	25.0%	0%	17.7%	19.1%	32.0%	19.7%	21.9%
Potential influence of missing data on results mentioned in discussion	18.1%	15.4%	25.0%	0%	17.7%	14.3%	16.0%	14.5%	15.6%

*** The studies considered differences between those with complete and missing data in terms of participant (baseline) characteristics.

2.8.4 Subset of articles using the relevant PROM as a primary endpoint

The above summaries considered publications utilising the relevant PROMs as either a primary or secondary outcome, as per the inclusion criteria for this review of the literature. When focussing on the subset of articles utilising the relevant PROMs as a primary outcome measure only (80 PROMs, approximately one third of all PROMs and 23.8% to 69.2% of each relevant PROMs category), the standard of reporting improved marginally.

For some of the PROMs, an increase in the proportion of studies mentioning methods to reduce the amount of missing data within the studies could be observed, along with an increase in the number of studies clarifying how much PROMs data was available at the primary follow-up point, and an overall decrease of the amount of missing data at follow-up. Overall, a slight increase could also be observed in the proportion of articles that perform and report a sensitivity analysis. On the other hand, the proportion of studies using LOCF in their primary analysis and not clearly stating their analysis population increased when only considering the studies using the relevant PROMs as a primary outcome measure. Full details of this investigation can be seen in Table 2-7.

Table 2-7: Results from the literature review for the subset of articles using the relevant PROM as a primary endpoint

Questionnaires	EQ-5D-3L index	HUI	OHS	OKS	PDQ	QLQ-C30	SF-12	SF-36	Overall
Number of studies	28	9	1	4	7	5	6	20	80
Methods to limit missing data described	28.6%	22.2%	100.0%	50.0%	0.0%	20.0%	50.0%	30.0%	28.75%
Differential missingness assessed* **	32.1%	22.2%	0%	25.0%	0.0%	60.0%	50.0%	15.0%	26.25%
Assumed missing data mechanism									
Not described	89.3%	100.0%	100.0%	100%	71.4%	100%	83.3%	100.0%	92.50%
Missing At Random	7.1%	-	-	-	28.6%	-	16.7%	-	6.25%
Missing Completely At Random	3.6%	-	-	-	-	-	-	-	1.25%
Missing data mentioned in methods/analysis section	71.4%	66.7%	100.0%	25.0%	71.4%	80.0%	50.0%	60.0%	66.25%
Analysis population									
Intention To Treat	46.4%	-		25.0%	28.6%	20.0%	50.0%	20.0%	26.25%
Modified Intention To Treat	32.1%	44.4 %	100.0%	75.0%	28.6%	40.0%	-	50.0%	42.50%
Per Protocol	-	-	-	-	14.3%	-	-	-	1.25%
Unclear	21.4%	55.6%	-	-	28.6%	40.0%	50.0%	30.0%	30.00%
Primary method of handling with missing data									
Complete Cases	28.6%	44.4%	100.0%	25.0%	14.3%	20.0%	33.3%	45.0%	33.75%
Last Observation Carried Forward	17.9%	-	-	25.0%	28.6%	40.0%	10.0%	10.0%	15.00%
Mean imputation	14.3%	-	-	-	-	-	5.0%	5.0%	7.50%
Regression imputation	-	-	-	-	-	-	-	-	-
Direct likelihood analysis	-	-	-	-	14.3%	-	-	-	1.25%
Repeated measures model	3.6%	-	-	-	14.3%	20.0%	25.0%	25.0%	13.75%
Multiple Imputation	21.4%	22.2%	-	-	-	-	10.0%	5.0%	11.25%
unclear	14.3%	33.3%	-	50.0%	28.6%	20.0%	20.0%	10.0%	17.50%
Justification provided for primary method of handling missing data	14.3%	11.1%	100.0%	0%	14.3%	0%	16.7%	0.0%	10.00%
Sensitivity analysis was performed	35.7%	33.3%	0.0%	0.0%	14.3%	60.0%	33.3%	15.0%	27.50%
Potential influence of missing data on results mentioned in discussion	21.4%	11.1%	0.0%	0.0%	14.3%	40.0%	0.0%	15.0%	16.25%

*** The studies considered differences between those with complete and missing data in terms of participant (baseline) characteristics.

2.9 Discussion

The results of the review, as presented in section 2.8, showed that the overall quality of the handling, analysis and reporting of missing data in RCTs is lacking with regards to current methodology and guidance. Many authors did not comply with basic advice about the reporting of missing outcome data in RCTs, as also found in the previous reviews[22, 23, 63]. Failure to report adequately on the attrition in RCTs was also reported by Hopewell et al[82].

Particularly noticeable was the failure of many publications to describe clearly the extent of missing data. The lack of clarity on how missing data were handled in the analysis made it impossible for the reader to assess how much the reported results may be at risk of bias arising from missing data. Particularly if missing data occurs partly by design, authors should ensure that results and interpretations are provided within this context, instead of extrapolating the conclusions, potentially inappropriately, to the entire trial population. Missing data by design may occur if only a subgroup of participants is included into the PROMs research or because participants with disease progression or other patient characteristics are excluded from the PROMs collection. High mortality rates in certain studies also makes the collection of PROMs impossible for a large proportion of participants. In addition, the continued use of imputation methods that are known to introduce bias, such as LOCF[38, 83] further put into question the reliability of the study results. Differential drop-out was identified in numerous studies, but rarely discussed or considered in the subsequent analyses.

The importance of sensitivity analysis has been highlighted repeatedly to assess the robustness of the study results with regards to the untestable assumptions about the underlying missing data mechanism. The results in Table 2-6 showed that sensitivity

analysis was only described in up to approximately 32% of articles. However, for the majority of identified PROMs, the percentage of studies reporting sensitivity analyses was around 20%. None were reported for studies reporting on trials using the OKS as an outcome.

Even where sensitivity analyses were performed, they often did not investigate the sensitivity of the assumptions made about missing data in the primary analysis, as suggested in the current literature[18]. More specifically, where the primary analysis utilised the CC population, assuming a MAR mechanism in an adjusted analysis, reported sensitivity analyses included single or multiple imputation or repeated measures models, also assuming a MAR mechanism, or vice versa. On other occasions, the LOCF method was used as a primary analysis, with the above described methods utilised in sensitivity analyses, or vice versa. In other cases, sensitivity analyses included adding all variables that had been identified to be predictive of missing data into the model, presumably in an attempt to make the MAR assumption more plausible, although this was not justified in the text.

There were very few examples where the underlying missing data mechanism was actually varied in the sensitivity analysis. In one example, missing values in the EQ-5D-3L were substituted with values of zero, which is often describe as a health state equal to death[46], for those who died. This can be seen as a sensitivity analysis varying the assumptions at least partly, assuming worse outcomes for those who have missing data due to death (the average EQ-5D-3L score in the trial was above zero). Similarly, a trial using the QLQ-C30 imputed missing data with the worst values observed at that outcome time point, while a study using the SF-36 imputed the best and worst scores for missing data in their sensitivity analysis. However, the uncertainty around the imputed values was not appropriately taken into account in these single imputation techniques.

The potential influence of missing data on study results was discussed infrequently, possibly leaving the readers to overestimate the robustness of the results presented.

Finally, the number of publications reporting on methods to minimise the occurrence of missing data used in planning and conducting the study was found to be low. This is disappointing since no statistical analysis, however advanced, can replace information obtained by more complete follow-up. Therefore, researchers should be aware that in handling missing data 'the single best approach is to prospectively prevent missing data occurrence'[18].

The lack of detail and clarity about how missing data in the PROMs for RCTs was handled and analysed raised the question if word-limits imposed on some articles and publications may contribute to this substandard level of reporting. The inclusion of researchers trained in statistics amongst the authors, and involvement of those in the study design and analysis may contribute to an improvement in reporting standards.

2.9.1 Strength of the study

By focussing on a set of eight widely used outcome measures, this review was able to include a broad range of literature, rather than focussing on specific journals, as done in some of the previous reviews[22, 23, 63]. Therefore, a more generalisable picture of current practice was created, without necessitating assumptions on whether the identified standard of reporting was representative of the current practice, or over-reporting it.

This review adds to the current literature in that it considered more recent publications, as well as offering additional, and very important aspects to the review. Particularly, steps taken to minimising the occurrence of missing data during the design stage of the trial,

consideration of differential missing data by trial arm, justification of chosen methods for handling missing data in the analysis and an in-depth assessment of the use of sensitivity analysis were included in this review.

2.9.2 Limitations

In addition to its strengths, the study also has a number of limitations. By attempting to create as broad a picture of the current practice as possible, and including publications from a wide variety of journals, it was necessary to limit the review to a certain number of outcome measures. It is hoped that the reporting practice observed in the subset of representative outcome measures is generalisable to other PROMs. However, it is possible that there are PROMs for which the handling, analysis and reporting of missing data in PROMs is different to the estimates presented here.

Only very few eligible studies were identified for some PROMs, especially the OHS and OKS, with four and nine studies respectively included into the review. Reasons for this include that these site-specific measurements are just two of many other PROMs designed and available to be used in this disease area[84-86]. Therefore, the pool of studies utilising these PROMs is naturally smaller than for PROMs designed to measure a broader range of outcomes. Arguably, the low number of publications identified results in a less generalisable picture of the handling of missing data in these studies.

The NHS EED database was included into the search strategy for the EQ-5D-3L and HUI, as it was considered to be more reliable in identifying the utility questionnaires. However, NHS EED relies on articles having been reviewed by the York team, and therefore the

entries for 2013 may not have been as up-to-date at the time of the review as the entries for earlier years would have been.

This review did not include a quantitative assessment of the reporting of any differences between the primary analysis and any sensitivity analysis that may have been performed. This is because insufficient numbers of appropriate analyses were included into this review to justify such assessments.

The relationship between the quality of reporting and word limits imposed by journals, which may contribute to important details about missing data being omitted in favour of other relevant information could not be assessed within this review. However, tables and well-designed CONSORT flow charts can be used to report much of the information on data availability and analysis populations. Details of assumptions about missing data mechanisms, analysis strategy and sensitivity analysis can be reported briefly with one or two sentences in the main text.

The review presented has not formally assessed temporal changes in the handling and reporting of missing data, although it was found conclusions were very much in line with previously reported similar studies. The focus of this research was to identify potential insufficiencies in current practice that need to be improved. However, Fielding et al reported little change in the reporting and handling of missing data over the last decade, except for a slight increase in the use of imputation[22, 61].

2.9.3 Areas for future research

This review has identified a number of shortcomings in the current handling and reporting of missing PROMS outcome data in RCT publications. Particularly the low uptake of MI, the focus on a single follow-up time point as opposed to statistical analyses utilising the longitudinal structure of the data, as well as a lack of appropriate sensitivity analyses are considered important areas of further research. It is possible that there is a perceived lack of clarity about the implementation of these approaches that causes their low uptake demonstrated here. Therefore, these aspects of handling missing PROMs outcome data are further investigated in the following chapters using both simulation studies and case studies. Variations in missing data pattern between different RCTs and PROMs were also noted within this review, and are further investigated in Chapter 3.

2.10 Conclusions

The review provides evidence of considerate discrepancies between guidance on and current practice of the handling, analysis and reporting of missing PROMs data in the publications of RCTs. Shortcomings include the use of inappropriate methodology, lack of clarity on the extent of missing data, the assumptions made about the missing data mechanisms and the methods used to handle missing data, as well as the lack of appropriate sensitivity analyses. These factors make it challenging for clinicians, researchers, health care providers and policy makers to assess how reliable the results from RCTs are, and may even lead to health care decisions to be based on sub-optimal information.

Greater awareness needs to be created of the potential biases introduced by the inappropriate handling of missing data, and the importance of sensitivity analysis. The handling and reporting of missing data, as well as the detailed and consistent reporting of it needs to be improved to be in line with current methodology to enable an appropriate assessment of any treatment effects and the associated conclusions.

Chapter 3 : Identifying common rates of and possible predictors for missing patient reported outcome measures

3.1 Introduction

This chapter aims to generate a better understanding of missing data in PROMs by investigating the rates and patterns of missing data in three large-scale RCTs, considering PROMs used as primary endpoints, i.e. OKS[41, 42] and PDQ-39[75, 87, 88], as well as secondary endpoints, i.e. EQ-5D-3L[46, 89] and SF-12[43, 90].

The chapter starts by introducing the relevant RCTs, giving a brief overview of the research question, trial design and recruitment numbers, and clarifies which subset of participants is to be used in the following work (section 3.2). Subsequently (section 3.3), patterns of missing data within the PROMs composite scores, as well as within the items, i.e. the subscales for the PDQ-39, are examined at the five year follow-up time point, as are the missing data patterns in the composite scores over time. Possible predictors for the probability of the composite scores being missing at follow-up are investigated using summary statistics and univariate analyses in the first instance. Patient demographics and other covariate variables are considered as possible predictors (section 3.4). Finally (section 3.5), the information from the univariate analyses is combined via multivariate logistic regression models to create a more comprehensive picture of factors that are indicative of the probability of PROMs outcome data being missing.

This work aims to further understand missing data patterns, possible factors that influence the availability of data and likely missing data mechanisms in RCTs utilising PROMs data. This preliminary work is important in preparation for appropriate analyses of datasets including missing PROMs outcome data by informing multiple imputation models. In

addition, this investigation informs the simulation work in subsequent chapters by identifying realistic missing data patterns and variables that may be important in explaining missing data occurrence.

3.1 Objectives for this chapter

Through the investigation and analysis of three RCT datasets, this research aims to further understand the patterns of missing data in RCTs. This aim is addressed by investigating:

- The availability of PROMs scores, subscales and items at the five-year follow-up point
- Longitudinal patterns of missing data of the relevant PROMs, focussing on composite scores
- Baseline characteristics that may be predictive of missing PROMs outcome data

3.2 Introduction of the RCT datasets

Three trials are considered in this chapter; their key characteristics, where relevant to this chapter, are summarised in the following sections.

3.2.1 The KAT trial

The Knee Arthroplasty Trial (KAT) was initiated in 1999 in the UK to address uncertainties about the clinical and cost effectiveness of new developments in knee replacements. Specifically, the trial was designed to consider four aspects of knee arthroplasty, namely patellar resurfacing, mobile bearings, all-polyethylene tibial components and unicompartmental replacement.

The study was set up as a partial factorial, multicentre RCT, allowing participants to be randomised to up to two of the above mentioned comparisons. However, this chapter focuses on the patella resurfacing component of the study, with participants being randomised (on a 1:1 ratio) to either patellar resurfacing or no patellar resurfacing, ignoring any additional randomisations the participants may have undergone.

Participants were eligible to enrol in the trial if they were to undergo primary knee replacement surgery and the surgeon was in equipoise about at least one of the comparisons.

The primary outcome of the trial was the OKS, but the trial also collected data on baseline characteristics, EQ-5D-3L, SF-12, complications, hospital admissions, operation details and costs. Outcome data were collected post-operatively at three months, then annually.

1715 participants were randomised to either patellar resurfacing or no patellar resurfacing. In the context of this chapter, follow-up data up to five years post operation are considered, although the trial is continuing to follow participants beyond this point. The

trial protocol and outcomes have been published in the peer-reviewed literature[91-93]. Relevant information from the case record forms (CRFs) used for data collection can be seen in Appendix 4.

3.2.2 The PD MED trial

PD Med is a pragmatic RCT which recruited participants between 2000 and 2009 to assess the relative clinical and cost-effectiveness of different classes of drugs for Parkinson's disease (PD).

Early stage Parkinson's randomisation:

Participants recently diagnosed with PD were randomised to receive either levadopa alone (LD), dopamine agonists (DA ± LD) or monoamine oxidase type B inhibitors (MAOBI ± LD), i.e. levodopa could be added as required to the DA and MAOBI treatment arms by the investigators, particularly if symptoms were not controlled by the standard dose of MAOBI or the maximum dose of DA.

Either MAOBI or LD could be omitted from the randomisation if considered inappropriate treatment for a particular participant.

Late stage Parkinson's randomisation:

Participants who had developed motor complications uncontrolled by LD were randomised to DA, MAOBI or catechol-O-methyltransferase inhibitors (COMTI), with the option of omitting either MAOBI or DA if they were currently treated with these classes of drugs.

During each randomisation, the chances of being allocated to any of the relevant interventions were equal (i.e. 1:1 or 1:1:1 allocation ratios). Eligibility was based on "uncertainty" about their best treatment option, i.e. patients were eligible for either the early or late stage randomisation if the treating clinician was not aware of a clear indication for, or contraindication against, a particular class of PD drug included in one of the above comparisons.

A total of 1620 participants were entered into the trial via the early stage randomisation, and 500 via the late stage randomisation. Participants from the early stage randomisation were encouraged to be randomised again into the later stage randomisation when symptoms were no longer controlled by the allocated class of drugs.

The PDQ-39 was the primary outcome measure. In addition, data was collected on patient demographics at baseline, functional status and QoL (including the EQ-5D-3L), side effects, health economics, disease status (assessed by Hoehn & Yahr stage) and carer well-being. Participants were assessed at baseline, six months and one year post randomisation, then yearly to five years and every two years thereafter until at least year 10.

For this chapter, five-year follow-up data from the early stage randomisations, specifically from participants randomised of LD vs. LD sparing interventions (DA and MAOBI), is used. This excludes the 214 participants who were randomised only between DA and MAOBI, leaving data for 1406 participants, 528 of whom were randomised to LD class drugs and 878 to LD sparing classes of drugs (525 to DA, 353 to MAOBI).

Trial outcomes have been published in the peer-reviewed literature[94], the protocol can be found online[95], and relevant CRFs are shown in Appendix 5.

3.2.3 The PD SURG trial

PD SURG is a randomised, long-term assessment of the relative effectiveness of surgery in combination with best medical care compared to best medical care alone for patients with advanced PD. Between 2000 and 2006, 366 participants were randomised on a 1:1 basis to either of the two interventions.

Inclusion criteria were a diagnosis of Parkinson's disease (UK Brain Bank criteria), an age adjusted dementia rating scale-II score of greater than 5 and fitness for surgery.

The primary outcome of the trial was the PDQ-39, which was collected at baseline, annually until year three and every two years thereafter until year nine. Secondary outcomes included QoL (Unified PD Rating Scale, SF-36 (for carers only), EQ-5D-3L), disease status (Hoehn & Yahr Staging System, Neuropsychological evaluation), dementia progression and resource utilisation and costs. Within the remit of this project, only follow-up data to the five year assessment is considered.

Trial outcomes have been published in the peer-reviewed literature[96], the protocol can be found online[97], and relevant CRFs are shown in Appendix 6.

3.3 Availability of PROMs over the five year follow-up period

In the preceding chapters, missing data were defined as information that was intended to be collected within the remit of the study, considered relevant for the statistical analysis and interpretation of the results, but was unavailable at the time of the analysis. Little et al[17] also define missing data as unavailable values that 'would have been meaningful for the analysis if they were observed'. Therefore, it is argued that quality of life values missing due to participant death should not be considered as missing data within this chapter, as these measures cannot be interpreted easily. For this reason, the remainder of this chapter considers the subset of participants who were alive at five years post randomisation.

3.3.1 KAT trial: missing data patterns

Of the 1715 participants randomised to the patellar comparison in the KAT trial, 189 were known to have died by the five-year follow-up (11.02% of the trial population). Death rates were similar between the treatment arms, with 96 reported deaths in the no patellar resurfacing group vs. 93 within those randomised to patellar resurfacing. These participants are excluded from all future summaries, leaving the data for 1526 participants to be analysed in the following sections.

3.3.1.1 KAT: OKS missing data patterns

Table 3-1 shows missing OKS data, i.e. participants for whom the OKS cannot be calculated, at the different assessment time points considered in this chapter. As discussed previously, according to its scoring manual[42], the OKS cannot be calculated when more than two items have been left unanswered.

The amount of missing data increases steadily from 5.50% of participants without a valid OKS at baseline to 16.51% of participants without a valid OKS at the five year assessment, with a slightly higher amount of missing data at the three month assessment compared to the one-year assessment. Rates of missing data are similar in both treatment arms.

Table 3-1: Missing OKS composite scores by treatment arm

	OKS composite score cannot be calculated/ is missing		
Time point	No patellar resurfacing (N = 758)	Patella resurfacing (N = 768)	Total (N = 1526)
Baseline	38 (5.01%)	46 (5.99%)	84 (5.50%)
3 months	77 (10.16%)	98 (12.76%)	175 (11.47%)
1 year	80 (10.55%)	84 (10.94%)	164 (10.75%)
2 years	95 (12.53%)	116 (15.1%)	211 (13.83%)
3 years	105 (13.85%)	109 (14.19%)	214 (14.02%)
4 years	109 (14.38%)	113 (14.71%)	222 (14.55%)
5 years	130 (17.15%)	122 (15.89%)	252 (16.51%)

Figure 3-1 provides details on the individual items that have not been completed within each questionnaire and at each time point. It can be seen that item 7 (“During the past 4 weeks could you kneel down and get up again afterwards?”) is missing more frequently than other items. This may be partly due to some patients not being comfortable kneeling, or having been advised not to kneel after their knee replacement, although the question is phrased in terms of ability rather than reporting actual incidences of kneeling. Because of the possibility of missing data arising from participants misunderstanding this question, data was collected on whether or not participants of the KAT trial had been advised against kneeling. However, this data collection was sporadic, with inconsistent proportions of participants being asked this question at each follow-up time point (approximately 61% at

year two, 48% and 24% at years one and three respectively, 5% in year four; with no data available at baseline, three months and five years). As this data is not available for all participants at all time points, it is not taken into account in the subsequent analyses.

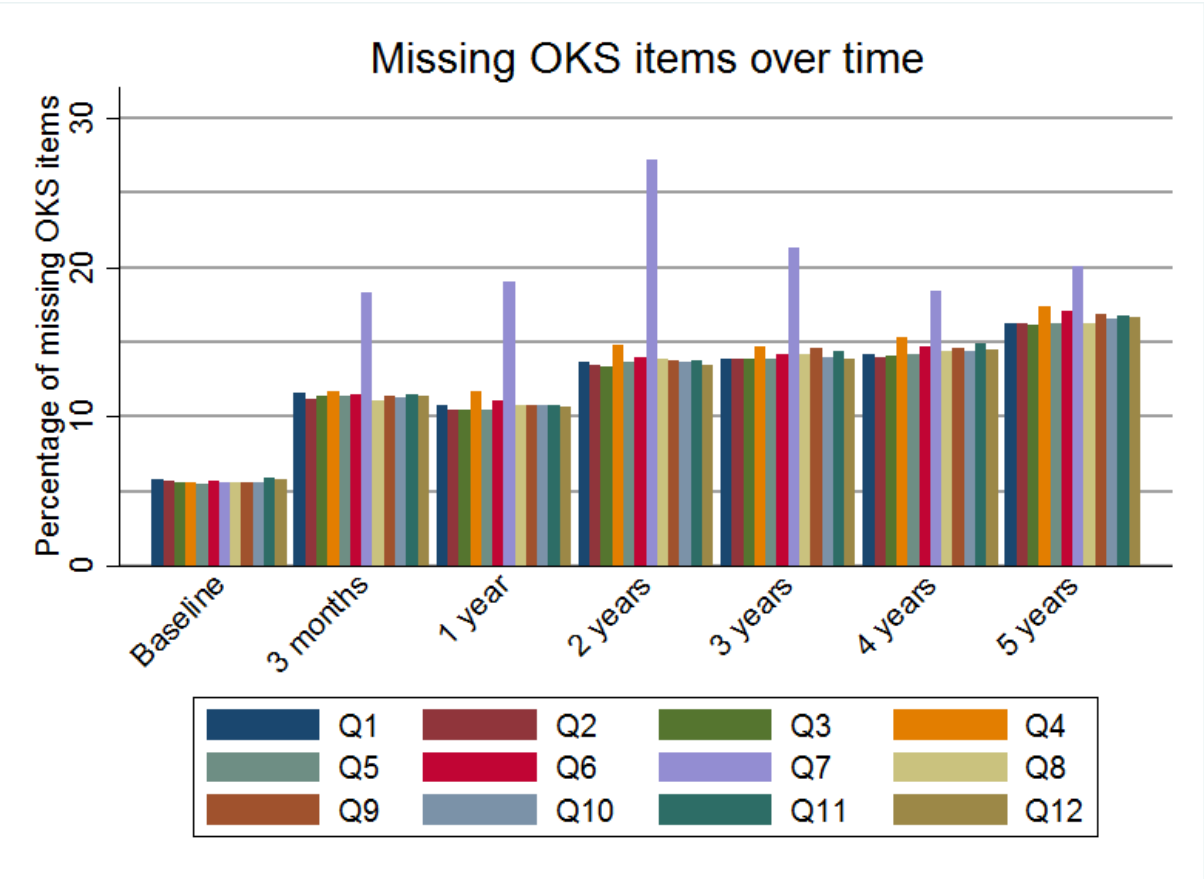


Figure 3-1: Missing OKS items over time

Investigating cases where the OKS scores are not available, it is noted that this is primarily due to non-response, where none of the 12 OKS items have been answered. This applies to all of the baseline cases for which the OKS is missing, and over 94% of the follow-up assessments.

As reiterated above, the OKS can be calculated for up to two missing items. Where the OKS can be calculated, data is usually available for all items, although this figure ranges from almost 98% at baseline to 80.84% at two years post randomisation. Where only one item is missing, usually item 7 has been left unanswered. The remaining items are missing with

very similar probabilities, except perhaps item 4 (“During the past 4 weeks, for how long have you been able to walk before the pain from your [..] knee becomes severe?”), which tends to be missing slightly more frequently than the other items (except item 7) at the one to five year follow-up visits. There are very few occasions where two items have been left unanswered. No particular combinations of missing items stood out in a further investigation into the pattern of missing items at the follow-up assessments.

The most prominent response patterns for the OKS items at the five year follow-up are shown in Table 3-2. Missing data patterns summarised under “other” were each identified in less than 1% of the data.

Table 3-2: Most prominent response patterns for the OKS items at the five year follow-up

	OKS composite scores cannot be calculated/ are missing		
Missing data patterns	No patellar resurfacing (N = 758)	Patella resurfacing (N = 768)	Total (N = 1526)
All 12 items completed	593 (78.23%)	570 (74.22%)	1163 (76.21%)
All 12 items missing	118 (15.57%)	126 (16.41%)	244 (15.99%)
Item 7 missing	27 (3.56%)	25 (3.26%)	52 (3.41%)
Other combination of missing items	20 (3.64%)	47 (6.12%)	47 (4.39%)

Figure 3-2 displays the information on the percentage of missing OKS items over assessment time point split by randomisation allocation. The data presented are in line with the overall OKS missingness, as discussed above.

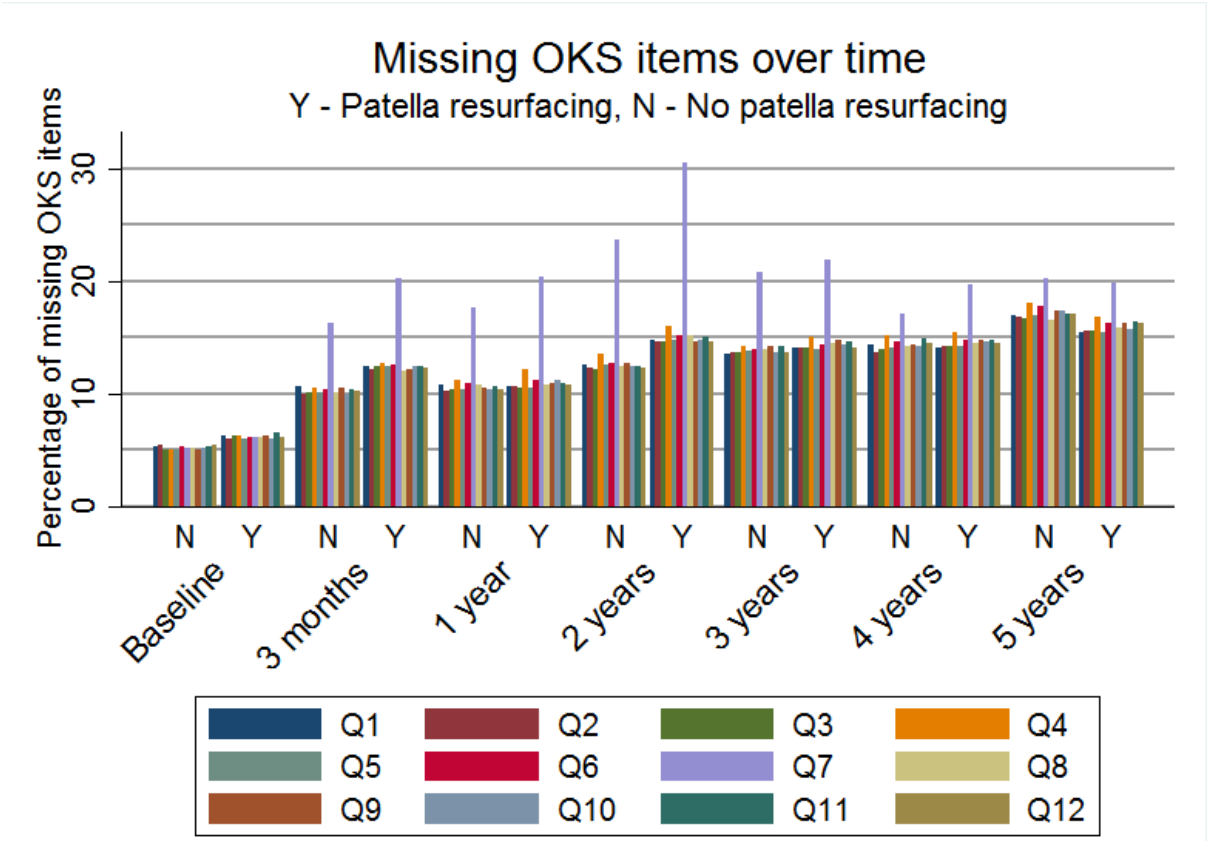


Figure 3-2: Missing OKS items over time by treatment arm

As shown above, 16.51% of OKS scores are missing at the five year follow-up time point. However, looking at the longitudinal missing data patterns, provided in Table 3-3, only 2.16% of participants have no valid OKS data at all, and an additional 1.90% of participants have provided only baseline but no follow-up OKS data. Therefore, while only 69.20% of participants have provided OKS data at all relevant time points, some post-randomisation data is available for the vast majority of the trial population. The missing data for 10.75% of the participants can be classed as monotone missing data patterns, with data available up to a certain point after which all data are missing. For 17.89% of the participants, data are missing intermittently, with missing assessments being followed by subsequent available data[20, 98, 99]. Over 12% of the participants show intermittent missing data patterns with a maximum of two missed assessments. The fact that there is some outcome data

available for almost all participants could be very important for analyses taking into account the longitudinal structure of the outcomes, or when referring to earlier observations in the context of multiple imputation (see Chapter 5). More information on the longitudinal patterns of missingness is provided in Table 3-3, and a graphical representation can be seen in Figure 3-3.

Table 3-3: OKS composite scores - longitudinal missingness patterns

	No patellar resurfacing (N = 758)	Patella resurfacing (N = 768)	Total (N = 1526)
OKS can be calculated at all time points	531 (70.05%)	525 (68.36%)	1056 (69.20%)
Complete missingness – no OKS data available	16 (2.11%)	17 (2.21%)	33 (2.16%)
Monotone missingness			
Data available for baseline only	15 (1.98%)	14 (1.82%)	29 (1.90%)
Data available until three months	7 (0.92%)	7 (0.91%)	14 (0.92%)
Data available until one year	16 (2.11%)	17 (2.21%)	33 (2.16%)
Data available until two years	15 (1.98%)	10 (1.3%)	25 (1.64%)
Data available until three years	9 (1.19%)	10 (1.3%)	19 (1.25%)
Data available until four years	23 (3.03%)	21 (2.73%)	44 (2.88%)
Total (monotone missingness)	85 (11.21%)	79 (10.29%)	164 (10.75%)
Intermittent missingness			
OKS missing intermittently at one time point	69 (9.1%)	64 (8.33%)	133 (8.72%)
OKS missing intermittently at two time points	19 (2.51%)	34 (4.43%)	53 (3.47%)
OKS missing intermittently at three time points	21 (2.77%)	28 (3.65%)	49 (3.21%)
OKS missing intermittently at four time points	9 (1.19%)	11 (1.43%)	20 (1.31%)
OKS missing intermittently at five time points	7 (0.92%)	9 (1.17%)	16 (1.05%)
OKS missing intermittently at six time points	1 (0.13%)	1 (0.13%)	2 (0.13%)
Total (intermittent missingness)	126 (16.62%)	147 (19.14%)	273 (17.89%)
Total	758 (100%)	768 (100%)	1526 (100%)

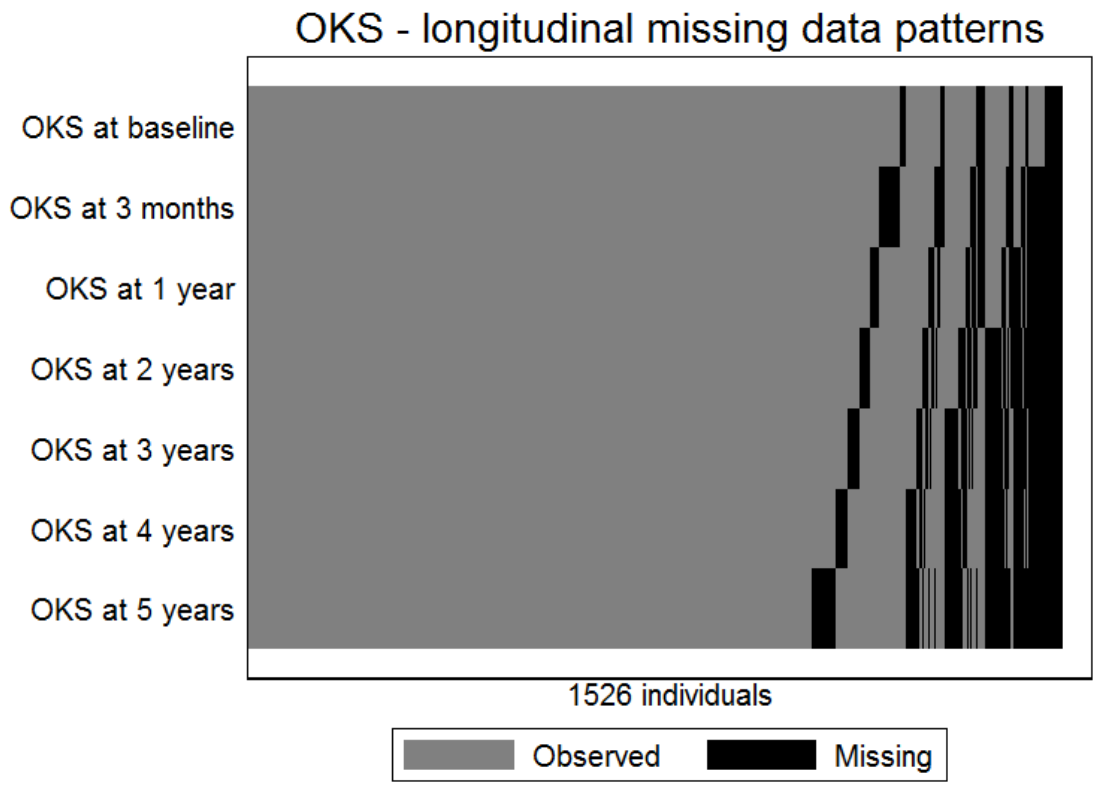


Figure 3-3: OKS longitudinal missingness patterns - graphical representation (KAT)

3.3.1.2 KAT: EQ-5D-3L missing data patterns

Table 3-4 shows the frequency and percentage of participants for whom the EQ-5D-3L cannot be calculated at the different assessment time points. As discussed previously, according to its scoring manual[46], the EQ-5D-3L composite score[i] cannot be calculated when any one item has been left unanswered. The amount of missing data increased from 7.14% of participants without a valid EQ-5D-3L composite score at baseline to 18.02% of participants at the five year assessment. The figures are slightly higher than those observed for the OKS, again with a slight increase in missing data at the three month assessment compared to the one-year assessment. As before, there are no pronounced differences in the amounts of missing data between the treatment arms.

Table 3-4: Missing EQ-5D-3L composite scores by treatment arm

	EQ-5D-3L composite score cannot be calculated/ is missing		
Time point	No patellar resurfacing (N = 758)	Patella resurfacing (N = 768)	Total (N = 1526)
Baseline	44 (5.8%)	65 (8.46%)	109 (7.14%)
3 months	92 (12.14%)	108 (14.06%)	200 (13.11%)
1 year	94 (12.4%)	100 (13.02%)	194 (12.71%)
2 years	108 (14.25%)	133 (17.32%)	241 (15.79%)
3 years	117 (15.44%)	130 (16.93%)	247 (16.19%)
4 years	128 (16.89%)	129 (16.80%)	257 (16.84%)
5 years	147 (19.39%)	128 (16.67%)	275 (18.02%)

Figure 3-4 provides more detail on the individual items that have not been completed within each questionnaire and at each time point. Item 5, the question on anxiety and depression, seems to be missing marginally more frequently than the others, possibly due

[i] Instead of items and composite scores, the terms domain and index are commonly used in EQ-5D-3L research. However, for consistency, the former terminology is used throughout the chapter.

to participants being less comfortable with this question. Again, where the EQ-5D-3L is missing, this is predominantly due to unit non-response, i.e. 77% of missing EQ-5D-3L composite scores at baseline, and between 81 and 87% at follow-up are due to unit-nonresponse. The remaining cases are missing mainly due to one question having been left unanswered; the probability of item 5 (anxiety and depression) being missing is only slightly higher than the probability of missing data for the other items; combinations of more than one item being omitted are observed on isolated occasions only.

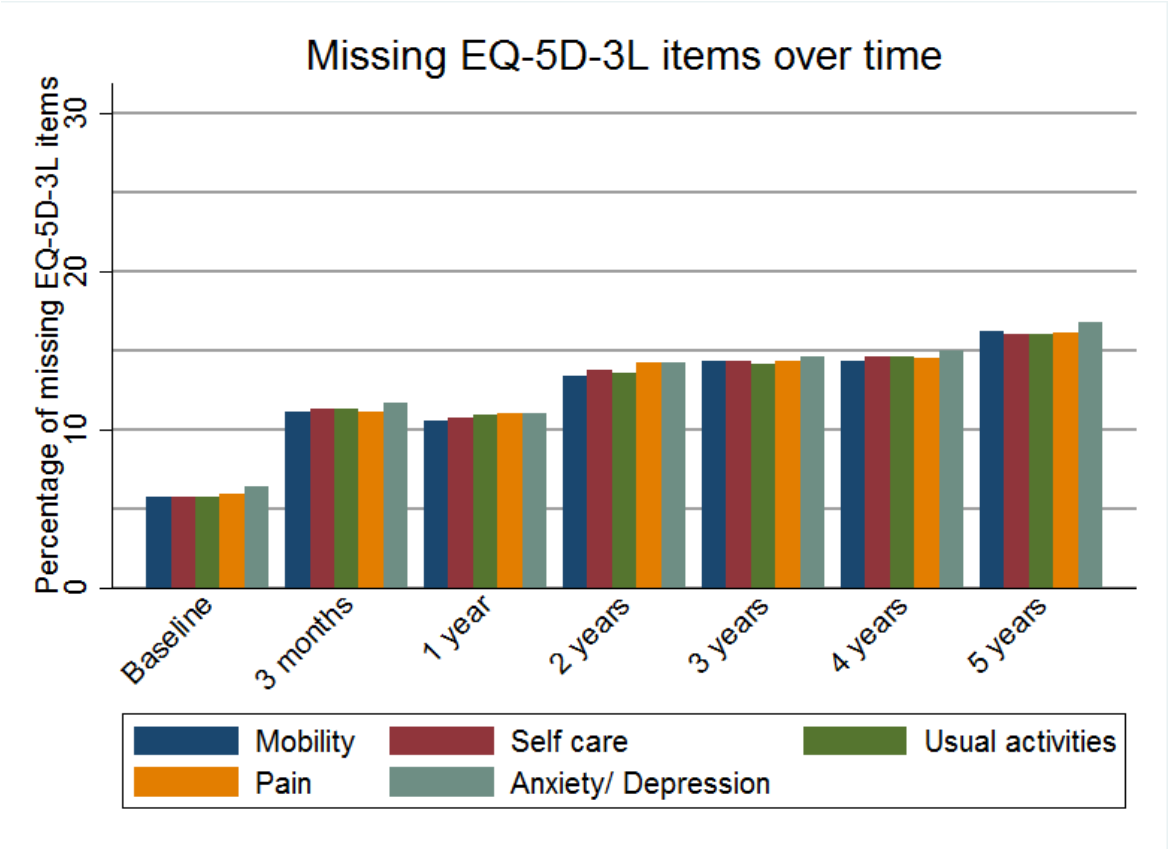


Figure 3-4: Missing EQ-5D-3L items over time

Figure 3-5 displays the same information by randomisation allocation; overall, rates and patterns of missingness are similar in both trial arms.

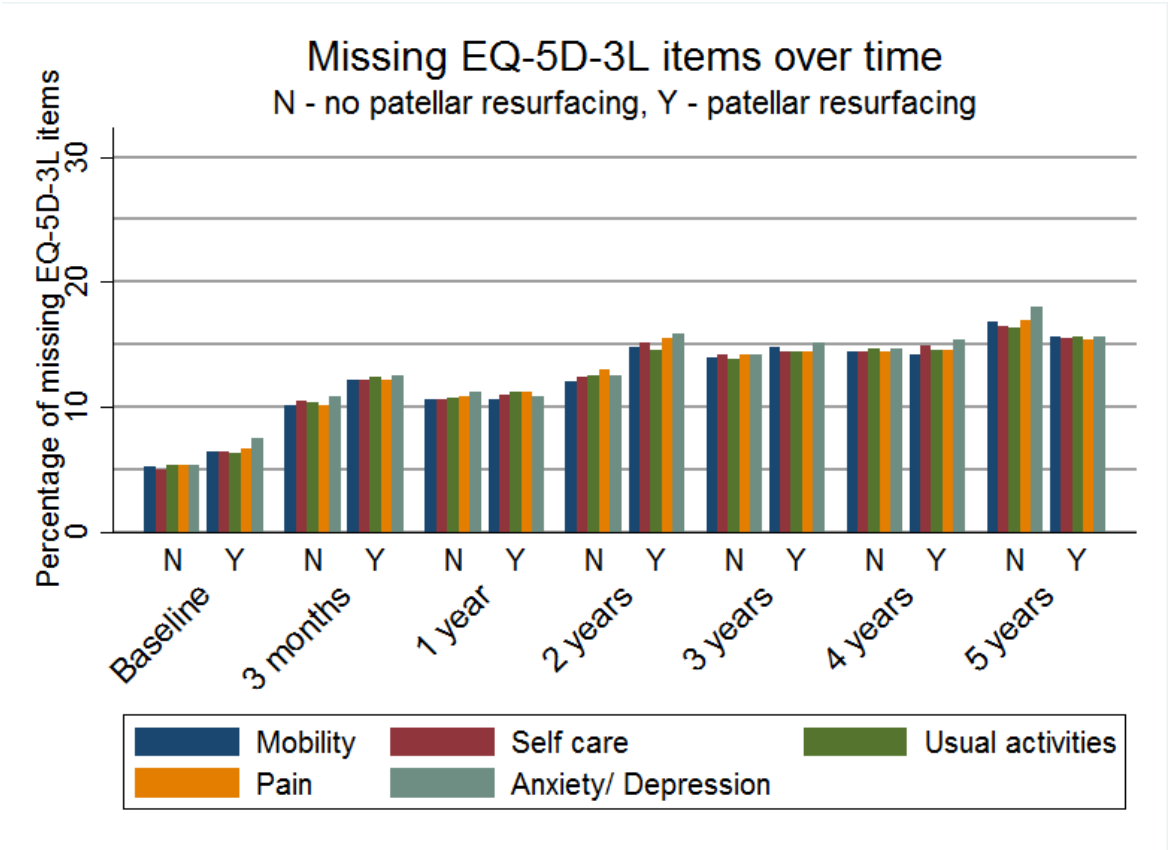


Figure 3-5: Missing EQ-5D-3L items over time by treatment arm

Details of the longitudinal missing data patterns are provided in Table 3-5. Only 2.16% of all participants have no observed EQ-5D-3L data, and an additional 1.97% of participants have provided only baseline but no follow-up EQ-5D-3L data. While only 61.80% of participants have provided valid EQ-5D-3L data at all relevant time points, some post-randomisation data is available for the vast majority of the trial population. The missing data for 10.94% of the participants can be classed as monotone missing data patterns. For 25.10% of the participants, data is missing intermittently. Over 17% of the participants show intermittent missing data patterns with a maximum of two missed assessments; more information on the longitudinal patterns of missingness is provided in Figure 3-6.

Table 3-5: EQ-5D-3L composite scores - longitudinal missingness patterns

	No patellar resurfacing (N = 758)	Patella resurfacing (N = 768)	Total (N = 1526)
EQ-5D-3L can be calculated at all time points	473 (62.4%)	470 (61.2%)	943 (61.80%)
Complete missingness – no EQ-5D-3L data available	16 (2.11%)	17 (2.21%)	33 (2.16%)
Monotone missingness			
Data available for baseline only	15 (1.98%)	15 (1.95%)	30 (1.97%)
Data available until three months	10 (1.32%)	6 (0.78%)	16 (1.05%)
Data available until one year	14 (1.85%)	17 (2.21%)	31 (2.03%)
Data available until two years	13 (1.72%)	11 (1.43%)	24 (1.57%)
Data available until three years	7 (0.92%)	7 (0.91%)	14 (0.92%)
Data available until four years	34 (4.49%)	18 (2.34%)	52 (3.41%)
Total (monotone missingness)	93 (12.27%)	74 (9.64%)	167 (10.94%)
Intermittent missingness			
EQ-5D-3L missing intermittently at one time point	96 (12.66%)	97 (12.63%)	193 (12.65%)
EQ-5D-3L missing intermittently at two time points	35 (4.62%)	46 (5.99%)	81 (5.31%)
EQ-5D-3L missing intermittently at three time points	23 (3.03%)	37 (4.82%)	60 (3.93%)
EQ-5D-3Lmissing intermittently at four time points	12 (1.58%)	15 (1.95%)	27 (1.77%)
EQ-5D-3L missing intermittently at five time points	8 (1.06%)	11 (1.43%)	19 (1.25%)
EQ-5D-3L missing intermittently at six time points	2 (0.26%)	1 (0.13%)	3 (0.2%)
Total (intermittent missingness)	176 (23.22%)	207 (26.95%)	383 (25.10%)
Total	758 (100%)	768 (100%)	1526 (100%)

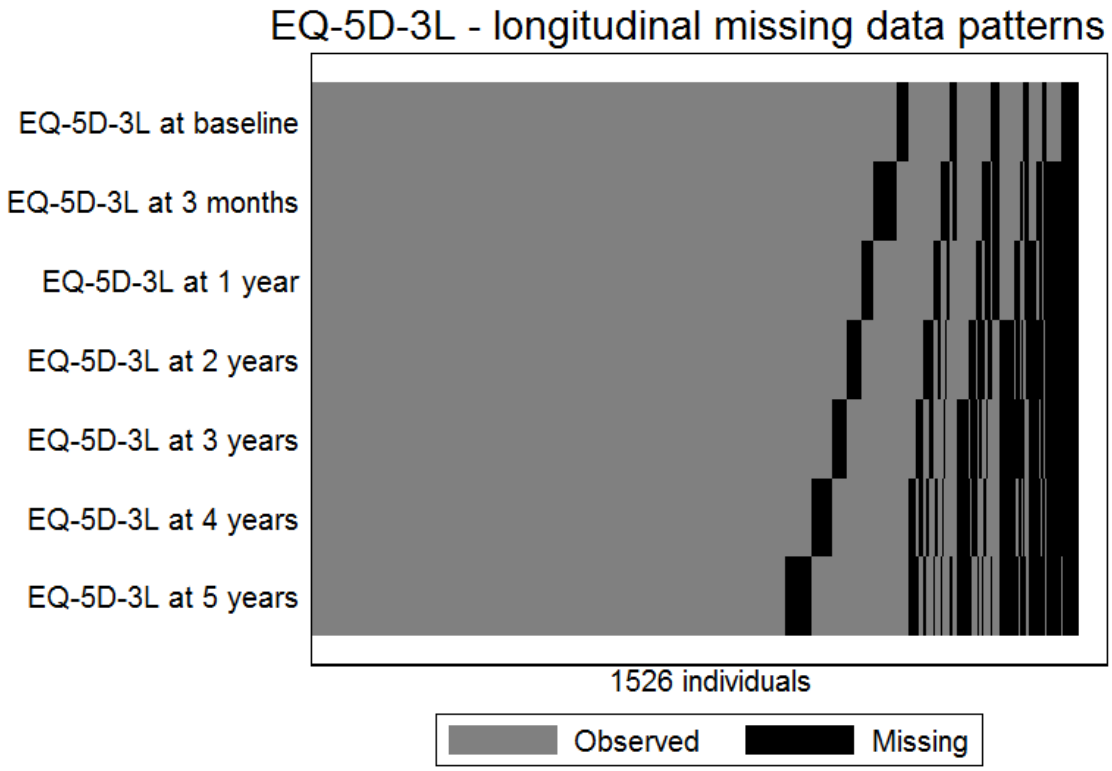


Figure 3-6: EQ-5D-3L composite scores - longitudinal missingness patterns - graphical representation (KAT)

3.3.1.3 KAT: SF-12 version 2 missing data patterns

In the above sections, 1526 participants randomised either to patellar resurfacing or no patellar resurfacing were included in the summaries. Further participants need to be excluded from the missing data summaries for the SF-12. During the recruitment and follow-up of the KAT trial, version 2 of the SF-12 was used for the majority of participants. However, early on in the trial, some participants completed version 1 of the SF-12 at baseline and/or three month follow-up. Participants who completed version 1 instead of version 2 at any point of the trial are excluded from the following summaries and analyses. This is because some of the items, and the categories within the items, have changed from version 1 to 2, and not all conclusions and results from the following sections may be transferrable between the versions. For this reason, an additional 104 participants (50 in the no patellar resurfacing arm and 54 in the patellar resurfacing arm) are excluded from the subsequent summaries considering the SF-12 within the KAT trial, leaving a total of 1422 participants to be analysed, 708 of those in the no patellar resurfacing group and 714 in the patellar resurfacing group.

Table 3-6 shows the frequency and percentage of participants for whom the SF-12 Mental Health Component Summary (MCS) score and the Physical Health Component Summary (PCS) score cannot be calculated at the different assessment time points. As discussed previously, according to its original scoring manual[43, 90], the SF-12 subscales[ii] cannot be calculated when any one item has been left unanswered, although algorithms that can handle missing data have been proposed[43, 78]. For the purposes of this work, only the

[ii] For consistency, the MCS and PCS are referred to as SF-12 subscales in this chapter.

original scoring manual are considered. The MCS and PCS scores have the same patterns of missingness, as all items contribute to both subscales using different weights.

Table 3-6 shows that the percentage of participants with missing SF-12 subscales increases from 16.53% at baseline to 33.33% at the five year assessment. The figures are similar in both trial arms, again with a slightly higher amount of missing data at the three month assessment compared to the one-year assessment; the rates of missing data are higher than those observed both for the OKS and EQ-5D-3L.

Table 3-6: Missing SF-12 subscales by treatment arm

	SF-12 subscales cannot be calculated/ is missing		
Time point	No patellar resurfacing (N = 708)	Patella resurfacing (N = 714)	Total (N = 1422)
Baseline	114 (16.1%)	121 (16.95%)	235 (16.53%)
3 month	193 (27.26%)	191 (26.75%)	384 (27%)
1 year	192 (27.12%)	192 (26.89%)	384 (27%)
2 years	211 (29.8%)	194 (27.17%)	405 (28.48%)
3 years	220 (31.07%)	223 (31.23%)	443 (31.15%)
4 years	233 (32.91%)	221 (30.95%)	454 (31.93%)
5 years	244 (34.46%)	230 (32.21%)	474 (33.33%)

Figure 3-7 provides further detail on the individual items that have not been completed within each questionnaire and at each time point. It can be observed that the percentages of items being missing are lower than the percentages of missing SF-12 subscales. Across the different assessment time points, on average approximately 70% of the participants have provided answers to all SF-12 items (i.e. the subscales can be calculated). Between 10 and 15% of the subscales are missing due to unit nonresponse. The remaining 15 to 20% of SF-12 subscales are missing due to different combinations of items being missing. Item

2b (current health state limits climbing several flights of stairs), as well as 3b (physical health limits the kind of work and other activities) and 4b (emotional problems result in work and other activities not being performed as carefully as usual) are missing with higher frequency than other items across the follow-up. Other patterns of missing items are less frequent. Item missingness is more prominent in the SF-12 version 2 in the KAT trial than in the OKS and EQ-5D-3L.

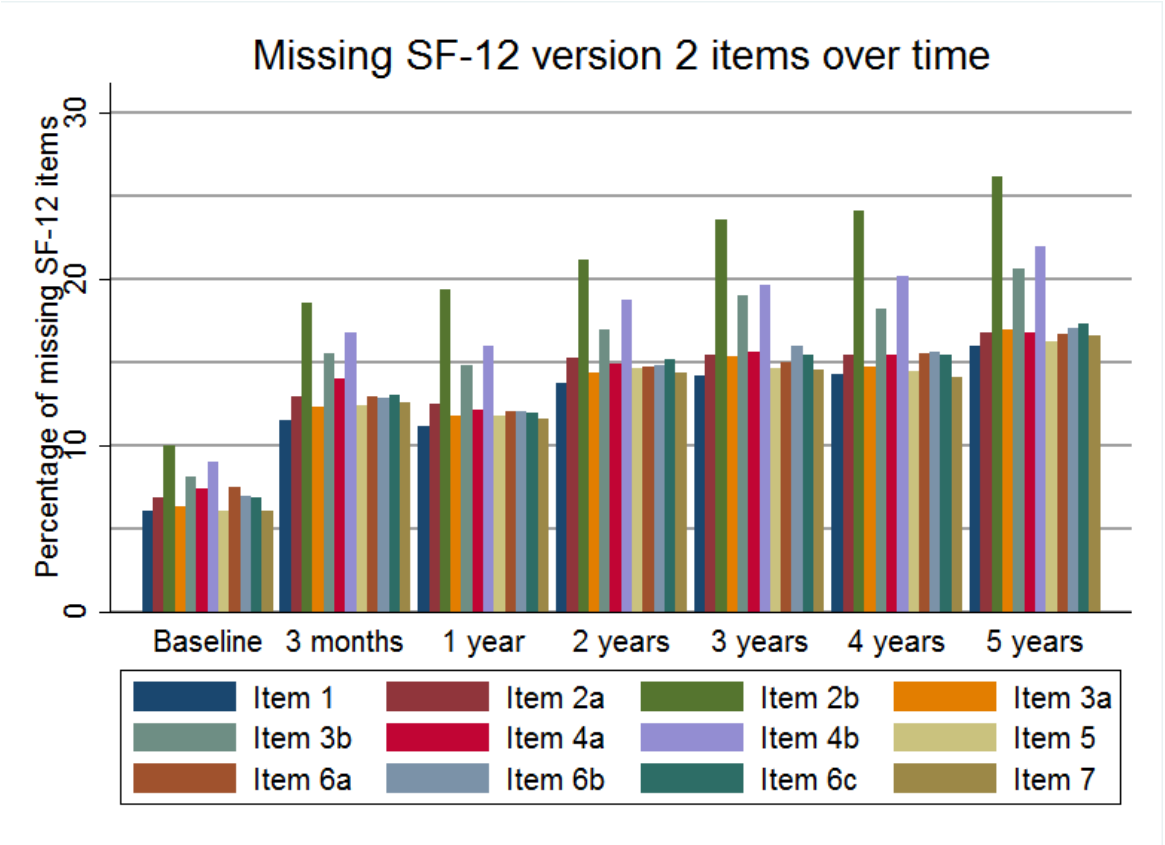


Figure 3-7: Missing SF-12 version 2 items over time

Figure 3-8 displays information on missing items by randomisation allocation. Overall, rates and patterns of missingness are similar in both trial arms.

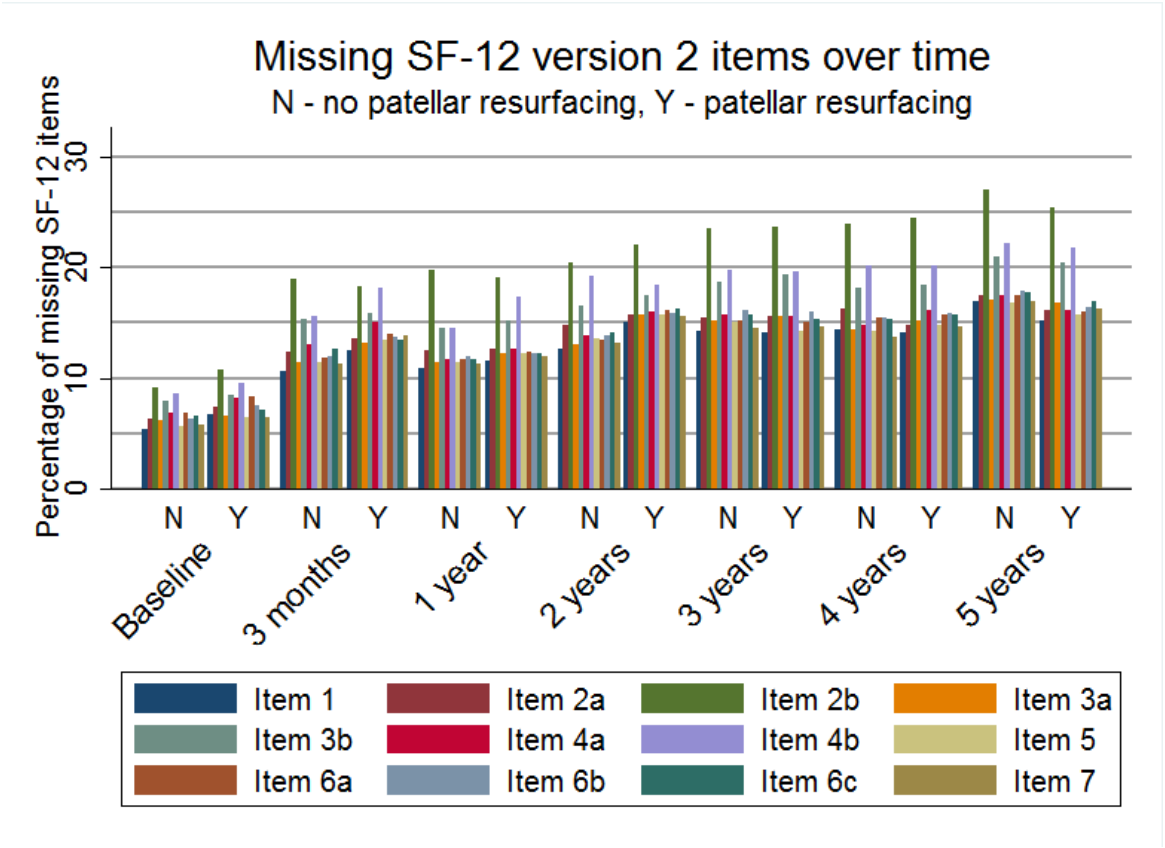


Figure 3-8: Missing SF-12 items over time by treatment arm

The longitudinal missing data patterns are considered in Table 3-7. 3.16% of all participants have missing SF-12 subscales at all time points, and an additional 3.31% of participants have provided only baseline but no follow-up data. While only 30.24% of participants have provided valid SF-12 subscale data at all relevant time points, some post-randomisation data is available for the vast majority of the trial population. The missing data for 13.34% of the participants can be classed as monotone missing data patterns. For 53.16% of the participants, data are missing intermittently. Over 30% of the participants show an intermittent missing data patterns with a maximum of two missed assessments.

Table 3-7: SF-12 version 2 - longitudinal missingness patterns

	No patellar resurfacing (N = 708)	Patella resurfacing (N = 714)	Total (N = 1422)
SF-12 can be calculated at all time points	196 (27.68%)	234 (32.77%)	430 (30.24%)
Complete missingness – no SF-12 data available	22 (3.11%)	23 (3.22%)	45 (3.16%)
Monotone missingness			
Data available for baseline only	22 (3.11%)	25 (3.5%)	47 (3.31%)
Data available until three months	6 (0.85%)	12 (1.68%)	18 (1.27%)
Data available until one year	18 (2.54%)	12 (1.68%)	30 (2.11%)
Data available until two years	13 (1.84%)	7 (0.98%)	20 (1.41%)
Data available until three years	9 (1.27%)	14 (1.96%)	23 (1.62%)
Data available until four years	30 (4.24%)	23 (3.22%)	53 (3.73%)
Total (monotone missingness)	98 (13.84%)	93 (13.03%)	191 (13.43%)
Intermittent missingness			
SF-12 missing intermittently at one time point	147 (20.76%)	120 (16.81%)	267 (18.78%)
SF-12 missing intermittently at two time points	93 (13.14%)	94 (13.17%)	187 (13.15%)
SF-12 missing intermittently at three time points	65 (9.18%)	68 (9.52%)	133 (9.35%)
SF-12 missing intermittently at four time points	43 (6.07%)	46 (6.44%)	89 (6.26%)
SF-12 missing intermittently at five time points	32 (4.52%)	31 (4.34%)	63 (4.43%)
SF-12 missing intermittently at six time points	12 (1.69%)	5 (0.7%)	17 (1.2%)
Total (intermittent missingness)	392 (55.37%)	364 (50.98%)	756 (53.16%)
Total	708 (100%)	714 (100%)	1422 (100%)

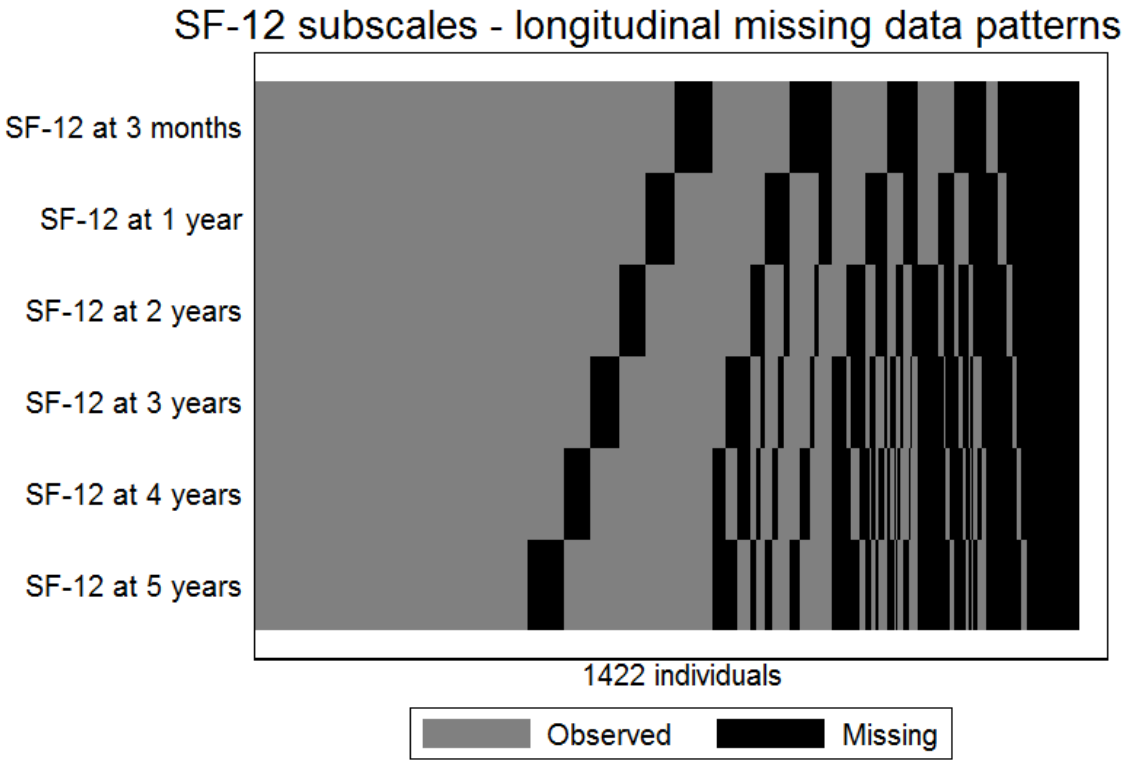


Figure 3-9: SF-12 longitudinal missingness patterns - graphical representation (KAT)

3.3.2 PD MED trial: missing data patterns

Within PD MED, 293 participants (20.84%) were confirmed to have died before their five year assessment (18.94% of those randomised to LD, and 21.98% of those randomised to LD sparing medications). As discussed previously, these participants are excluded from the further analyses; 1113 participants are included in the following summaries.

3.3.2.1 PD MED: PDQ-39 missing data patterns

According to its scoring manual[48], the PDQ-39 subscales and composite scores (PDQ-39-SI)[iii] cannot be calculated if any items have been left unanswered, with the exception that participants without a partner are not expected to answer item 28 (Due to having Parkinson's disease, how often during the last month have you lacked support in the ways you need from your spouse or partner?). For pragmatic reasons, the social support subscale is still calculated for anyone with missing answer for item 28, provided that the other relevant items have valid, non-missing responses. The use of an expectation maximisation algorithm has been proposed for the imputation of missing dimension scores[100]; however, these imputation techniques are not considered within this chapter.

The data in Table 3-8 shows missing PDQ-39 data, i.e. percentage of participants for whom the PDQ-39 composite score cannot be calculated, is increasing steadily from 11.23% at baseline to 37.11% at the five year assessment. There are no pronounced differences in the amounts of missing data between the treatment arms, although missing data rates in the LD arm are marginally lower until year four.

[iii] The PDQ-39 composite score is also referred to as the PDQ-39 single index score, or short PDQ-39-SI. However, throughout this chapter, the terminology composite score is used for consistency.

Table 3-8: Missing PDQ-39 by treatment arm (PD MED)

Time point	LD (N=428)	LD sparing (N=685)	Total (N=1113)
Baseline	48 (11.21%)	77 (11.24%)	125 (11.23%)
6 months	62 (14.49%)	109 (15.91%)	171 (15.36%)
1 year	70 (16.36%)	126 (18.39%)	196 (17.61%)
2 years	88 (20.56%)	149 (21.75%)	237 (21.29%)
3 years	104 (24.3%)	167 (24.38%)	271 (24.35%)
4 years	131 (30.61%)	203 (29.64%)	334 (30.01%)
5 years	146 (34.11%)	267 (38.98%)	413 (37.11%)

Dividing data availability into the different subscales (Figure 3-10) shows the set of questions on activities of daily living and on emotional wellbeing are slightly more likely to be missing than the other subscales.

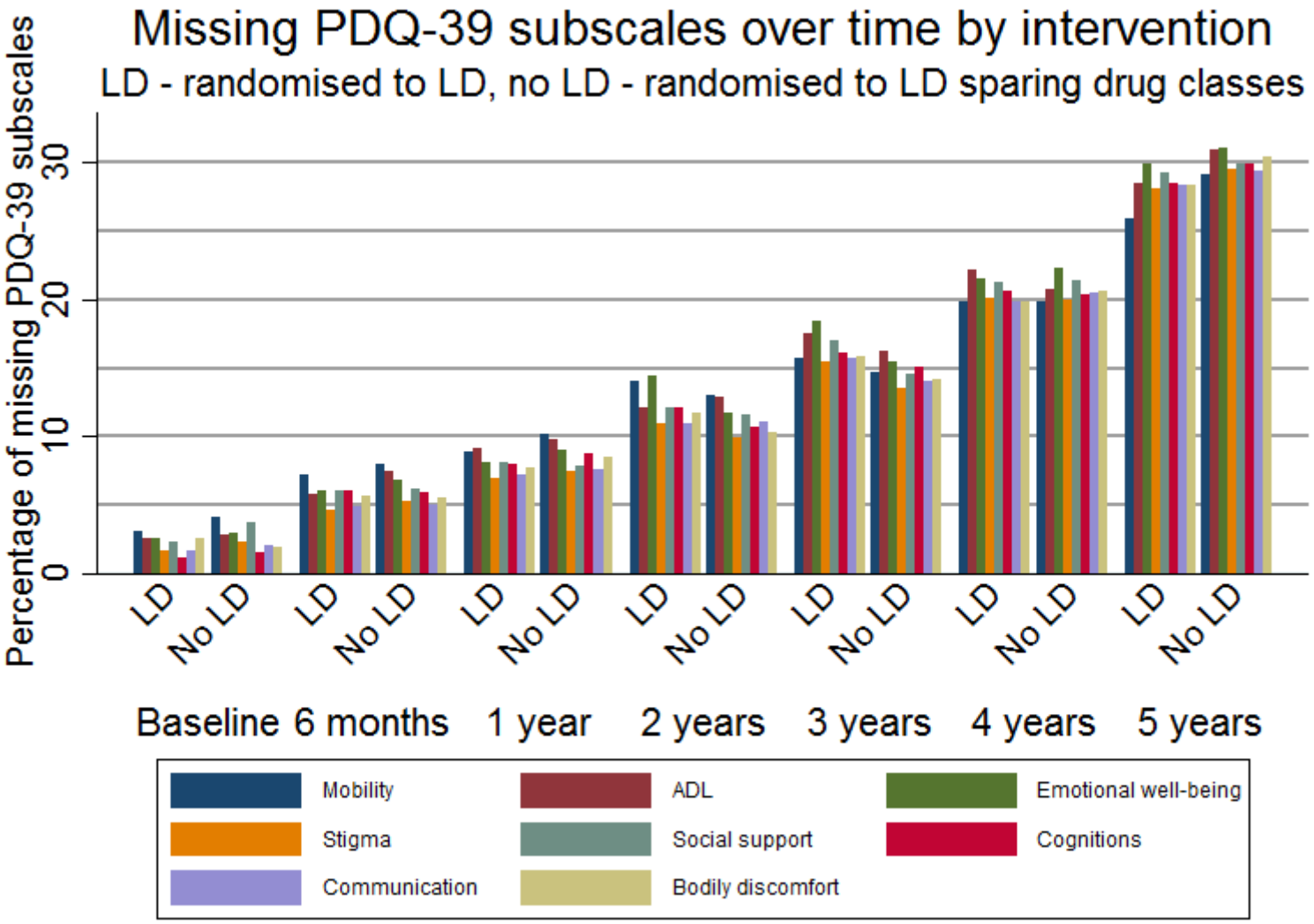


Figure 3-10: Missing PDQ-39 subscales by treatment arm over time (PD MED)

Table 3-9 presents the longitudinal missing data patterns of the PDQ-39-SI in the PD MED trial. It can be seen that only 35.04% of all participants have valid PDQ-39 data for all assessment time points. Less than 1% of participants have missing data for the PDQ-39-SI across all assessments. Approximately 20% of the trial population follows monotone missing data patterns, while a total of around 45% have intermitting missing data, though the majority of those (approximately 30% of all participants) have missing PDQ-39-SI data for one or two assessments only. The information on longitudinal missing data patterns is shown graphically in Figure 3-11.

Table 3-9: PDQ-39-SI - longitudinal missingness patterns (PD MED)

	LD (N = 428)	LD sparing (N = 685)	Total (N = 1113)
PDQ-39-SI can be calculated at all time points	160 (37.38%)	230 (33.58%)	390 (35.04%)
Complete missingness – no PDQ-39-SI data available	5 (1.17%)	3 (0.44%)	8 (0.72%)
Monotone missingness			
Data available for baseline only	6 (1.4%)	17 (2.48%)	23 (2.07%)
Data available until six months	6 (1.4%)	9 (1.31%)	15 (1.35%)
Data available until one year	7 (1.64%)	17 (2.48%)	24 (2.16%)
Data available until two years	14 (3.27%)	18 (2.63%)	32 (2.88%)
Data available until three years	19 (4.44%)	20 (2.92%)	39 (3.5%)
Data available until four years	27 (6.31%)	57 (8.32%)	84 (7.55%)
Total (monotone missingness)	79 (18.46%)	138 (20.15%)	217 (19.5%)
Intermittent missingness			
PDQ-39-SI missing intermittently at one time point	76 (17.76%)	122 (17.81%)	198 (17.79%)
PDQ-39-SI missing intermittently at two time points	48 (11.21%)	85 (12.41%)	133 (11.95%)
PDQ-39-SI missing intermittently at three time points	24 (5.61%)	41 (5.99%)	65 (5.84%)
PDQ-39-SI missing intermittently at four time points	19 (4.44%)	39 (5.69%)	58 (5.21%)
PDQ-39-SI missing intermittently at five time points	9 (2.1%)	22 (3.21%)	31 (2.79%)
PDQ-39-SI missing intermittently at six time points	8 (1.87%)	5 (0.73%)	13 (1.17%)
Total (intermittent missingness)	184 (42.99%)	314 (45.84%)	498 (44.74%)
Total	428 (100%)	685 (100%)	1113 (100%)

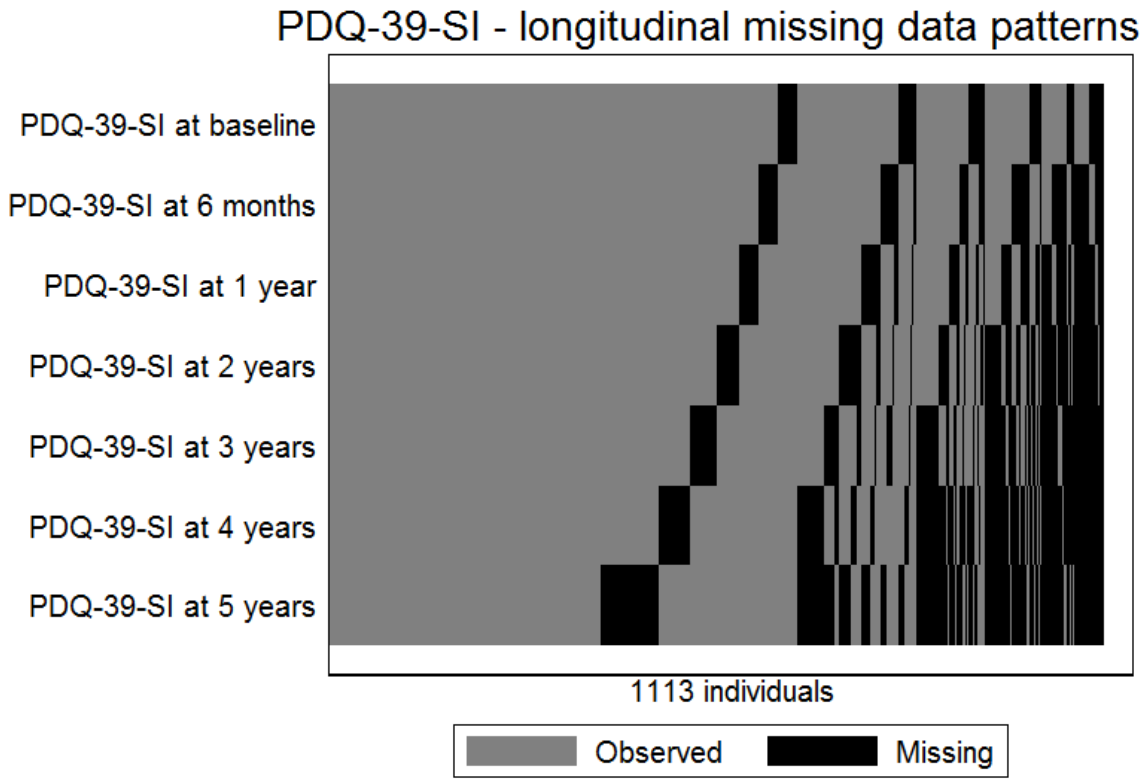


Figure 3-11: PDQ-39-SI longitudinal missingness patterns - graphical representation (PD MED)

3.3.2.2 PD MED: EQ-5D-3L missing data patterns

The data in Table 3-10 shows missing EQ-5D-3L data, i.e. percentage of patients for whom the EQ-5D-3L composite scores cannot be calculated, which is increasing steadily from 1.71% at baseline to 26.33% at the five year assessment. As mentioned previously, according to its scoring manual[48], the EQ-5D-3L composite score cannot be calculated if any items have been left unanswered. There are no pronounced differences in the missing data rates between the treatment arms. However, it can be seen that the rates of missing EQ-5D-3L composite scores are around 10% lower than those for the PDQ-39-SI at each time point.

Table 3-10: Missing EQ-5D-3L composite scores by treatment arm (PD MED)

Time point	LD (N=428)	LD sparing (N=685)	Total (N=1113)
Baseline	8 (1.87%)	11 (1.61%)	19 (1.71%)
6 months	22 (5.14%)	34 (4.96%)	56 (5.03%)
1 year	33 (7.71%)	58 (8.47%)	91 (8.18%)
2 years	47 (10.98%)	66 (9.64%)	113 (10.15%)
3 years	61 (14.25%)	85 (12.41%)	146 (13.12%)
4 years	75 (17.52%)	121 (17.66%)	196 (17.61%)
5 years	111 (25.93%)	182 (26.57%)	293 (26.33%)

Figure 3-12 considers the availability of the individual items at each assessment time point. Rates of missingness are similar for all items within the different follow-up time points; the graph reflects the higher probability of data being missing for participants randomised to LD at two and three years, as also observed in Table 3.10.

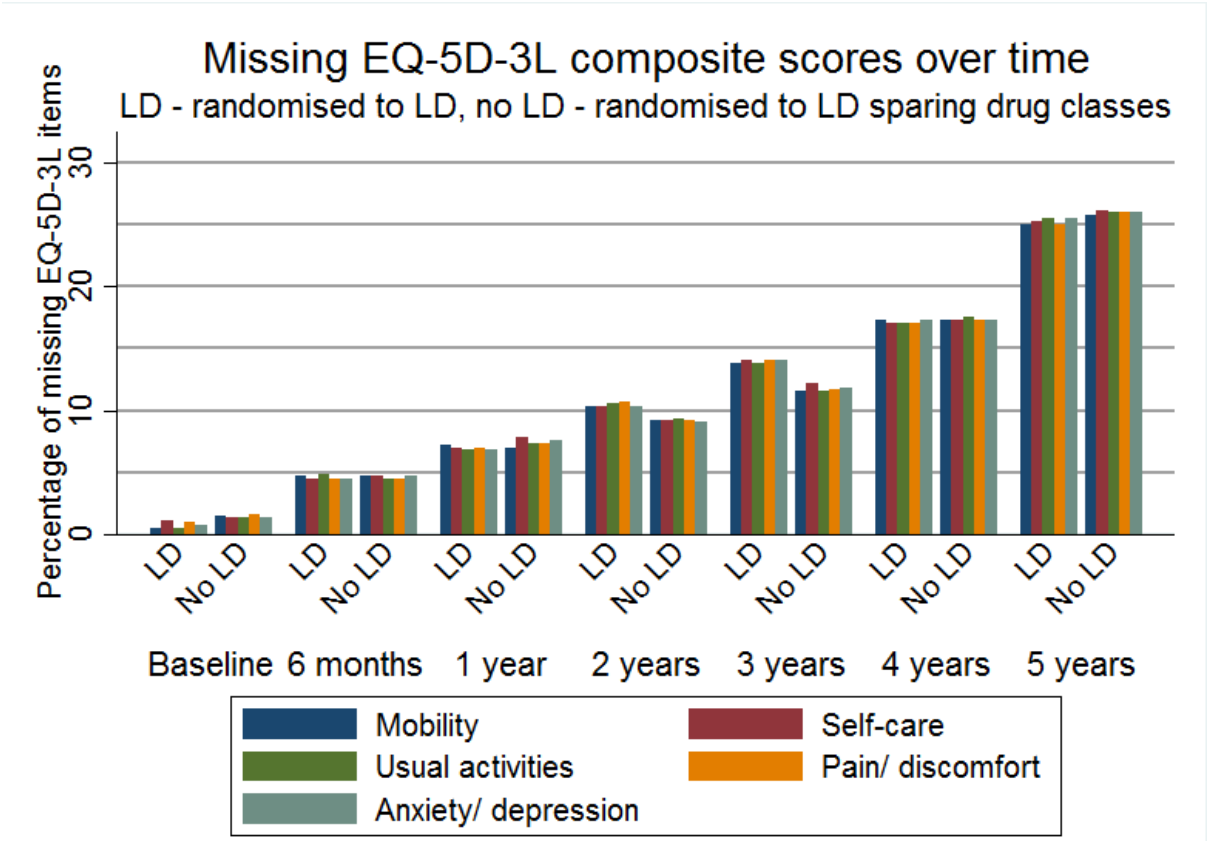


Figure 3-12: Missing EQ-5D-3L items by treatment arm over time (PD MED)

Table 3-11 presents the longitudinal missing data patterns of the EQ-5D-3L composite scores in the PD MED trial. It can be seen that almost 64% of participants have valid EQ-5D-3L composite scores for all assessment time points within the initial five years of follow-up, nearly twice as many as those with valid PDQ-39-SI at all follow-up time points. Only one participant has missing EQ-5D-3L composite scores across all assessments. Approximately 21% of the trial population follow monotone missing data patterns, while a total of around 15% have intermittently missing data, which is also much lower than the figure observed for the PDQ-39 SI. The majority of those participants (almost 12% of all participants) have missing EQ-5D-3L data for one or two assessments only. These longitudinal missing data patterns are displayed graphically in Figure 3-13.

Table 3-11: EQ-5D-3L composite score longitudinal missingness patterns (PD MED)

	LD (N = 428)	LD sparing (N = 685)	Total (N = 1113)
EQ-5D-3L can be calculated at all time points	268 (62.62%)	439 (64.09%)	707 (63.52%)
Complete missingness – no EQ-5D-3L available	1 (0.23%)	0 (0%)	1 (0.09%)
Monotone missingness			
Data available for baseline only	7 (1.64%)	10 (1.46%)	17 (1.53%)
Data available until six months	9 (2.1%)	11 (1.61%)	20 (1.8%)
Data available until one year	9 (2.1%)	16 (2.34%)	25 (2.25%)
Data available until two years	12 (2.8%)	15 (2.19%)	27 (2.43%)
Data available until three years	15 (3.5%)	27 (3.94%)	42 (3.77%)
Data available until four years	40 (9.35%)	62 (9.05%)	102 (9.16%)
Total (monotone missingness)	92 (21.5%)	141 (20.58%)	233 (20.93%)
Intermittent missingness			
EQ-5D-3L missing intermittently at one time point	36 (8.41%)	50 (7.3%)	86 (7.73%)
EQ-5D-3L missing intermittently at two time points	19 (4.44%)	25 (3.65%)	44 (3.95%)
EQ-5D-3L missing intermittently at three time points	6 (1.4%)	10 (1.46%)	16 (1.44%)
EQ-5D-3L missing intermittently at four time points	2 (0.47%)	13 (1.9%)	15 (1.35%)
EQ-5D-3L missing intermittently at five time points	3 (0.7%)	7 (1.02%)	10 (0.9%)
EQ-5D-3L missing intermittently at six time points	1 (0.23%)	0 (0%)	1 (0.09%)
Total (intermittent missingness)	67 (15.65%)	105 (15.33%)	172 (15.45%)
Total	428 (100%)	685 (100%)	1113 (100%)

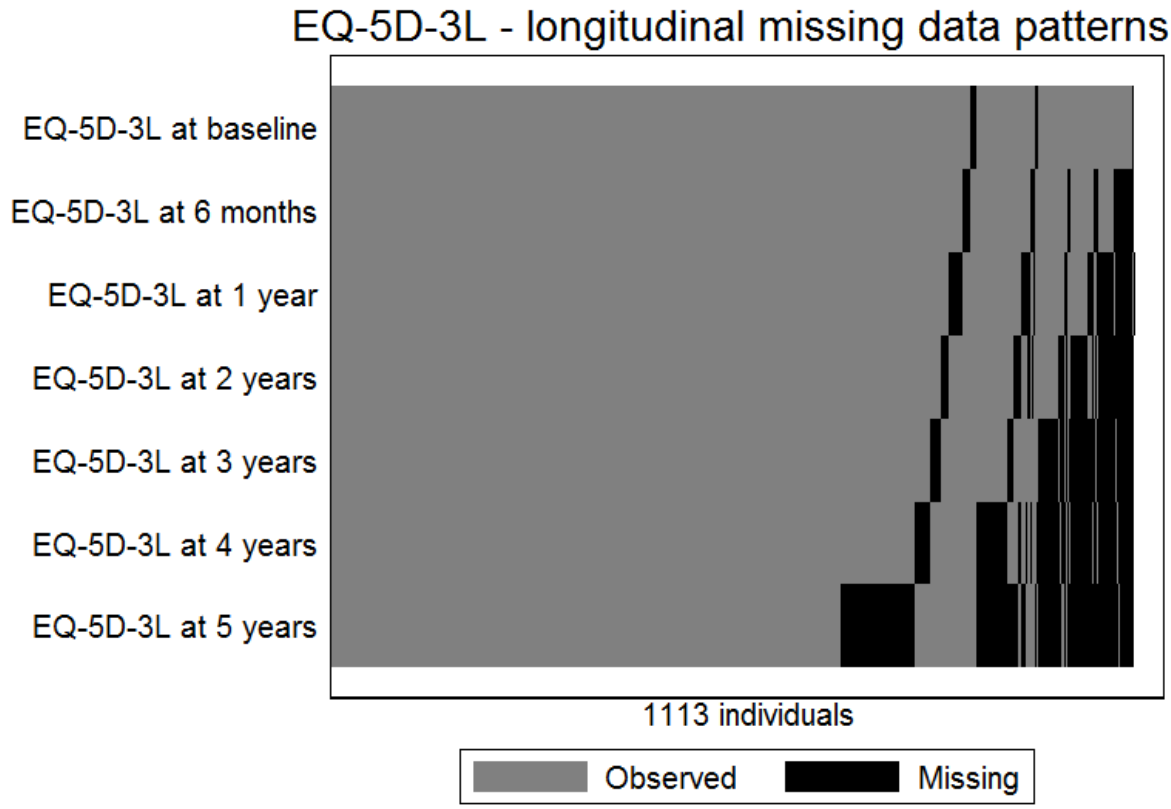


Figure 3-13: EQ-5D-3L composite scores longitudinal missingness patterns - graphical representation (PD MED)

3.3.3 PD SURG trial: missing data patterns

As discussed previously, participants are excluded from the subsequent analyses if their death was confirmed prior to the five-year follow-up. Therefore, 41 of the 366 participants (11.20%) are excluded (11.48% of those randomised to drugs, and 10.93% of those randomised to surgery), and 325 participants are included in the following summaries.

3.3.3.1 PD SURG: PDQ-39 missing data patterns

Table 3-12 shows that the amount of missing PDQ-39-SI in the PD SURG trial increases from around 15% of scores being unavailable at baseline to almost 52% of participants having missing data at the five year follow-up. At baseline, more missing data is observed in the group randomised to surgery, while at three years those randomised to drugs have more missing PDQ-39-SI data. However, missing data levels are similar at the other follow-up time points.

Table 3-12: Missing PDQ-39-SI by treatment arm (PD SURG)

Time point	Randomised to drugs (N = 162)	Randomised to surgery (N = 163)	Total (N = 325)
Baseline	20 (12.35%)	30 (18.4%)	50 (15.38%)
1 year	41 (25.31%)	35 (21.47%)	76 (23.38%)
2 years	46 (28.40%)	45 (27.61%)	91 (28.00%)
3 years	68 (41.98%)	60 (36.81%)	128 (39.38%)
5 years	85 (52.47%)	83 (50.92%)	168 (51.69%)

Figure 3-14 breaks up the data availability into the eight PDQ-39 subscales by treatment arm. The amount of missing data for the subscales appears to be lower than the amount of missing data for the PDQ-39-SI score. This difference can be explained by the PDQ

scoring algorithm, which does not allow for the composite score to be calculated if information for any one of the subscales is missing. For example, around 48% of all participants have information available for all subscales at the five year follow-up. For 43% of participants, information for all subscales is missing, leaving 9% of participants with the information for often only one or two subscales missing. Similarly, at one, two and three years post randomisation, 10%, 11% and 10% of participants respectively have valid data for some but not all PDQ-39 subscales, meaning that the PDQ-39-SI cannot be calculated.

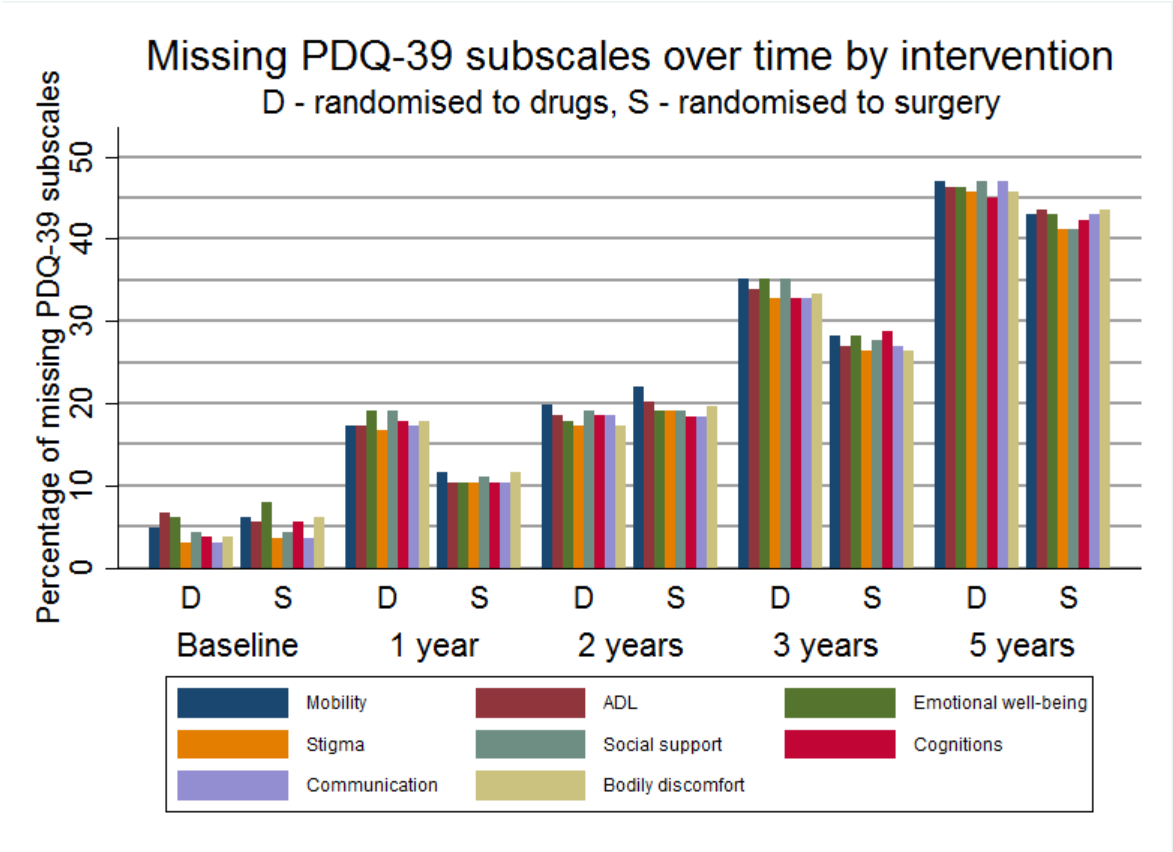


Figure 3-14: Missing PDQ-39 subscales over time by randomised treatment allocation

Table 3-13 shows longitudinal missing data patterns for thePDQ-39-SI. Only around 24% of the total population have valid data for the PDQ-39-SI for all time points. The percentage of participants with no PDQ-39-SI data at all is negligible with only 3 participants (less than 1% of the sample) falling into this category. Around 35% of the sample have monotone

missing data patterns, and approximately 41% intermittent missing data patterns. Figure 3-15 displays the longitudinal missing data patterns graphically.

Table 3-13: PDQ-39-SI - longitudinal missingness patterns (PD SURG)

	Drugs (N = 162)	Surgery (N = 163)	Total (N = 325)
PDQ-39-SI can be calculated at all time points	39 (24.07%)	38 (23.31%)	77 (23.69%)
Complete missingness – no PDQ-39-SI data available	2 (1.23%)	1 (0.61%)	3 (0.92%)
Monotone missingness			
Data available for baseline only	11 (6.79%)	8 (4.91%)	19 (5.85%)
Data available until one year	7 (4.32%)	5 (3.07%)	12 (3.69%)
Data available until two years	14 (8.64%)	19 (11.66%)	33 (10.15%)
Data available until three years	26 (16.05%)	23 (14.11%)	49 (15.08%)
Total (monotone missingness)	58 (35.8%)	55 (33.74%)	113 (34.77%)
Intermittent missingness			
PDQ-39-SI missing intermittently at one time point	24 (14.81%)	27 (16.56%)	51 (15.69%)
PDQ-39-SI missing intermittently at two time points	14 (8.64%)	22 (13.5%)	36 (11.08%)
PDQ-39-SI missing intermittently at three time points	21 (12.96%)	11 (6.75%)	32 (9.85%)
PDQ-39-SI missing intermittently at four time points	4 (2.47%)	9 (5.52%)	13 (4%)
Total (intermittent missingness)	63 (38.89%)	69 (42.33%)	132 (40.62%)
Total	162 (100%)	163 (100%)	325 (100%)

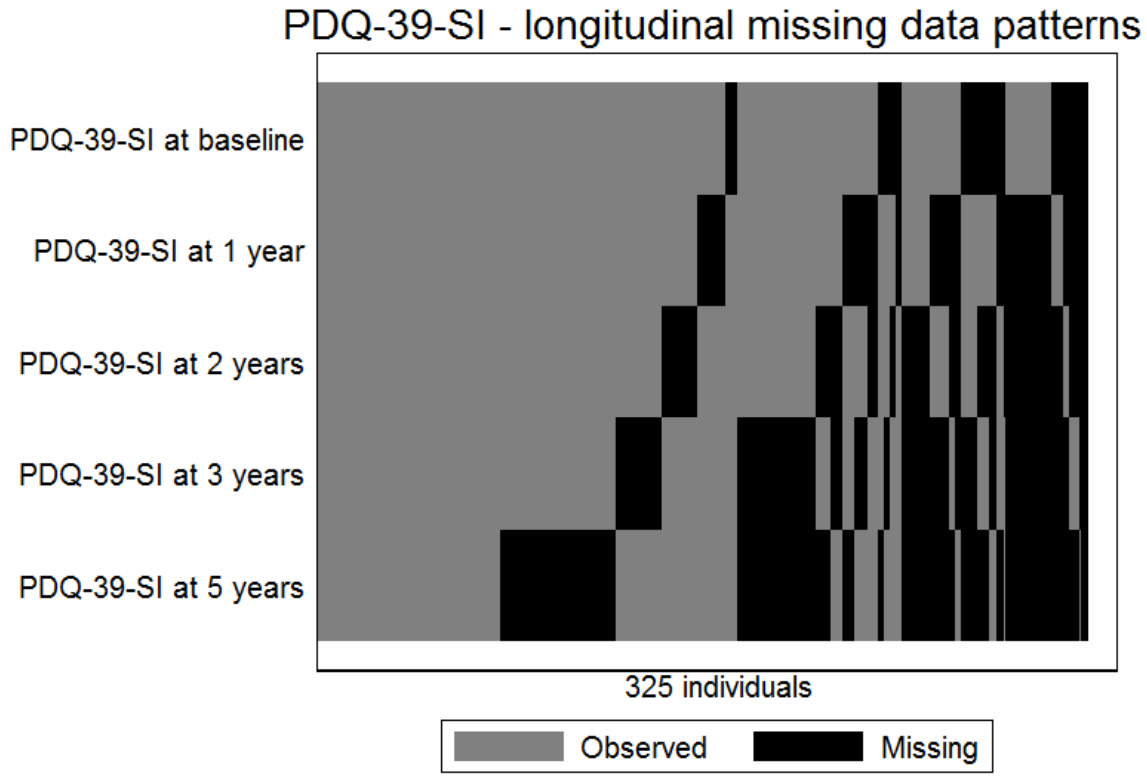


Figure 3-15: PDQ-39-SI longitudinal missingness patterns - graphical representation (PD SURG)

3.3.3.2 PD SURG: EQ-5D-3L missing data patterns

Unlike the PDQ-39, the EQ-5D-3L was not measured at the two year follow-up. Therefore, the following summaries consider missing data at baseline, as well as at the one, three and five year follow-up points. The amount of missing EQ-5D-3L data increases from approximately 4% of scores being unavailable at baseline to just over 43% missing data at the five year follow-up, as shown in Table 3-14. At baseline, there is marginally more missing data in the group randomised to surgery, while at one and three years more missing EQ-5D-3L data is observed in those randomised to drugs. Missing data levels are similar at the five year follow-up time point.

Table 3-14: Missing EQ-5D-3L composite scores by treatment arm (PD SURG)

Time point	Randomised to drugs (N = 162)	Randomised to surgery (N = 163)	Total (N = 325)
Baseline	6 (3.7%)	8 (4.91%)	14 (4.31%)
1 year	27 (16.67%)	17 (10.43%)	44 (13.54%)
3 years	53 (32.72%)	46 (28.22%)	99 (30.46%)
5 years	69 (42.59%)	71 (43.56%)	140 (43.08%)

Figure 3-16 breaks up the data availability into the five EQ-5D-3L items by treatment arm. The amount of missing data for the items is similar, albeit a little lower than the amount of missing data for the EQ-5D composite scores.

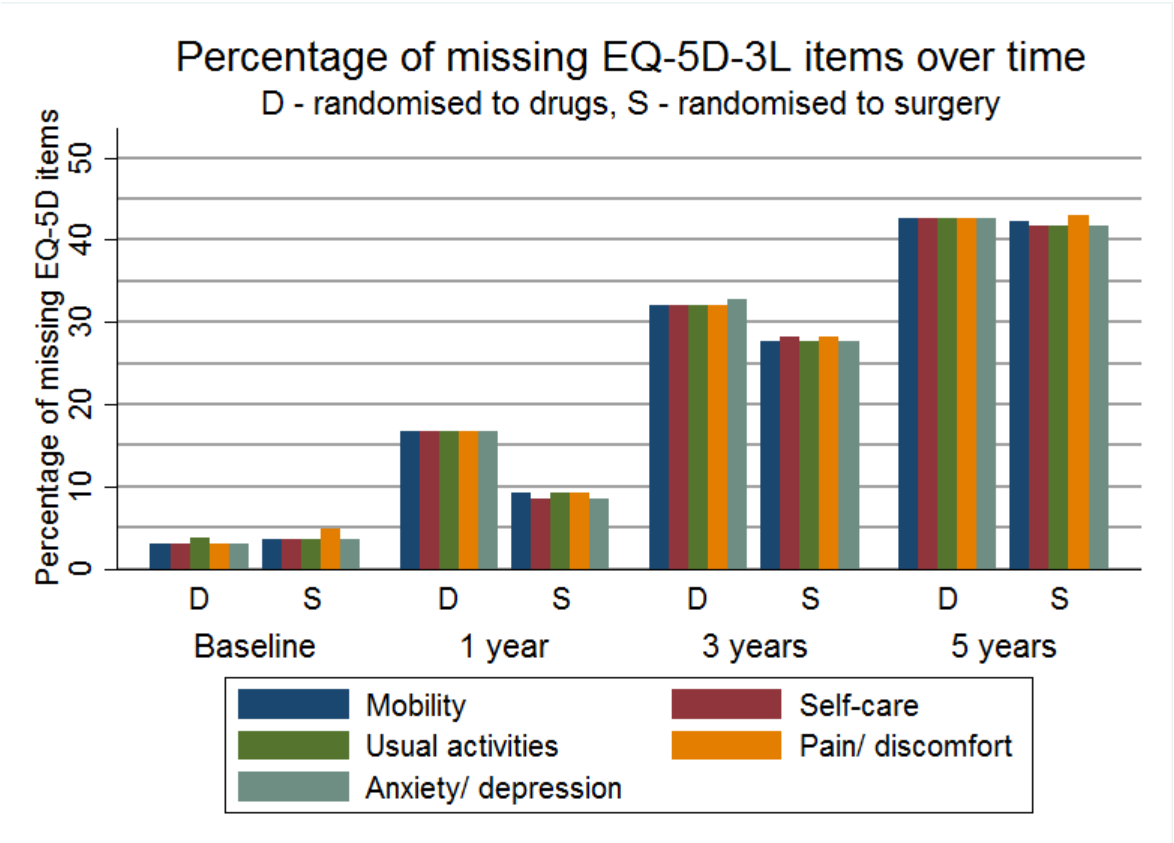


Figure 3-16: Missing EQ-5D-3L items over time by randomised treatment allocation

Table 3-15 shows longitudinal missing data patterns for the EQ-5D-3L composite scores. Around 44% of the total population have valid EQ-5D-3L data for all time points. The percentage of participants with no EQ-5D-3L data at any assessments is negligible with only two participants (less than 1% of the sample) falling into this category. Around 39% of the sample have monotone missing data patterns, and approximately 16% intermittent missing data patterns.

Table 3-15: EQ-5D-3L composite scores - longitudinal missingness patterns (PD SURG)

	Drugs (N = 162)	Surgery (N = 163)	Total (N = 325)
EQ-5D-3L can be calculated at all time points	70 (43.21%)	74 (45.4%)	144 (44.31%)
Complete missingness – no EQ-5D-3L data available	0 (0%)	2 (1.23%)	2 (0.62%)
Monotone missingness			
Data available for baseline only	15 (9.26%)	9 (5.52%)	24 (7.38%)
Data available until one year	19 (11.73%)	19 (11.66%)	38 (11.69%)
Data available until three years	29 (17.9%)	37 (22.7%)	66 (20.31%)
Total (monotone missingness)	63 (38.89%)	65 (39.88%)	128 (39.38%)
Intermittent missingness			
EQ-5D-3L missing intermittently at one time point	18 (11.11%)	15 (9.2%)	33 (10.15%)
EQ-5D-3L missing intermittently at two time points	8 (4.94%)	4 (2.45%)	12 (3.69%)
EQ-5D-3L missing intermittently at three time points	3 (1.85%)	3 (1.84%)	6 (1.85%)
Total (intermittent missingness)	29 (17.9%)	22 (13.5%)	51 (15.69%)
Total	162 (100%)	163 (100%)	325 (100%)

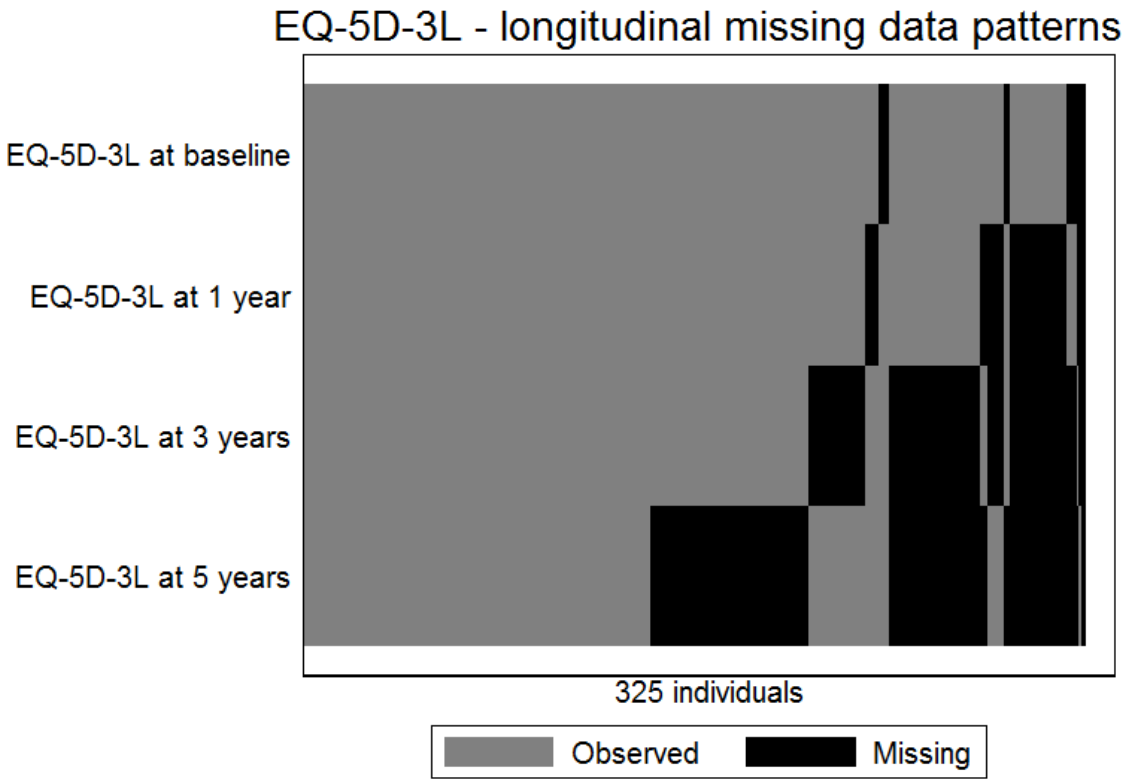


Figure 3-17: EQ-5D-3L composite scores - longitudinal missingness patterns - graphical representation (PD SURG)

3.3.4 Comparison of missing data patterns within the three RCTs

This section summarises the information on missing data presented above. Table 3-16 shows that missing data rates, as well as the change in missing data from baseline to the final follow-up, differ not only between trials, but also between the PROMs used within trials.

Focussing on longitudinal missing data (Table 3-17) and using the EQ-5D-3L as an example, the composite score can be calculated at all relevant trial assessments for approximately 62% and 64% in the KAT and PD MED trials, respectively. However, in the PD SURG trial, a valid EQ-5D-3L composite score is available at all relevant assessment time points for approximately 44% of participants. Similarly, within the KAT trial, the percentage of participants with PROMs data at all assessment time points is variable (69% for the OKS, 62% for the EQ-5D-3L and 30% for the SF-12). Likewise, the percentages of participants with intermittently missing data differs both within and between trials, while the percentages of participants with monotone missing are more similar within the different RCTs.

Table 3-16: Overview of the missing PROMs dates across RCTs

	KAT			PDMED		PD SURG	
	OKS (N = 1526)	EQ-5D-3L (N = 1526)	SF-12 (N = 1422)	PDQ-39-SI (N = 1113)	EQ-5D-3L (N = 1113)	PDQ-39-SI (N = 325)	EQ-5D-3L (N = 325)
PROM missing at baseline	5.50%	7.14%	16.53%	11.23%	1.71%	15.38%	4.31%
PROM missing at three*/six** months	11.47%	13.11%	27.00%	15.36%	5.03%	n/a	n/a
PROM missing at one year	10.75%	12.71%	27.00%	17.61%	8.18%	23.38%	13.54%
PROM missing at two years	13.83%	15.79%	28.48%	21.29%	10.15%	28.00%	n/a
PROM missing at three years	14.02%	16.19%	31.15%	24.35%	13.12%	39.38%	30.46%
PROM missing at four years	14.55%	16.84%	31.93%	30.01%	17.61%	n/a	n/a
	16.51%	18.02%	33.33%	37.11%	26.33%	51.69%	43.08%

*KAT trial; **PD MED trial

Table 3-17: Longitudinal missingness patterns for all RCTs and PROMs

	KAT			PDMED		PD SURG	
	OKS (N = 1526)	EQ-5D-3L (N = 1526)	SF-12 (N = 1422)	PDQ-39-SI (N = 1113)	EQ-5D-3L (N = 1113)	PDQ-39-SI (N = 325)	EQ-5D-3L (N = 325)
PROM can be calculated at all time points	69.20%	61.8%	30.24%	35.04%	63.52%	23.69%	44.31%
Complete missingness – no PROM data available	2.16%	2.16%	3.16%	0.72%	0.09%	0.92%	0.62%
Monotone missingness							
PROM available for baseline only	1.90%	1.97%	3.31%	2.07%	1.53%	5.85%	7.38%
PROM available until three*/six** months	0.92%	1.05%	1.27%	1.35%	1.8%	n/a	n/a
PROM available until one year	2.16%	2.03%	2.11%	2.16%	2.25%	3.69%	11.69%
PROM available until two years	1.64%	1.57%	1.41%	2.88%	2.43%	10.15%	n/a
PROM available until three years	1.25%	0.92%	1.62%	3.5%	3.77%	15.08%	20.31%
PROM available until four years	2.88%	3.41%	3.73%	7.55%	9.16%	n/a	n/a
Total (monotone missingness)	10.75%	10.94%	13.43%	19.5%	20.93%	34.77%	39.38%
Intermittent missingness							
PROM missing intermittently at one time point	8.72%	12.65%	18.78%	17.79%	7.73%	15.69%	10.15%
PROM missing intermittently at two time points	3.47%	5.31%	13.15%	11.95%	3.95%	11.08%	3.69%
PROM missing intermittently at three time points	3.21%	3.93%	9.35%	5.84%	1.44%	9.85%	1.85%
PROM missing intermittently at four time points	1.31%	1.77%	6.26%	5.21%	1.35%	4%	n/a
PROM missing intermittently at five time points	1.05%	1.25%	4.43%	2.79%	0.9%	n/a	n/a
PROM missing intermittently at six time points	0.13%	0.2%	1.2%	1.17%	0.09%	n/a	n/a
Total (intermittent missingness)	17.89%	25.1%	53.16%	44.74%	15.45%	40.62%	15.69%
Total	100%	100%	100%	100%	100%	100%	100%

*KAT trial; **PD MED trial

3.4 Predictors of missing PROMs data at five years: univariate models

3.4.1 KAT trial: univariate missing data patterns

3.4.1.1 KAT: OKS univariate missing data patterns

In an attempt to identify variables that may be predictive of missing OKS data at the five year follow-up, the trial participants were split into two categories indicating whether the OKS at five years was missing or available. The summary in Table 3-18 shows that there may be differences in the demographics between participants with and without missing data at the five year follow-up, particularly with regards to gender, type of knee arthritis, previous knee surgeries, ASA physical status (a system to assess patients' fitness prior to surgery)[101, 102] and post-operative complications.

Table 3-18: Patient characteristics split by availability of the OKS at 5 years

	OKS at 5 years missing (N = 252)	**OKS at 5 years available (N = 1274)**	**Total (N = 1526)**
Randomisation allocation patella resurfacing vs. no patella resurfacing*			
No patella resurfacing	130 (51.59%)	628 (49.29%)	758 (49.67%)
Patella resurfacing	122 (48.41%)	646 (50.71%)	768 (50.33%)
Gender*			
Female	157 (62.30%)	717 (56.28%)	874 (57.27%)
Male	95 (37.70%)	557 (43.72%)	652 (42.73%)
Weight (kg)^	78 (69, 90), (42, 150)	80 (70, 90), (43, 195)	80 (70, 90), (42, 195)
Height (cm)^	161 (155, 168), (120, 190)	165 (158, 172), (106, 199)	164 (157, 172), (106, 199)
BMI^	29.41 (26.49, 33.79), (17.04, 54.17)	29.07 (26.09, 32.42), (17.15, 77.43)	29.14 (26.12, 32.46), (17.04, 77.43)
Primary type of knee arthritis*			
Rheumatoid	10 (3.97%)	49 (3.85%)	59 (3.87%)
Osteoarthritis	212 (84.13%)	1203 (94.43%)	1415 (92.73%)
Both	1 (0.40%)	2 (0.16%)	3 (0.20%)
Missing	29 (11.51%)	20 (1.57%)	49 (3.21%)
Arthritis location*			
Single knee	54 (21.43%)	340 (26.69%)	394 (25.82%)
Both knees	92 (36.51%)	509 (39.95%)	601 (39.38%)
General	77 (30.56%)	405 (31.79%)	482 (31.59%)
Missing	29 (11.51%)	20 (1.57%)	49 (3.21%)
Previous knee surgery*			
No	151 (59.92%)	820 (64.36%)	971 (63.63%)
Yes	69 (27.38%)	432 (33.91%)	501 (32.83%)
Missing	32 (12.70%)	22 (1.73%)	54 (3.54%)
ASA physical status*			
completely fit and healthy	34 (13.49%)	244 (19.15%)	278 (18.22%)
some illness but no effect on daily activity	125 (49.60%)	776 (60.91%)	901 (59.04%)
symptomatic illness present, minimal restriction	48 (19.05%)	178 (13.97%)	226 (14.81%)
symptomatic illness, severe restriction	0 (0.00%)	5 (0.39%)	5 (0.33%)
Missing	45 (17.86%)	71 (5.57%)	116 (7.60%)

	OKS at 5 years missing (N = 252)	**OKS at 5 years available (N = 1274)**	**Total (N = 1526)**
Post-operative complications*			
yes	38 (15.08%)	183 (14.36%)	221 (14.48%)
no	170 (67.46%)	1065 (83.59%)	1235 (80.93%)
Missing	44 (17.46%)	26 (2.04%)	70 (4.59%)
OKS*	15 (11, 22), (1, 39)	18 (13, 23), (0, 44)	18 (13, 23), (0, 44)

**Frequency and percentages are displayed*
^Median, IQR and range are displayed

To further assess the potential differences between participants with and without missing data at the five year follow-up, univariate logistic regression models were used to estimate the odds ratio (OR) for non-response with 95% confidence intervals, with each covariate used as a single predictor. It should be noted that some baseline data is missing. Imputation was not performed to obtain data for the missing covariates, as reflected by the varying amount of observations included into each univariate model (N).

The results from the univariate models are shown in Table 3-19, with statistically significant results highlighted in bold. The reference categories are displayed with parameter estimates equalling one and standard errors of these estimates of zero (categorical variables only). The two most severe ASA physical status (symptomatic illness causing severe restrictions, and moribund) have been excluded from the logistic regression models as no participants within these categories have OKS data available at five years. The univariate analyses (Table 3-19) indicate that especially baseline OKS, together with ASA physical status, age and height may be predictive of the probability of outcome data being missing at the five year follow-up assessment.

Table 3-19: Univariate analysis to identify possible predictors of OKS data being missing at the five year assessment - KAT

Parameter label	values	N	Parameter estimate (OR)	SE of parameter estimate	Standard normal deviate	P-value	Lower 95% confidence limit	Upper 95% confidence limit
Randomisation allocation patella resurfacing vs. no patella resurfacing	No patella resurfacing	1526	1	0			1	1
Randomisation allocation patella resurfacing vs. no patella resurfacing	Patella resurfacing	1526	0.912	0.126	-0.665	0.506	0.696	1.196
Gender	Female	1526	1	0			1	1
Gender	Male	1526	0.779	0.110	-1.763	0.078	0.590	1.028
Age at operation	Age (years)	1482	1.029	0.010	2.956	**0.003**	1.010	1.049
Weight	Weight (kg)	1464	0.996	0.005	-0.856	0.392	0.987	1.005
Height	Height (cm)	1444	0.971	0.007	-3.870	**0.000**	0.956	0.985
BMI	BMI	1438	1.021	0.013	1.585	0.113	0.995	1.047
Primary type of knee arthritis	Rheumatoid	1477	1	0			1	1
Primary type of knee arthritis	Osteoarthritis	1477	0.864	0.306	-0.413	0.679	0.431	1.731
Primary type of knee arthritis	Both	1477	2.450	3.119	0.704	0.481	0.202	29.696
Arthritis location	Single knee	1477	1	0			1	1
Arthritis location	Both knees	1477	1.138	0.211	0.698	0.485	0.792	1.636
Arthritis location	General	1477	1.197	0.230	0.936	0.349	0.821	1.744
Previous knee surgery	No previous knee surgery	1472	1	0			1	1
Previous knee surgery	Previous knee surgery	1472	0.867	0.136	-0.906	0.365	0.638	1.180

Parameter label	values	N	Parameter estimate (OR)	SE of parameter estimate	Standard normal deviate	P-value	Lower 95% confidence limit	Upper 95% confidence limit
ASA physical status	Completely fit and healthy	1405	1	0			1	1
ASA physical status	Some illness but no effect on daily activity	1405	1.156	0.239	0.701	0.483	0.771	1.734
ASA physical status	Symptomatic illness present, minimal restriction	1405	1.935	0.474	2.696	**0.007**	1.198	3.127
Post-operative complications	No operative complications	1456	1	0			1	1
Post-operative complications	Operative complications	1456	1.301	0.256	1.339	0.181	0.885	1.912
OKS at baseline	OKS at baseline	1442	0.956	0.010	-4.221	**0.000**	0.937	0.976

3.4.1.2 KAT: EQ-5D-3L univariate missing data patterns

Following the principles outlined above, information from the univariate models to predict the probability of the EQ-5D-3L being missing is presented in Table 3-20. Similar to the univariate analyses modelling the missingness in the OKS at five years post randomisation, baseline EQ-5D-3L, ASA physical status, age and height also seem to be significant in predicting the availability of the EQ-5D-3L at the five year follow-up.

Table 3-20: Univariate analysis to identify possible predictors of EQ-5D-3L composite score data being missing at the five year assessment - KAT

Parameter label	values	N	Parameter estimate (OR)	SE of parameter estimate	Standard normal deviate	P-value	Lower 95% confidence limit	Upper 95% confidence limit
Randomisation allocation patella resurfacing vs. no patella resurfacing	No patella resurfacing	1526	1	0			1	1
Randomisation allocation patella resurfacing vs. no patella resurfacing	Patella resurfacing	1526	0.831	0.111	-1.384	0.166	0.640	1.080
Gender	Female	1526	1	0			1	1
Gender	Male	1526	0.737	0.101	-2.216	0.027	0.563	0.965
Age at operation	Age (years)	1482	1.029	0.010	3.064	**0.002**	1.010	1.048
Weight	Weight (kg)	1464	0.997	0.005	-0.739	0.460	0.988	1.006
Height	Height (cm)	1444	0.970	0.007	-4.167	**0.000**	0.956	0.984
BMI	BMI	1438	1.025	0.013	1.983	0.047	1.000	1.050
Primary type of knee arthritis	Rheumatoid	1477	1	0			1	1
Primary type of knee arthritis	Osteoarthritis	1477	0.976	0.346	-0.069	0.945	0.487	1.954
Primary type of knee arthritis	Both	1477	2.450	3.119	0.704	0.481	0.202	29.697
Arthritis location	Single knee	1477	1	0			1	1
Arthritis location	Both knees	1477	1.056	0.187	0.308	0.758	0.747	1.493
Arthritis location	General	1477	1.147	0.210	0.748	0.455	0.801	1.641
Previous knee surgery	No previous knee surgery	1472	1	0			1	1
Previous knee surgery	Previous knee surgery	1472	0.822	0.125	-1.289	0.198	0.611	1.107

Parameter label	values	N	Parameter estimate (OR)	SE of parameter estimate	Standard normal deviate	P-value	Lower 95% confidence limit	Upper 95% confidence limit
ASA physical status	Completely fit and healthy	1405	1	0			1	1
ASA physical status	Some illness but no effect on daily activity	1405	1.182	0.233	0.847	0.397	0.803	1.739
ASA physical status	Symptomatic illness present, minimal restriction	1405	1.703	0.406	2.232	**0.026**	1.067	2.719
Post-operative complications	No operative complications	1456	1	0			1	1
Post-operative complications	Operative complications	1456	1.261	0.240	1.217	0.223	0.868	1.830
EQ-5D-3L at baseline	EQ-5D-3L at baseline	1417	0.468	0.112	-3.186	**0.001**	0.294	0.747

3.4.1.3 KAT: SF-12 version 2 univariate missing data patterns

Similar to the univariate models for the missingness in the OKS and EQ-5D-3L at five years post randomisation, the availability of the SF-12 version 2 subscales also seems to be significantly associated with ASA physical status, age and height, as well as the baseline MCS. The baseline PCS, on the other hand, does not seem to be predictive of SF-12 version 2 subscale availability at the five year follow-up (Table 3-21).

Table 3-21: Univariate analysis to identify possible predictors of SF-12 version 2 subscale data being missing at the five year assessment - KAT

Parameter label	values	N	Parameter estimate (OR)	SE of parameter estimate	Standard normal deviate	P-value	Lower 95% confidence limit	Upper 95% confidence limit
Randomisation allocation patella resurfacing vs. no patella resurfacing	No patella resurfacing	1422	1	0			1	1
Randomisation allocation patella resurfacing vs. no patella resurfacing	Patella resurfacing	1422	0.904	0.102	-0.900	0.368	0.725	1.127
Gender	Female	1422	1	0			1	1
Gender	Male	1422	1.076	0.122	0.644	0.520	0.861	1.344
Age at operation	Age (years)	1378	1.040	0.008	5.056	**0.000**	1.024	1.055
Weight	Weight (kg)	1360	0.999	0.004	-0.299	0.765	0.992	1.006
Height	Height (cm)	1340	0.985	0.006	-2.493	**0.013**	0.974	0.997
BMI	BMI	1334	1.007	0.011	0.684	0.494	0.986	1.029
Primary type of knee arthritis	Rheumatoid	1372	1	0			1	1
Primary type of knee arthritis	Osteoarthritis	1372	1.405	0.443	1.079	0.280	0.758	2.606
Primary type of knee arthritis	Both	1372	1	0			1	1
Arthritis location	Single knee	1373	1	0			1	1
Arthritis location	Both knees	1373	1.035	0.148	0.243	0.808	0.782	1.371
Arthritis location	General	1373	0.962	0.146	-0.258	0.797	0.715	1.294
Previous knee surgery	No previous knee surgery	1368	1	0			1	1
Previous knee surgery	Previous knee surgery	1368	1.022	0.125	0.176	0.860	0.804	1.298

Parameter label	values	N	Parameter estimate (OR)	SE of parameter estimate	Standard normal deviate	P-value	Lower 95% confidence limit	Upper 95% confidence limit
ASA physical status	Completely fit and healthy	1312	1	0			1	1
ASA physical status	Some illness but no effect on daily activity	1312	1.309	0.208	1.695	0.090	0.959	1.786
ASA physical status	Symptomatic illness present, minimal restriction	1312	1.892	0.376	3.210	**0.001**	1.282	2.793
ASA physical status		1312	1	0			1	1
Post-operative complications	No operative complications	1352	1	0			1	1
Post-operative complications	Operative complications	1352	1.294	0.206	1.616	0.106	0.947	1.767
SF-12 version 2 subscales at baseline	SF-12 PCS at baseline	1187	1.003	0.007	0.367	0.714	0.988	1.017
	SF-12 MCS at baseline	1187	0.986	0.005	-2.554	**0.011**	0.975	0.997

3.4.2 PD MED trial: univariate missing data models

3.4.2.1 PD MED: PDQ-39–SI univariate missing data patterns

Table 3-22 shows the results from the univariate analyses using the probability of the PDQ-39-SI being missing at the five year assessment as the outcome variable. As before, the number of observations included into each univariate model depends on the availability of explanatory variables. The initial analyses indicate that gender, age, Hoehn & Yahn stage, baseline PDQ-39-SI and time since initial PD diagnosis may be predictive of availability of the PDQ-39-SI at the five-year follow-up assessment.

Table 3-22: Univariate analysis to identify possible predictors of the PDS-39-SI data being missing at the five year assessment– PD MED trial

Parameter label	values	N	Parameter estimate (OR)	SE of parameter estimate	Standard normal deviate	P-value	Lower 95% confidence limit	Upper 95% confidence limit
Randomised treatment (LD vs. LD sparing)	LD	1113	1	0			1	1
Randomised treatment (LD vs. LD sparing)	LD sparing	1113	1.234	0.159	1.634	0.102	0.959	1.587
Gender	Female	1113	1	0			1	1
Gender	Male	1113	0.755	0.096	-2.197	**0.028**	0.588	0.970
Age at randomisation	Age (years)	1113	1.026	0.008	3.141	**0.002**	1.010	1.043
Hoehn & Yahr stage	Hoehn & Yahr Stage 1.0	1111	1	0			1	1
Hoehn & Yahr stage	Hoehn & Yahr Stage 1.5	1111	1.114	0.199	0.603	0.546	0.785	1.580
Hoehn & Yahr stage	Hoehn & Yahr Stage 2.0	1111	1.109	0.183	0.629	0.530	0.803	1.532
Hoehn & Yahr stage	Hoehn & Yahr Stage 2.5	1111	0.978	0.195	-0.113	0.910	0.661	1.445
Hoehn & Yahr stage	Hoehn & Yahr Stage 3.0	1111	1.734	0.467	2.042	**0.041**	1.022	2.941
Hoehn & Yahr stage		1111	1	0			1	1
Previous PD therapy	No	1113	1	0			1	1
Previous PD therapy	Less than 1 month	1113	0.977	0.438	-0.052	0.959	0.406	2.350
Previous PD therapy	1 to 3 months	1113	1.211	0.392	0.592	0.554	0.642	2.283
Previous PD therapy	3 to 6 months	1113	1.099	0.475	0.219	0.827	0.471	2.564
PDQ-39-SI at baseline	PDQ-39-SI (baseline)	988	1.013	0.005	2.637	**0.008**	1.003	1.023
Does the participant have a regular carer?	No	1113	1	0			1	1
Does the participant have a regular carer?	Yes	1113	0.806	0.105	-1.654	0.098	0.624	1.041
Time since initial diagnosis of PD (years)	less than one year ago	1113	1	0			1	1
Time since initial diagnosis of PD (years)	1 to 2 years ago	1113	0.631	0.097	-2.985	**0.003**	0.466	0.854
Time since initial diagnosis of PD (years)	2 to 3 years ago	1113	0.872	0.205	-0.583	0.560	0.549	1.383
Time since initial diagnosis of PD (years)	More than 3 years ago	1113	0.906	0.236	-0.380	0.704	0.544	1.509

3.4.2.2 PD MED: EQ-5D-3L composite score univariate missing data patterns

Table 3-23 shows the results from the univariate analyses using the probability of the EQ-5D-3L being missing at the five year assessment as the outcome variable. The initial analyses indicate that age, baseline EQ-5D-3L, availability of a regular carer and time since initial PD diagnosis may be predictive of availability of the EQ-5D-3L at the five-year follow-up assessment.

Table 3-23: Univariate analysis to identify possible predictors of the EQ-5D-3L data being missing at the five year assessment – PD MED trial

Parameter label	values	N	Parameter estimate (OR)	SE of parameter estimate	Standard normal deviate	P-value	Lower 95% confidence limit	Upper 95% confidence limit
Randomised treatment (LD vs. LD sparing)	LD	1113	1	0			1	1
Randomised treatment (LD vs. LD sparing)	LD sparing	1113	1.033	0.145	0.234	0.815	0.785	1.360
Gender	Female	1113	1	0			1	1
Gender	Male	1113	0.932	0.131	-0.499	0.618	0.708	1.228
Age at randomisation	Age (years)	1113	1.018	0.009	2.040	**0.041**	1.001	1.036
Hoehn & Yahr stage	Hoehn & Yahr Stage 1.0	1111	1	0			1	1
Hoehn & Yahr stage	Hoehn & Yahr Stage 1.5	1111	1.087	0.209	0.433	0.665	0.746	1.584
Hoehn & Yahr stage	Hoehn & Yahr Stage 2.0	1111	0.918	0.166	-0.476	0.634	0.644	1.308
Hoehn & Yahr stage	Hoehn & Yahr Stage 2.5	1111	0.843	0.186	-0.771	0.441	0.547	1.300
Hoehn & Yahr stage	Hoehn & Yahr Stage 3.0	1111	1.225	0.357	0.696	0.487	0.692	2.168
Hoehn & Yahr stage		1111	1	0			1	1
Previous PD therapy	No	1113	1	0			1	1
Previous PD therapy	Less than 1 month	1113	1.062	0.514	0.124	0.901	0.411	2.742
Previous PD therapy	1 to 3 months	1113	1.315	0.451	0.798	0.425	0.671	2.576
Previous PD therapy	3 to 6 months	1113	1.000	0.480	-0.001	0.999	0.390	2.562
EQ-5D-3L at baseline	PDQ-39-SI (baseline)	1094	0.304	0.084	-4.305	**0.000**	0.177	0.523
Does the participant have a regular carer?	No	1113	1	0			1	1
Does the participant have a regular carer?	Yes	1113	0.683	0.096	-2.704	**0.007**	0.518	0.900
Time since initial diagnosis of PD (years)	less than one year ago	1113	1	0			1	1
Time since initial diagnosis of PD (years)	1 to 2 years ago	1113	0.634	0.111	-2.597	**0.009**	0.450	0.894
Time since initial diagnosis of PD (years)	2 to 3 years ago	1113	1.047	0.264	0.183	0.855	0.639	1.715
Time since initial diagnosis of PD (years)	More than 3 years ago	1113	1.385	0.369	1.221	0.222	0.821	2.336

3.4.3 PD SURG trial: univariate missing data models

3.4.3.1 PD SURG: PDQ-39-SI univariate missing data patterns

Table 3-24 shows the results from the univariate analyses using the probability of PDQ-39-SI being missing as the outcome variable. The initial univariate analyses indicate that only previous COMT inhibitor use (i.e. use before randomisation into the trial) may be predictive of availability of the PDQ-39-SI at the five-year follow-up assessment.

Table 3-24: Univariate analysis to identify possible predictors of the PDS-39-SI data being missing at the five year assessment – PD SURG trial

Parameter label	values	N	Parameter estimate (OR)	SE of parameter estimate	Standard normal deviate	P-value	Lower 95% confidence limit	Upper 95% confidence limit
Randomised treatment	Drugs	325	1	0			1	1
Randomised treatment	Surgery	325	0.940	0.209	-0.279	0.780	0.608	1.452
Gender	Female	325	1	0			1	1
Gender	Male	325	0.946	0.232	-0.227	0.820	0.584	1.531
Age at randomisation	Age (years)	325	0.995	0.015	-0.342	0.733	0.966	1.025
Hoehn & Yahr Stage (categorised)	Hoehn & Yahr Stage 0, 1.0 or 1.5	325	0.873	0.379	-0.313	0.755	0.373	2.045
Hoehn & Yahr Stage (categorised)	Hoehn & Yahr Stage 2.0	325	1	0			1	1
Hoehn & Yahr Stage (categorised)	Hoehn & Yahr Stage 2.5	325	0.946	0.259	-0.204	0.838	0.553	1.617
Hoehn & Yahr Stage (categorised)	Hoehn & Yahr Stage 3.0 or 4.0	325	0.870	0.234	-0.518	0.604	0.513	1.475
PDQ-39-SI	PDQ-39-SI (baseline)	275	1.007	0.009	0.827	0.408	0.990	1.024
Regular carer	No	320	1	0			1	1
Regular carer	Yes	320	0.911	0.241	-0.351	0.726	0.543	1.530
Time since initial diagnosis of PD (years)	5 years ago or less	312	1	0			1	1
Time since initial diagnosis of PD (years)	6 to 10 years ago	312	2.483	1.346	1.678	0.093	0.858	7.183
Time since initial diagnosis of PD (years)	11 to 15 years ago	312	1.928	1.041	1.217	0.224	0.670	5.552
Time since initial diagnosis of PD (years)	More than 15 years ago	312	1.375	0.781	0.561	0.575	0.452	4.184
Previous Dopamine	No	325	1	0			1	1
Previous Dopamine	Yes	325	1.349	0.917	0.440	0.660	0.356	5.116
Previous Selegiline	No	325	1	0			1	1
Previous Selegiline	Yes	325	1.036	0.230	0.157	0.875	0.670	1.600
Previous COMTI	No	325	1	0			1	1
Previous COMTI	Yes	325	1.604	0.363	2.088	**0.037**	1.029	2.501
Previous Apomorphine	No	325	1	0			1	1
Previous Apomorphine	Yes	325	1.431	0.328	1.562	0.118	0.913	2.242

Parameter label	values	N	Parameter estimate (OR)	SE of parameter estimate	Standard normal deviate	P-value	Lower 95% confidence limit	Upper 95% confidence limit
Tremor	No	325	1	0			1	1
Tremor	Yes	325	0.691	0.157	-1.632	0.103	0.443	1.077
Dyskinsia	No	325	1	0			1	1
Dyskinsia	Yes	325	0.890	0.219	-0.473	0.636	0.550	1.442
Severe on/off periods	No	325	1	0			1	1
Severe on/off periods	Yes	325	0.885	0.235	-0.462	0.644	0.526	1.488
Other reason for considering surgery	No	325	1	0			1	1
Other reason for considering surgery	Yes	325	1.013	0.422	0.032	0.974	0.448	2.293
Apomorphine prescribed (if randomised to medical therapy)	No	325	1	0			1	1
Apomorphine prescribed (if randomised to medical therapy)	Yes	325	0.922	0.241	-0.312	0.755	0.552	1.538

3.4.3.2 PD SURG EQ-5D-3L composite score univariate missing data patterns

Table 3-25 shows the results from the univariate analysis using the probability of the EQ-5D-3L being missing as the outcome variable. As before, the number of observations included into each univariate model depends on the availability of explanatory variables. Similarly to the univariate analysis for the PDQ-39-SI, the univariate models indicate that, again, only previous COMT inhibitor use (i.e. use before randomisation into the trial) may be predictive of availability of the EQ-5D-3L composite score at the five-year follow-up assessment.

Table 3-25: Univariate analysis to identify possible predictors of the EQ-5D-3L composite score data being missing at the five year assessment – PD SURG trial

Parameter label	values	N	Parameter estimate (OR)	SE of parameter estimate	Standard normal deviate	P-value	Lower 95% confidence limit	Upper 95% confidence limit
Randomised treatment	Drugs	325	1	0			1	1
Randomised treatment	Surgery	325	1.040	0.233	0.176	0.860	0.670	1.614
Gender	Female	325	1	0			1	1
Gender	Male	325	1.004	0.249	0.015	0.988	0.618	1.632
Age at randomisation	Age (years)	325	0.997	0.015	-0.182	0.856	0.968	1.027
Hoehn & Yahr Stage (categorised)	Hoehn & Yahr Stage 0, 1.0 or 1.5	325	0.969	0.429	-0.071	0.943	0.407	2.306
Hoehn & Yahr Stage (categorised)	Hoehn & Yahr Stage 2.0	325	1	0			1	1
Hoehn & Yahr Stage (categorised)	Hoehn & Yahr Stage 2.5	325	0.879	0.254	-0.446	0.656	0.499	1.548
Hoehn & Yahr Stage (categorised)	Hoehn & Yahr Stage 3.0 or 4.0	325	0.696	0.198	-1.272	0.203	0.398	1.217
EQ-5D-3L	EQ-5D-3L (baseline)	311	0.477	0.209	-1.691	0.091	0.202	1.125
Regular carer	No	320	1	0			1	1
Regular carer	Yes	320	0.815	0.216	-0.770	0.441	0.485	1.371
Time since initial diagnosis of PD (years)	5 years ago or less	312	1	0			1	1
Time since initial diagnosis of PD (years)	6 to 10 years ago	312	2.770	1.668	1.692	0.091	0.851	9.017
Time since initial diagnosis of PD (years)	11 to 15 years ago	312	2.889	1.735	1.767	0.077	0.890	9.374
Time since initial diagnosis of PD (years)	More than 15 years ago	312	1.744	1.098	0.883	0.377	0.507	5.994
Previous Dopamine	No	325	1	0			1	1
Previous Dopamine	Yes	325	1.531	1.096	0.594	0.552	0.376	6.230
Previous Selegiline	No	325	1	0			1	1
Previous Selegiline	Yes	325	1.047	0.235	0.207	0.836	0.675	1.625
Previous COMTI	No	325	1	0			1	1
Previous COMTI	Yes	325	1.649	0.380	2.170	**0.030**	1.050	2.590

Parameter label	values	N	Parameter estimate (OR)	SE of parameter estimate	Standard normal deviate	P-value	Lower 95% confidence limit	Upper 95% confidence limit
Previous Apomorphine	No	325	1	0			1	1
Previous Apomorphine	Yes	325	1.352	0.310	1.314	0.189	0.862	2.121
Tremor	No	325	1	0			1	1
Tremor	Yes	325	0.698	0.161	-1.563	0.118	0.444	1.096
Dyskinsia	No	325	1	0			1	1
Dyskinsia	Yes	325	0.888	0.219	-0.480	0.631	0.547	1.442
Severe on/off periods	No	325	1	0			1	1
Severe on/off periods	Yes	325	1.324	0.359	1.034	0.301	0.778	2.253
Other reason for considering surgery	No	325	1	0			1	1
Other reason for considering surgery	Yes	325	0.726	0.314	-0.741	0.459	0.311	1.694
Apomorphine prescribed (if randomised to medical therapy)	No	325	1	0			1	1
Apomorphine prescribed (if randomised to medical therapy)	Yes	325	1.012	0.267	0.045	0.964	0.604	1.696

3.5 Predictors of missing PROMs at five years: multivariate models

3.5.1 Model selection - methodology

Following the univariate analyses presented above, multivariate logistic regression models were devised to predict the probability of missingness in the subset of patients that are not known to have died before the five year follow-up. Randomised intervention was kept in the models, regardless of its statistical significance. For all other variables, backwards model selection was utilised. Using this approach, a logistic regression model was fitted including all possible covariates (i.e. all covariates used in the univariate analysis above). Explanatory variables were removed iteratively if they were statistically insignificant at the 5% level and contributed least to the model, based on the highest p-values[103-106]. As significance for categorical variables relies on the differences between a relevant category and the reference category in the statistical model, the category used as the reference category was varied where appropriate to confirm that a specific variable was in fact insignificant in the model and could be removed.

To ensure comparability of the nested models, a subset of data including only participants with data available for all possible independent variables was used in the model selection process. The final model was refitted to the trial population using an available cases approach excluding only those participants with missing data for the variables included in the final model.

3.5.2 KAT trial: multivariate predictors of missing outcome data at five years

3.5.2.1 KAT: multivariate models to predict missing OKS outcome data

In the model selection process for the KAT dataset, BMI was initially included instead of a combination of height and weight to avoid problems with multicollinearity. After BMI was not found to be significant, height was included as the univariate analysis had shown that this may be potentially predictive of missing OKS at the five year follow-up.

The backwards selection process resulted in baseline OKS, age, height and ASA physical status remaining in the final model together with the randomised intervention, which was forced to remain in the model despite not being statistically significant. It was thought that height may be a potential confounder for gender. However, gender remained insignificant in the model when substituted for the height variable. Results of the final regression are shown in Table 3-26; the number of observations included reflects exclusions due to missing baseline data; some ASA categories are omitted due to perfect predictions.

Table 3-26: Results from the logistic regression model predicting whether OKS data is missing at the five year follow-up (KAT)

	OR	SE	p-value	95% CI
Patella resurfacing	0.939	0.151	0.694	(0.686, 1.286)
Baseline OKS	0.952	0.011	<0.001	(0.93, 0.974)
Age	1.023	0.011	0.027	(1.003, 1.044)
Height	0.976	0.008	0.005	(0.96, 0.993)
ASA physical status				
completely fit and healthy vs. some illness but no effect on daily activity	1.051	0.236	0.826	(0.676, 1.633)
symptomatic illness present, minimal restriction vs. some illness but no effect on daily activity	1.529	0.308	0.035	(1.03, 2.27)
Constant	3.789	6.149	0.412	(0.158, 91.15)

With the given explanatory variables collected at baseline, the above results describe the best, most parsimonious model to predict the probability of data being missing at the five year follow-up. However, the model has low discriminative ability, as can be seen in the histogram (Figure 3-18), i.e. the model cannot distinguish well between those with and without missing data, shown by similar predicted probabilities for those with and without missing outcome data.

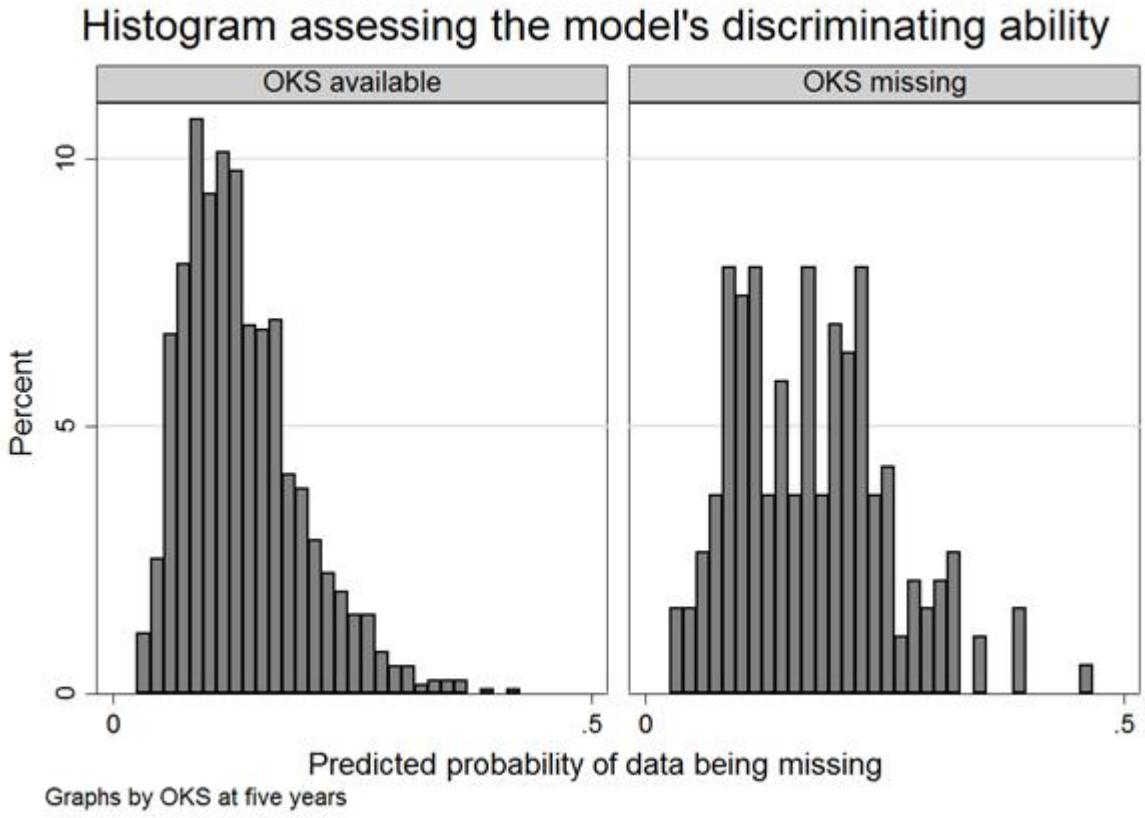


Figure 3-18: Assessment of the model predicting missing data in the OKS at five years (KAT)

This suggests that the probability of OKS data being missing at the five year follow-up may also be dependent on variables other than those included in the logistic regression model. These variables may be observed or unobserved in the RCT; a MAR assumption may not adequately represent the true underlying missing data mechanism.

The above models have used a binary variable indicating whether the OKS is missing at the five year follow-up. In order to more fully explore potential predictors of missingness in the OKS, the prediction models have been repeated using the following set of different outcome variables:

- Indicator of the OKS missing at any of the follow-up time points (binary variable)
- Indicator of the OKS missing at half the follow-up time points or more (binary variable)
- Number of missing OKS follow-up scores (ordinal variable, taking values 0 to 6)
- Number of missing OKS items across all follow-up (continuous variable, range 0 to 72)

The model selection process followed the same principles as described above. Details of the coefficients, corresponding 95% confidence intervals and p-values are displayed in Table 3-27. The prediction models were generated to explore possible predictors of missingness, and no checks of model fit and their underlying assumptions have been performed. The models are consistent in that baseline OKS and participant height seem to be predictive of OKS missingness at follow-up. Treatment allocation is not statistically significant in any of these additional models. Age or ASA physical status were also not statistically significant in any of the additional models, and were removed during the selection process.

Table 3-27: Prediction models of OKS missingness at follow-up (KAT)

Outcome variable		**Missing data at 5 years**	**Any missing data during follow-up**	**At least half of follow-up data missing**	**Number of follow-up scores missing (0-7)**	**Number of items missing across all follow-up (0-72)**
Model used		Logistic regression	Logistic regression	Logistic regression	Ordinal logit	Continuous regression
N		1334	1407	1407	1407	1407
Log-likelihood**		-519.42	-803.04	-464.63	-1375.14	0.0179
Explanatory variables:						
Patellar resurfacing vs. no resurfacing	OR/ regression coefficient	0.939	1.080	1.088	1.092	0.682
	95% CI	0.686, 1.286	0.850, 1.371	0.773, 1.531	0.863, 1.382	-0.951, 2.318
	p-value	0.694	0.529	0.628	0.462	0.413
Baseline OKS	OR/ regression coefficient	0.952	0.974	0.945	0.969	-0.238
	95% CI	0.930, 0.9745	0.958, 0.990	0.922, 0.969	0.953, 0.985	-0.349, -0.127
	p-value	<0.001	0.002	<0.001	<.001	<0.001
Age	OR/ regression coefficient	1.023	n/a	n/a	n/a	n/a
	95% CI	1.003, 1.044	n/a	n/a	n/a	n/a
	p-value	0.027	n/a	n/a	n/a	n/a
Height	OR/ regression coefficient	0.976	0.988	0.981	0.987	-0.097

Outcome variable		Missing data at 5 years	Any missing data during follow-up	At least half of follow-up data missing	Number of follow-up scores missing (0-7)	Number of items missing across all follow-up (0-72)
	95% CI	0.960, 0.993	0.976, 1.000	0.964, 0.999	0.975, 0.999	-0.182, -0.013
	p-value	0.005	0.056*	0.038	0.036	0.024
ASA physical status cat 1 vs. 2	OR/ regression coefficient	1.051	n/a	n/a	n/a	n/a
	95% CI	0.676, 1.633	n/a	n/a	n/a	n/a
	p-value	0.826	n/a	n/a	n/a	n/a
ASA physical status cat 3 vs. 2	OR/ regression coefficient	1.529	n/a	n/a	n/a	n/a
	95% CI	1.030, 2.270	n/a	n/a	n/a	n/a
	p-value	0.035	n/a	n/a	n/a	n/a
Constant	OR/ regression coefficient	3.789	4.067	6.833	n/a	28.165
	95% CI	0.157, 91.150	0.546, 30.285	0.375, 124.33	n/a	14.460, 41.870
	p-value	0.412	0.171	0.194	n/a	<0.001

* Was significant in selection model (complete cases), **R^2(adjusted) for the continuous regression model

3.5.2.2 KAT: multivariate models to predict missing EQ-5D-3L outcome data

In the multivariate model, baseline EQ-5D-3L, age and height were found to be predictive of the probability of missing EQ-5D-3L outcome data at baseline. As observed previously, treatment allocation is not statistically significant in the model. Here, ASA physical status at baseline was not found to be statistically significant at the 5% level in the model. There seems to be a strong association between height and gender. Gender would be statistically significant in the final model if substituted for height, but not once the model is also adjusted for height. The likelihood ratio test shows that the model fit is significantly improved if height is added to a model including the other statistically significant variables (and treatment allocation), rendering gender insignificant. The output from the logistic regression model, as well as the other models looking at varying aspects of missingness, as discussed above, are displayed in Table 3-28. As observed previously, the discriminative ability of the logistic regression model is low, as demonstrated in Figure 3-19.

Table 3-28: Prediction models of EQ-5D-3L missingness at follow-up (KAT)

Outcome variable		**Missing data at 5 years**	**Any missing data during follow-up**	**At least half of follow-up data missing**	**Number of follow-up scores missing (0-7)**	**Number of items missing across all follow-up (0-72)**
Model used		Logistic regression	Logistic regression	Logistic regression	Ordinal logit	Continuous regression
N		1372	1372	1.382	1372	1382
Log-likelihood**		-559.06	-855.33	-492.20	-1500.02	0.016
Explanatory variables:						
Patellar resurfacing vs. no resurfacing	OR/ regression coefficient	0.829	0.982	1.219	1.02	0.283
	95% CI	0.613, 1.122	0.782, 1.233	0.877, 1.694	0.814, 1.269	-0.410, 0.977
	p-value	0.224	0.877	0.239	0.888	0.423
Baseline EQ-5D-3L	OR/ regression coefficient	0.397	0.573	0.413	0.517	-2.338
	95% CI	0.240, 0.655	0.392, 0.838	0.240, 0.711	0.355, 0.751	-3.491, -1.186
	p-value	<0.001	0.004	0.001	0.001	<0.001
Age	OR/ regression coefficient	1.025	1.018	n/a	0.017	n/a
	95% CI	1.005, 1.045	1.003, 1.033	n/a	0.003, 1.032	n/a
	p-value	0.013	0.020	n/a	0.020	n/a
Height	OR/ regression coefficient	0.975	0.988	0.976	0.987	-0.045

Outcome variable		**Missing data at 5 years**	**Any missing data during follow-up**	**At least half of follow-up data missing**	**Number of follow-up scores missing (0-7)**	**Number of items missing across all follow-up (0-72)**
	95% CI	0.960, 0.990	0.977, 1.000	0.959, 0.992	0.975, 0.998	-0.080, -0.009
	p-value	0.002	0.047	0.005	0.021	0.013
Constant	OR/ regression coefficient	3.113	1.283	9.438	n/a	11.321
	95% CI	0.149, 65.264	0.131, 12.564	0.593, 150.20	n/a	5.534, 17.108
	p-value	0.464	0.830	0.112	n/a	<0.001

* Was significant in selection model (complete cases), **R^2(adjusted) for the continuous regression model

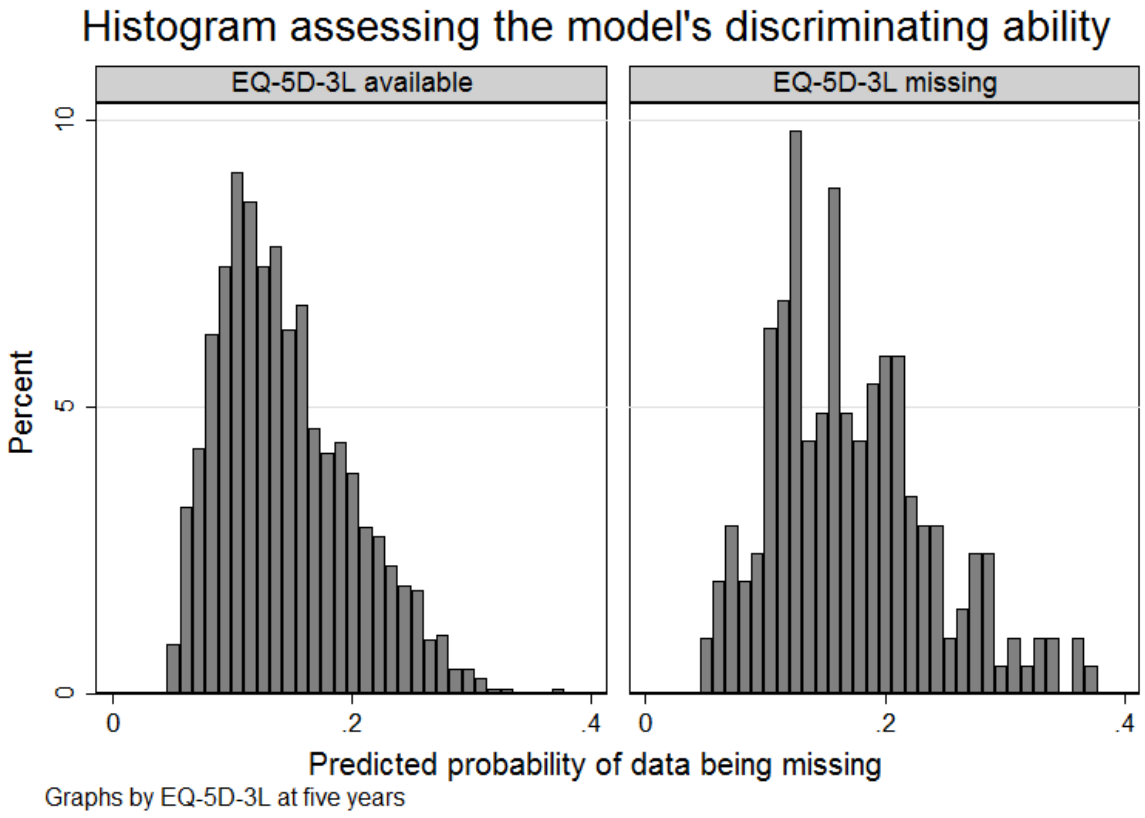


Figure 3-19: Assessment of the model predicting missing data in the EQ-5D-3L at five years (KAT)

3.5.2.3 KAT: multivariate models to predict missing SF-12 outcome data

Both baseline SF-12 subscales were included in the selection model; correlations between these variables was sufficiently low to not raise any concerns about multicollinearity. The variable selection results in age and baseline SF-12 MCS score remaining in the final model as statistically significant variables. The results for the different missingness models are shown in Table 3-29. Again, low discriminatory ability of the main logistic regression model is demonstrated in Figure 3-20.

Table 3-29: Prediction models of missingness at follow-up in the SF-12 subscales (KAT)

Outcome variable		**Missing data at 5 years**	**Any missing data during follow-up**	**At least half of follow-up data missing**	**Number of follow-up scores missing (0-7)**	**Number of items missing across all follow-up (0-72)**
Model used		Logistic regression	Logistic regression	Logistic regression	Ordinal logit	Continuous regression
N		1172	1145	1172	1145	1154
Log-likelihood**		-701.89	-735.68	-645.01	-1865.77	0.031
Explanatory variables:						
Patellar resurfacing vs. no resurfacing	OR/ regression coefficient	0.897	0.770	0.985	0.873	1.041
	95% CI	0.696, 1.155	0.603, 0.983	0.723, 1.236	0708, 1.077	-0.760, 2.842
	p-value	0.400	0.036	0.681	0.206	0.257
Baseline SF-12 MSC score	OR/ regression coefficient	0.984	0.983	0.975	0.977	-0.225
	95% CI	0.673, 0.995	0.972, 0.994	0.963, 0.986	0.968, 0.986	-0.305, -0.145
	p-value	0.04	0.003	<0.001	<0.001	>0.001
Age	OR/ regression coefficient	1.035	1.031	1.026	1.030	
	95% CI	1.018, 1.052	1.015, 1.047	1.008, 1.043	1.017, 1.045	

Outcome variable		**Missing data at 5 years**	**Any missing data during follow-up**	**At least half of follow-up data missing**	**Number of follow-up scores missing (0-7)**	**Number of items missing across all follow-up (0-72)**
	p-value	<0.001	<0.01	0.004	<0.001	
Height	OR/ regression coefficient		0.983		0.986	-0.101
	95% CI		0.971, 0.996		0.986, 0.997	-0.191, -0.010
	p-value		0.008		0.013	0.030
Constant	OR/ regression coefficient	0.095	9.375	0.209		36.082
	95% CI	0.028, 0.331	0.866, 101.492	0.057, 0.764		21.064, 51.099
	p-value	<0.001	0.066	0.018		>0.001

* Was significant in selection model (complete cases), **R^2(adjusted) for the continuous regression model

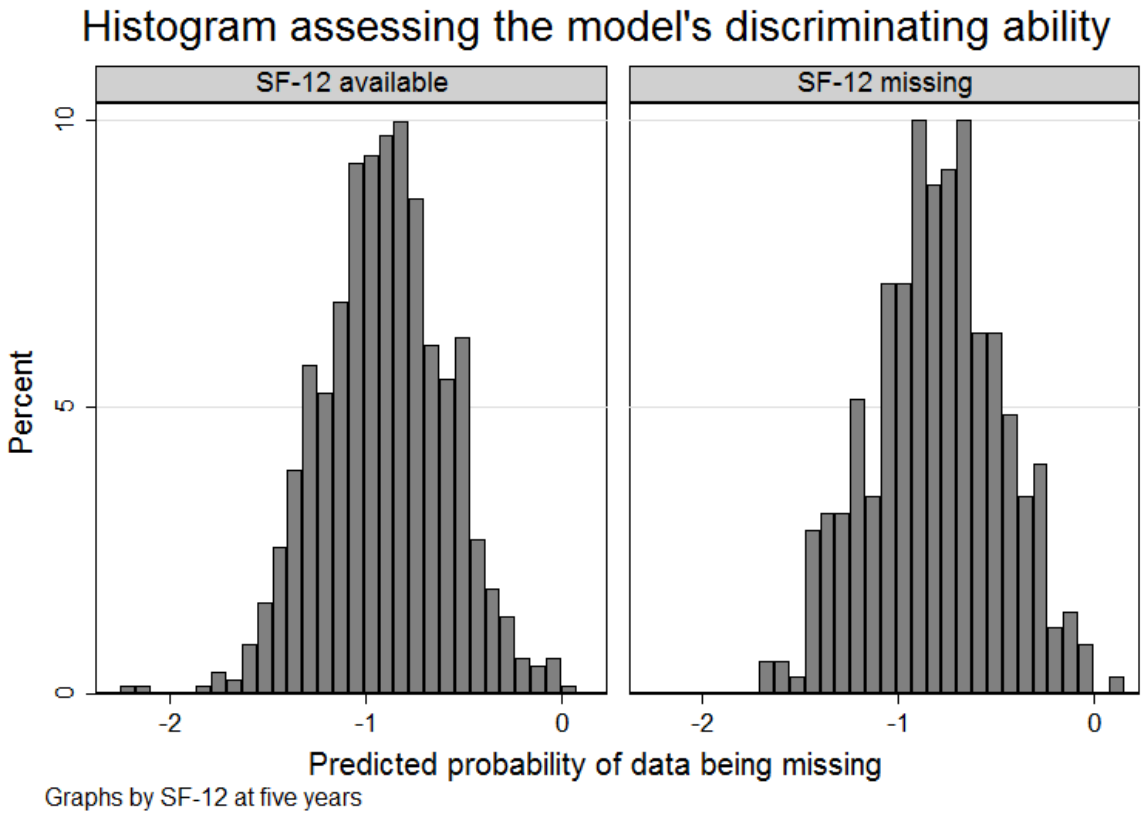


Figure 3-20: Assessment of the model predicting missing data in the SF-12 at five years (KAT)

3.5.3 PD MED data

Similarly to the process described above in section 3.5.1, a backwards selection model was used to identify possible predictors for the PROMs data being missing at the five year follow-up in a multivariate logistic regression model, using a subset of participants who are alive at the five year assessment and have available data for all relevant covariates.

3.5.3.1 PD MED: multivariate models to predict missing PDQ-39-SI outcome data

The randomised treatment, baseline PDQ-39-SI, age at randomisation and the time since initial PD diagnosis (less than one year, one to two years, two to three years, greater than three years) were statistically significant in the final logistic regression model using PDQ-39-SI availability at five years as the outcome. The Hoehn & Yahr stage, which was statistically significant in the univariate models ceased to be significant in the model once it was adjusted for other patient characteristics. The model results for this logistic regression model, as well as the models investigating other aspects of missingness are shown in Table 3-30.

Table 3-30: Prediction models of missingness within the PDQ-39-SI at follow-up (PD MED)

Outcome variable		**Missing data at 5 years**	**Any missing data during follow-up**	**At least half of follow-up data missing**	**Number of follow-up scores missing (0-7)**	**Number of items missing across all follow-up (0-72)**
Model used		Logistic regression	Logistic regression	Logistic regression	Ordinal logit	Continuous regression
N		988	988	988	988	988
Log-likelihood**		-623.85	-642.86	-473.51	-1531.68	0.044
Explanatory variables:						
LD sparing vs. LD medication	OR/ regression coefficient	1.347	1.267	1.356	1.229	0.657
	95% CI	1.021, 1.774	0.947., 1.616	0.970, 1.896	0.973, 1.553	-0.711, 2.025
	p-value	0.035	0.118	0.075	0.084	0.346
Baseline PDQ-39-SI	OR/ regression coefficient	1.015	1.024	1.036	1.028	0.148
	95% CI	1.005, 1.025	1.014, 1.035	1.024, 1.048	1.019, 1.037	0.098, 0.199
	p-value	0.003	<0.001	<0.001	<0.001	<0.001
Age	OR/ regression coefficient	1.035	1.030	1.031	1.030	n/a
	95% CI	1.017, 0.054	1.013, 1.048	1.009, 0.054	1.015, 1.046	n/a
	p-value	<0.001	0.001	0.005	<0.001	n/a
Symptoms 1-2 years vs. <1 year	OR/ regression coefficient	0.588	0.672	n/a	0.715	-1.669
	95% CI	0.422, 0.819	0.495, 0.913	n/a	0.544, 0.942	-3.257, -0.080

Outcome variable		Missing data at 5 years	Any missing data during follow-up	At least half of follow-up data missing	Number of follow-up scores missing (0-7)	Number of items missing across all follow-up (0-72)
	p-value	0.002	0.011	n/a	0.017	0.040
Symptoms 2-3 years vs. <1 year	OR/ regression coefficient	0.882	0.796	n/a	0.876	-0.403
	95% CI	0.539, 1.443	0.494, 1.282	n/a	0.571, 1.346	-2.881, 2.076
	p-value	0.618	0.347	n/a	0.547	0.750
Symptoms 3+ years vs. <1 year	OR/ regression coefficient	0.982	1.064	n/a	1.242	3.128
	95% CI	0.572, 1.686	0.613, 1.850	n/a	0.770, 2.004	0.353, 5.902
	p-value	0.948	0.825	n/a	0.374	0.027
Carer vs. not having a carer	OR/ regression coefficient	n/a	n/a	n/a	n/a	-1.915
	95% CI	n/a	n/a	n/a	n/a	-3.348, -0.481
	p-value	n/a	n/a	n/a	n/a	0.009
Constant	OR/ regression coefficient	0.033	0.114	0.010	n/a	4.862
	95% CI	0.008, 0.128	0.032, 0.406	0.002, 0.054	n/a	3.020, 6.704
	p-value	<0.001	0.001	<0.001	n/a	<0.001

* Was significant in selection model (complete cases), **R^2(adjusted) for the continuous regression model

3.5.3.2 PD MED: multivariate models to predict missing EQ-5D-3L outcome data

Using the backwards model selection approach, age, baseline EQ-5D-3L composite scores, availability of a carer and the duration of symptoms before trial entry appear to be significant in predicting the probability of EQ-5D-3L outcome data being available at the five year follow-up in the main logistic regression model. Contrasting this with the model for missing PDQ-39-SI data, the treatment allocation is not statistically significant in predicting the availability of the EQ-5D-3L. The results for the various models investigating aspects of missingness are shown in Table 3-31.

Table 3-31: Prediction models of missingness within the EQ-5D-3L at follow-up (PD MED)

Outcome variable		**Missing data at 5 years**	**Any missing data during follow-up**	**At least half of follow-up data missing**	**Number of follow-up scores missing (0-7)**	**Number of items missing across all follow-up (0-72)**
Model used		Logistic regression	Logistic regression	Logistic regression	Ordinal logit	Continuous regression
N		1094	1094	1094	1094	1094
Log-likelihood**		-607.41	-695.79	-380.91	-1273.45	0.034
Explanatory variables:						
LD sparing vs. LD medication	OR/ regression coefficient	1.065	0.941	1.043	0.980	-0.021
	95% CI	0.802, 1.415	0.727, 1.218	0.708, 1.536	0.764, 1.257	-0.855, 0.812
	p-value	0.622	0.644	0.832	0.872	0.960
Baseline PDQ-39-SI	OR/ regression coefficient	0.282	0.356	0.206	0.335	-4.280
	95% CI	0.162, 0.490	0.212, 0.599	0.104, 0.408	0.205, 0.546	-5.980, -2.580
	p-value	<0.001	<0.001	<0.001	<0.001	<0.001
Age	OR/ regression coefficient	1.023	n/a	0.993	1.004	n/a
	95% CI	1.005. 1.042	n/a	0.969, 1.017	0.989, 1.020	n/a
	p-value	0.014	n/a	0.458	0.587	n/a
Symptoms 1-2 years vs. <1 year	OR/ regression coefficient	0.651	0.694	0.857	0.737	-0.791

Outcome variable		Missing data at 5 years	Any missing data during follow-up	At least half of follow-up data missing	Number of follow-up scores missing (0-7)	Number of items missing across all follow-up (0-72)
	95% CI	0.459, 0.879	0.509, 0.956	0.534, 1.375	0.545, 0.997	-1.760, 0.177
	p-value	0.016	0.021	0.523	0.048	0.109
Symptoms 2-3 years vs. <1 year	OR/ regression coefficient	1.089	0.940	0.822	0.965	-0.181
	95% CI	0.655, 1.813	0.584, 1.153	0.534, 1.780	0.609, 1.529	-1.724, 1.363
	p-value	0.742	0.799	0.620	0.878	0.818
Symptoms 3+ years vs. <1 year	OR/ regression coefficient	1.432	1.162	2.187	1.419	2.157
	95% CI	0.835, 2.453	0.692, 1.950	1.163, 1.780	0.859, 2.346	0.438, 3.875
	p-value	0.192	0.570	0.015	0.172	0.015
Carer vs. not having a carer	OR/ regression coefficient	0.660	0.669	n/a	n/a	-1.336
	95% CI	0.458, 0.923	0.514, 0.872	n/a	n/a	-2.200, -0.472
	p-value	0.004	0.003	n/a	n/a	0.002
Constant	OR/ regression coefficient	0.212	1.576	0.571	n/a	7.627
	95% CI	0.054, 0.838	1.005, 2.470	0.100, 3.254	n/a	6.149, 9.105
	p-value	0.027	0.047	0.528	n/a	<0.001

* Was significant in selection model (complete cases), **R^2(adjusted) for the continuous regression model

3.5.4 PD SURG data

3.5.4.1 PD SURG: multivariate models to predict missing PDQ-39-SI outcome data

The model selection process followed the rules outline in section 3.5.1, using the subset of participants without missing data for all possible explanatory variables. None of the investigated explanatory variables were statistically significant in a multivariate logistic regression model attempting to explain the probability of the PDQ-39-SI being missing at the five year follow-up; the final model only included the constant and the randomised treatment, which was forced to remain in the model regardless of statistical significance. However, when fitting the final model to the available cases population (treatment allocation and previous COMTI use are available for all 325 participants), previous COMTI use was statistically significant at the 5% level. This indicates that the significance of the variables is highly dependent on the sample size of the subpopulation used. Of note, during the model selection process, as in the previous univariate analysis, some explanatory variables, namely time since diagnosis and baseline PDQ-39-SI had p-values of just over 5%. The PD SURG trial was not powered for this analysis, and it is possible that some of the explanatory variables would have met the 5% significance level in the adjusted model if a larger sample size had been available.

3.5.4.2 PD SURG: multivariate models to predict missing EQ-5D-3L outcome data

The variable selection process to identify variables predictive of the probability of the EQ-5D-3L composite scores being missing at five years produces results similar to the above model for the PDQ-39. Again, use of COMTI prior to trial enrolment appeared to be the only explanatory variable predictive of outcome data being missing; treatment allocation is not statistically significant in the model. As mentioned previously, a larger sample size

may have increased the likelihood of explanatory variables found to be statistically significant in the model.

The results for these regression models are not shown as they do not add to those presented above.

3.6 Discussion

The work described in this chapter identified patterns of missing data within PROMs in three different trials, with regards to item missingness, and longitudinal missing data patterns. It is important to note that the rates and patterns of missing data differ between PROMs as well as between trials, due to the different patient populations. Although it is unclear if the findings are generalisable to other trials as a whole, or trials with a comparable set-up in comparable populations, the identified patterns can be used as a basis to inform realistic simulation models in the comparison of different approaches to handling missing data in the statistical analysis.

The univariate and multivariate models identify variables collected at baseline that may be predictive of the probability of outcome data being missing. Therefore, it may be important to include these variables into the simulation of missing data, as well as MI models to analyse missing data. It is hence recommend to undertake similar investigative work ahead of analysing any studies with missing outcome data, in order to better understand missing data patterns and inform potential imputation models. Researchers should also bear in mind that a lack of evidence for a variable being statistically significantly associated with the probability of outcome data being missing does not mean that there may not be such a link, as the same explanatory variable may well have been statistically significant in a larger sample. Finally, the models investigated here had low discriminatory power to distinguish between participants with and without missing data, also indicated by the low values for R^2 adjusted (Table 3-27 to Table 3-31). Therefore, it is possible that additional, potentially unobserved variables play an important role in the missing data mechanism, reiterating the need to support any primary analysis with adequate sensitivity analysis[18, 107]. The reasons for missing data and participant withdrawal from follow-up could also

further support assumptions made about the missing data mechanism[108], but unfortunately were not available for this work. These may indicate if participants withdrew from the trial due to side effects, worsening or resolution of symptoms, which may indicate that data are MNAR. On the other hand, if missing data is due to staffing issues at sites, or participants being lost to follow-up due to a change of address, this may support a MAR assumption.

3.7 Conclusions

Missing data rates and patterns differ widely between trials and PROMs, indicating that no universal rules of handling missing data can be derived. Thus, it is important to investigate each data set carefully with regards to missing data in order to make appropriate assumptions about the underlying missing data mechanism and to evaluate which variables should be included in any imputation mechanisms. However, it is imperative to also remember that missing data patterns can also be associated with unobserved data, data collected after the baseline visit. The lack of statistical significance in the prediction models should be considered with caution, as variables may still be relevant for potential prediction models. The sensitivity of the results with regards to the assumed missing data mechanism should always be investigated through appropriate sensitivity analyses.

Chapter 4 : Multiple imputation for missing patient reported outcome measures in randomised controlled trials: Advantages and disadvantages of imputing at the item, subscale and composite score level

4.1 Introduction

Traditionally, research concerning missing data in PROMs has focussed on how the missing PROMs composite scores should be handled, with multiple imputation (MI) methods considered to be one of the most reliable methods[29, 109, 110]. Multi-item PROMs consist of multiple questions, referred to as items, which are combined into an overall composite scores, and sometimes subscales[7, 111]. Therefore, different types of imputation are possible, e.g. imputation of composite scores, subscales (where available) or separate items, the latter of which may yield additional information and therefore improve the accuracy of such imputations. Research has not commonly been performed on the comparison between these approaches.

This chapter presents an overview of the research that has been performed in this area to date. Such research is limited to specific questionnaires such as the EQ-5D-3L questionnaire[112], and the Pain Coping Inventory (PCI-active), a 12-item questionnaire[52]. The findings from these studies are put into the context of the research performed within the remit of this chapter, which aims to further validate the existing research, and investigates whether the results are generalisable to a wider range of PROMs. Advantages and limitations of multiple imputations at the item, subscale (where appropriate) and composite score level are considered in a range of real-life scenarios for three different PROMs, namely the OKS, EQ-5D-3L and SF-12. The impact on bias and precision of the

different imputation approaches for handling missing PROMs outcome data, as well as on the treatment coefficients of linear regression models in an RCT context are evaluated.

The content of this chapter follows approved and peer-reviewed guidance for simulation studies[113]. Relevant background is presented in section 4.2, with the hypotheses for the project, along with the research aims of this chapter outlined in section 4.3. Section 4.4 describes the rationale for using simulation, and details how the datasets and the missing PROMs outcome data were simulated, and how the different imputation approaches were implemented. In section 4.5, the different approaches are assessed for feasibility by considering convergence rates, and compared using root mean square error (RMSE) and mean absolute error (MAE).

A discussion of the methodology and results is provided in section 4.6. Here, the results and observed trends are summarised and compared to previous research. Recommendations on how missing PROMs outcome data in RCTs should be handled in each of the investigated scenarios are also provided. Emphasis is also put on the development of guidance on how to construct feasible multiple imputation models while circumventing problems with non-convergence. Novel aspects of this research, together with limitations are also discussed. Findings are summarised and brought together in the chapter conclusions in section 4.7.

4.2 Overview of the existing research

As mentioned in the introduction, very little research has been published to date comparing the performance of applying MI at the item versus composite score level. Research by Simons et al[112] compared imputation at the item and composite score level for estimating EQ-5D-3L composite scores in the presence of missing at random (MAR) data[iv]. Their base case simulations included 1814 observations and primarily followed a unit-nonresponse pattern of missingness (88.7% of missing data), with low rates of item missingness, often due to one missing item. For these base case scenarios, both MI approaches performed almost identically in terms of accuracy for all different proportions of missing data investigated (i.e. 5-40% of missing EQ-5D-3L data). As the sample size was decreased to 500 observations or fewer, both approaches performed similarly for up to 10% of missing data, however, MI at the composite score was found to be more accurate for 20% and 40% of missing data within these smaller sample sizes. MI at the item level was found to be performing better as the proportion of unit-nonresponse decreased, i.e. more missing data was due to individual items being unavailable. The authors recommend further research to assess generalisability of these findings to other PROMS with potentially different psychometric properties.

Eekhout et al[52] applied a number of different methods to account for missing data in the Pain Coping Inventory (PCI), which is a 12-item PROM. In their work, the PCI was used as a covariate in a regression model, as opposed to the outcome variable, and MI approaches were compared in terms of accuracy and precision of the fitted PCI regression coefficients. In this scenario, MI at the item level achieved the best results, while MI applied to the

[iv] Instead of items and composite scores, the terms domain and index are commonly used in EQ-5D-3L research. However, for consistency, the former terminology is used throughout the chapter.

composite scores resulted in overestimated standard errors where large percentages (>50%) of study participants had missing data. The authors also found that complete cases analysis (CCA), which does not impute missing data, yielded acceptable results in terms of regression coefficients, but overestimated standard errors, especially when more than 10% of the study population had some missing PROMs data, and therefore advised against the use of CCA. Finally, the authors raised concerns about item mean imputation, which many scoring manuals, including those for the Oxford scores[13, 42], advise for handling data with small amounts of missing items. In their simulation study, item mean imputation resulted in biased estimates, particularly where more than 10% of participants had some missing PROMs data.

4.3 Hypotheses and objectives for this research

The existing research on the advantages and disadvantages of applying MI at the item, subscale or composite score level, as presented in section 4.2, is limited and stipulates the need for further research. In particular, only two different PROMs have been considered, which raises the question whether the conclusions are generalisable to other PROMs, which may have additional categories within each items, more items than the EQ-5D-3L or potentially different psychometric properties[112]. Furthermore, the recommendations differ depending on whether the estimation of the PROM composite score is of interest, or whether it is to be used as an explanatory variable in a regression model. Finally, it is important to validate the findings in different datasets.

4.3.1 Hypotheses for this chapter

The three PROMs considered in this chapter are the OKS, EQ-5D-3L and SF-12.

For the OKS, the composite score is calculated as the unweighted sum of all 12 items[42]:

$$OKS_{composite\ score} = \sum_{i=1}^{12} item_i \qquad \text{where } item_i \in (0,1,2,3,4)$$

Applying this formula produces a composite score ranging from 0 to 48. The validated pain and function subscales are calculated similarly[111], each from a mutually exclusive subset of items, with the subscales being standardised to range from 0 to 100.

In essence, the SF-12 and EQ-5D-3L are scored similarly, with the addition of a constant and weights that are applied to each of the item levels. The weights and constant are derived from a country specific value set[43, 46]. An additional constant term is also added to the EQ-5D-3L if the most severe health state was chosen for at least one item. The SF-12

consists of 12 items which all contribute, using different weights, to the calculation of the MCS and PCS scores, which are reported.

Generally, composite scores or subscales cannot be derived if at least one item is missing, however, the OKS scoring manual suggests that up to two items (one from each of the subscales) can be substituted by the mean item score of the available items[13, 111].

All items contribute to the calculation of the composite scores and subscales. Therefore, the research hypothesis is, similarly to Simons et al[112], that where the MAR data follows an item nonresponse pattern, imputation at the item level is superior to that at the composite score or subscale level as the proportion of item nonresponse increases, as those later approaches disregard some of the available data. Correspondingly, it is hypothesised that where the MAR data follows primarily a unit-nonresponse pattern, the different MI approaches perform similarly, as the MI at the item level in this case cannot utilise any additional information that is not available to the MI at the composite score or subscale level.

Within this chapter, the OKS is the only PROM for which data can be considered at three levels, i.e. the item, subscale and composite score level. Different items are used for the calculation of either subscale, and the subscales can be used to calculate the composite score, so it is conceivable that insufficient items are available to calculate the composite score, but that one of the subscales can still be derived. This subscale could then be utilised in the MI model. Therefore, imputation at the subscale level is also investigated, and it is hypothesised that there are benefits in terms of accuracy when imputing at the subscale level compared to imputing at the composite score level when sufficient data is available for the calculation of one of the subscales. No benefit of applying MI at the subscale level

over MI at the composite score level is expected in unit-nonresponse scenarios, or where neither of the subscales can be estimated due to missing data.

4.3.2 Objectives for this chapter

This chapter aims to compare different MI approaches for handling missing PROMs data, i.e. imputation at the composite score, subscale (where appropriate) or item level, while also exploring the benefits and disadvantages of these approaches. By addressing these research objectives, this chapter addresses outstanding questions and clarifies some of the ambiguities around the application of MI to PROMs outcome data.

For scenarios where the analysis includes complete baseline data and PROMs outcome data for a single follow-up time point within an RCT context (i.e. data from the KAT study[92, 93] are utilised), this chapter aims to extend the existing research by:

- Applying MI at the item or composite score level for the OKS, SF-12 and EQ-5D-3L
- Considering MI at the subscale levels, where appropriate
- Expanding the existing research to consider the impact of the different MI approaches on the estimate of the treatment coefficient in regression models using a PROM as the dependent variable
- Investigating whether the imputation recommendations are consistent where either the PROMs composite scores or the treatment coefficients from a regression model are of interest
- In a regression context, assessing the advantages and disadvantages of using CCA compared to the different imputation approaches
- Considering the performance of mean imputation of up to two missing items in the OKS in line with the scoring manual[42] to using the MI approaches for all missing data

- Consider whether results are consistent across different PROMs, or whether findings are dependent on the PROMs used
- Consider a wide range of sample sizes and proportions of missing data
- Provide guidance for researchers on how to address missing PROMs outcome data in a variety of real-life RCTs scenarios and how to balance the potentially complex MI models with steps to improve the probability of convergence of the MI models

4.4 Simulation methodology

4.4.1 Rationale for using simulations

The use of a single study to generate reliable performance estimators for statistical methods is considered insufficient, as it is unknown how reliable findings are. Instead, simulation studies enable researchers 'to assess the performance of current and novel statistical methods in pre-defined scenarios'[114] in relation to a known truth[113]. Here, the true results are calculated from a full dataset, before the simulation of missing data. Simulations are computationally intensive processes, during which the methods under investigation are applied repeatedly and compared to the 'true results', thus enabling a comparison of different approaches.

To achieve consistency and appropriate quality within the development and reporting of simulation studies, extensive guidance has been developed[113] and is adhered to in the analyses.

4.4.2 General simulation procedures

All programming for this chapter is performed using the statistical software Stata version 14[115]. 1,000 valid results are obtained for each simulation scenario, where feasible.

4.4.2.1 Simulation scenarios to be investigated

Multiple imputation at the composite score, subscale (where applicable) and item level are simulated in the following scenarios:

The percentage of participants with missing data are set to 5%, 10%, 20%, and 40%, in line with the missing data patterns observed in Chapter 3. The different sample sizes considered within this simulation study are 100, 200, 500, as well as the total number of

observations in the complete cases datasets, which varies depending on the PROMs (797 for the SF-12, 1030 for the OKS and 1160 for the EQ-5D-3L).

4.4.3 Generation of the datasets to be used in the simulation

4.4.3.1 Base case datasets

The complete cases subset of the KAT study introduced in Chapter 3 form the basis for the simulated datasets, i.e. only the data of the participants for whom all relevant variables are observed are utilised in the simulations, circumventing the need to impute baseline data. This dataset is referred to as the base case dataset.

For the simulation of smaller datasets, the required numbers of observations are sampled from the base case dataset (sampling without replacement) using the 'sample' command in Stata. This means that the sample sizes investigated within this chapter are limited to the size of the complete cases subsets for each PROM. Using the resampling approach means that the dataset is different at each iteration of the simulation for the smaller sample sizes. However, the use of pre-defined seeds that determine the random number generation in Stata ensure that the datasets are reproducible.

4.4.3.2 Simulation of missing data within the simulation datasets

Missing data are imposed onto PROMs items following the algorithm proposed by van Buuren et al[116]. Briefly, this method allows researchers to set the overall percentage of observations with some missing PROMs data, and to specify the patterns of missing data, as well as the MAR mechanism. The missing data patterns are simulated by specifying the probability with which each item within the PROM is missing. The MAR mechanism is simulated by specifying a set of variables that determine the likelihood of data being

missing. These patterns may be based on those observed in the dataset under investigation or related studies, as well as any other scenarios to be examined. The algorithm therefore allows the simulation of realistic MAR scenarios, which take into consideration the complex relationship between missing data patterns and other variables, unlike more simplistic methods of MAR generation.

More precisely, the steps shown in Figure 4-1 are followed to impose missing data into the relevant data set, as introduced by van Buuren[116] and also outlined in publications by Yu et al[117] and Simons et al[112]. Notation is kept consistent with that presented in existing publications. Missing data is imposed at the item level, based on the desired missing data patterns. The variables included in the logistic regression model (described in step 2 in Figure 4-1) to generate missing data include treatment allocation, age, baseline PROM, height, ASA physical status (a system to assess patients' fitness prior to surgery)[101, 102] and centre size (three categories). The code for the generation of the observed missing data patterns is shown in Appendix 7.

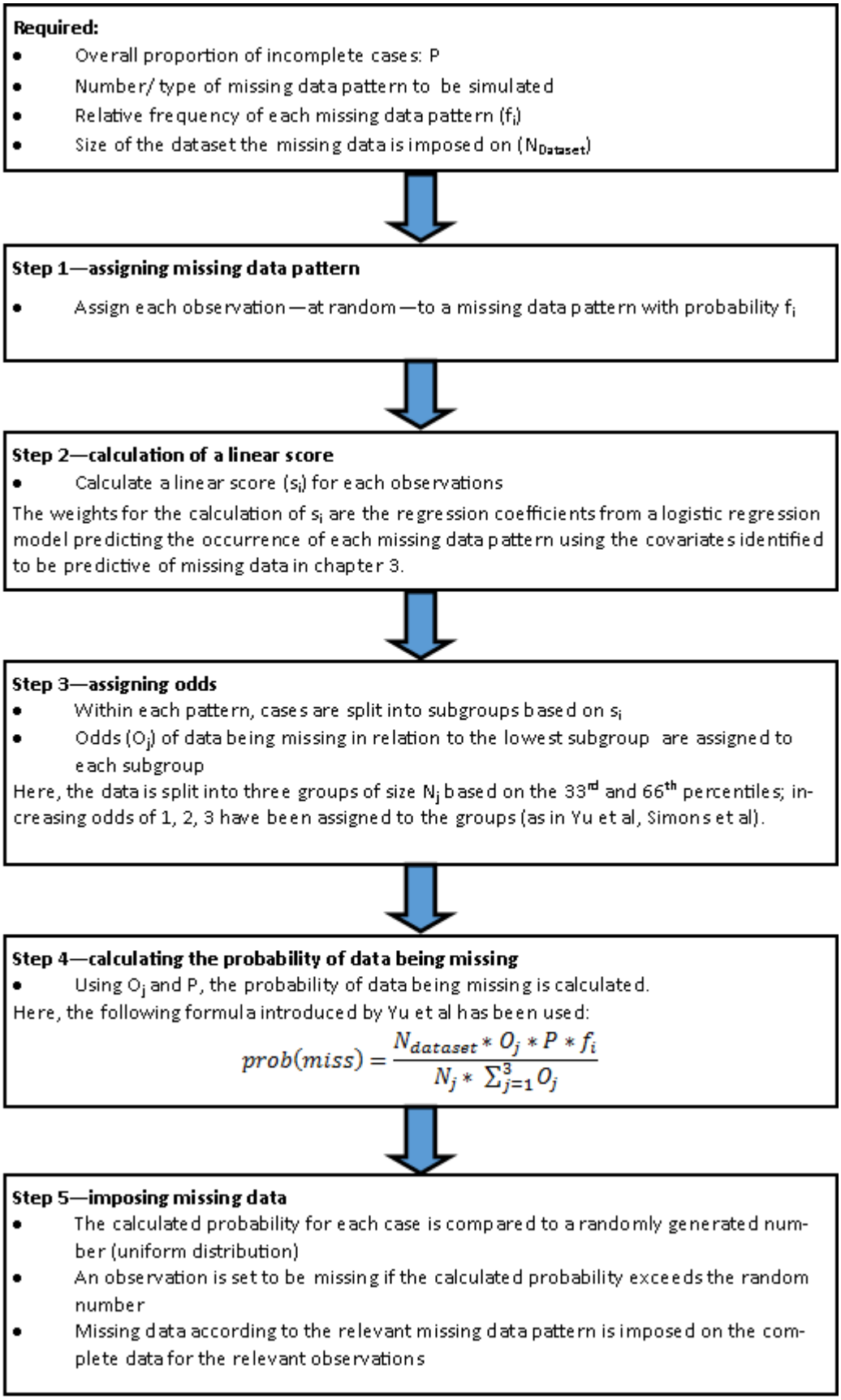


Figure 4-1: Depiction of the algorithm for the simulation of missing PROMs data within the complete cases dataset

4.4.3.3 Comments on the generation of missing data – formulae used

In the simulation model used in this chapter, whereby increasing odds of 1, 2 and 3 are assigned to three equally spaced intervals based on the s_j scores within each missing data pattern, the formula to calculate the probability of missing data for each case can be simplified as follows:

1) $N_{dataset} * f_i$ equals the number of cases assigned to the relevant missing data pattern (N_i)
2) $N_i = 3 * N_j$ as the cases within each missing data pattern are divided in three equally spaced intervals based on the 33rd and 66th percentile of the s_j scores
3) $\sum_{j=1}^{3} O_j = 6$ (based on the assigned odds being 1, 2, 3)

Applying the above listed condition, the formula for the probability of missing data within each of the odds subgroups can be simplified as follows:

$$prob(miss) = \frac{N_{dataset} * O_j * P * f_i}{N_j * \sum_{j=1}^{3} O_j}$$

Applying 1):

$$prob(miss) = \frac{N_i * O_j * P}{N_j * \sum_{j=1}^{3} O_j}$$

Applying 2):

$$prob(miss) = \frac{3 * N_j * O_j * P}{N_j * \sum_{j=1}^{3} O_j} = \frac{3 * O_j * P}{\sum_{j=1}^{3} O_j}$$

Applying 3):

$$prob(miss) = \frac{3 * O_j * P}{6} = \frac{O_j * P}{2}$$

The simplified formula indicates that the specified overall percentage of missing data is applied to each missing data pattern, adjusted for the odds depending on the s_j scores. These formulae clarify that on average, the required percentage of missing data is imposed

within each missing data pattern. Here, those within the lowest odds subgroup (odds = 1) are assigned a probability of missing data of half the rate of the overall percentage of missing data, while those in the highest odds subgroup (odds = 3) are assigned a probability of missing data of 1.5 times the overall percentage of missing data. The probability of missing data for those within the middle odds subgroup (odds = 2) is the same as the specified overall percentage of missing data.

The performance of formulae were compared in additional simulation work and sufficiently similar results were produced, with the percentages of participants with simulated missing data within each of the categories differing at the third or fourth decimal place. These slight differences can stem from the fact that the use of percentiles may not always result in odds subgroups of equal size being generated, depending on the values of s_j, and the impact of this difference will be more notable within missing data pattern with a lower relative frequency.

In other projects involving the simulation of missing data[52], a different formula has been used to calculate the probability of a specific case having missing data, resulting in slightly different probabilities for the odds categories. Van Buuren et al[116] did not clearly specify a formula in their paper, but conversations with the authors have shown that formulae similar to those referred to by Brand et al were used[118].

4.4.4 Application of multiple imputation in the datasets after the simulation of missing PROMs data

Within this chapter, missing data is imputed using multiple imputation by chained equation (MICE)[58], a practical approach to impute missing data for a number of variables, allowing the imputations for each variable to draw on information imputed for missing observations in other variables. In short, where several variables contain some missing observations, these are initially imputed by observed values from this variable (simple random sampling with replacement). The first variable with missing values is then regressed on all other variables, including all other variables with missing observations and additional variables included in the imputation model. This regression model only includes individuals with observed data for this first variable; missing data in this variable is subsequently imputed by draws from the posterior predictive distribution generated by the regression model. Missing data for subsequent variables is imputed in a similar manner, each time utilising the previously imputed data for other variables. Once imputations for all variables have been generated in this manner, the first cycle of the imputation process has been completed. Several cycles are repeated until the imputed values stabilise, and one set of imputed values has been generated. This procedure is repeated until the required number of imputations are obtained.

Predictive mean matching is used in combination with the above process[58]. This process restricts the choice of imputed values to values that were observed within the relevant variables. In the context of imputing PROMs data, predictive mean matching has the benefit that imputed scores lie within the range of valid responses (i.e. do not go beyond the possible range, i.e. from 0 to 48 for the OKS). Similarly, predictive mean matching ensures that valid imputed scores are obtained (i.e. no decimal places where a PROMs

scoring manual only results in integer scores). Predicted mean matching has also been recommended when the imputed variable may not be normally distributed or may have a non-linear relationship with the covariates[58]. This is deemed to be important particularly for the EQ-5D-3L, which is often bimodal[119].

Predictive mean matching can be implemented by picking the observed value that is the closest match in terms of the predicted scores, or the imputed value can be from a random draw of a pre-specified number of observations closest to the predicted value. In this thesis, the former approach (nearest neighbour) is used.

In this simulation work, the imputation models include all variables used in the generation of the MAR data (section 4.4.3.2). Additionally, gender is included in the MI models, as it is part of the analysis model[58] (described in section 4.4.5).

The baseline PROM composite scores are included in all MI models; additional data included for imputations at the subscale and item level, as discussed below. Imputations are performed separately by randomised treatment, where feasible. This approach allows factors such as the distribution of outcomes, their variance and relationship with any of the covariates to differ between treatment arms[120].

The sample size of the base cases for each simulation are defined by the number of participants for whom the relevant PROMs scores are available at baseline and follow-up, and therefore differ between the PROMs. For all imputation models in the base cases, the number of imputations is set to 50. A rule of thumb proposes that the number of imputations should be as least equal to the percentage of incomplete cases[58], while larger numbers of imputations also generate a larger degree of reproducibility.

In the subsequent simulations performed for smaller sample sizes, described in section 4.4.2.1, the number of imputations is reduced to the percentage of missing data within the

relevant simulation, in line with the recommendations[58]. The number of imputations for some of the exploratory simulations, which vary the observed missing data patterns, are reduced to 10. Details on the number of imputations used are provided in the results section.

To further reduce the length of time required to run the simulation models, the maximum number of iterations for each imputation model was set to 1,000. The length of time required for the simulation models to run are also discussed in the results.

All imputations are performed using Stata's *mi impute* command.

4.4.4.1 Scenario considering imputation at the score level

Full PROMs scores are calculated according to the relevant scoring manuals (and also without mean imputations for the OKS) based on the newly created PROMs items with some simulated missing data. As such, the calculations result in some PROMs scores being missing. An MI model for continuous data using predictive mean matching using the closest match is applied.

4.4.4.2 Scenario considering imputation at the subscale level

Similar to the process described in section 4.4.4.1, subscales are calculated based on the PROMs items with imposed missing data. Subscales are considered as continuous data within this simulation study, and are imputed using a MICE approach. Imputing the subscales jointly allows for missing subscales to be imputed based on the values of other available or imputed subscales at follow-up. In addition to the described imputation model, the imputation model for the subscales also includes the baseline values of each subscale.

For the OKS, the multiply imputed subscales are combined to generate composite scores for each of the multiply imputed datasets, in line with the proposed scoring[111]. The MCS and PCS scores for the SF-12 are imputed jointly, but the two components are then considered as individual composite scores.

4.4.4.3 Scenario considering imputation at the item level

Data for the levels of each PROM item are imputed. The OKS items have five levels each, the EQ-5D-3L items have three levels each and the SF-12 items have either three or five levels on a Likert scale. The levels are clearly related to factors such as pain and symptom severity, or physical functioning and can be considered as ordered categories. Therefore, ordered logit models are used to impute items, a method which was shown to have fewer problems with convergence than a multinomial logit model in a previous study[112], which was confirmed in preliminary work.

As for the subscales, a MICE approach is used to allow for missing items to be imputed based on the values of other available or imputed items. In addition to baseline composite scores and subscales (where available) the imputation model for the composite scores (section 4.4.4.1), the imputation model for the items also includes the baseline values of each item, where feasible. The multiply imputed items are used to calculate the composite PROMs scores for each of the multiply imputed datasets.

In the case of non-convergence of the imputation models, discussed in detail in section 4.4.6.4, additional simulations were run to ensure that results for 1000 valid simulations were obtained where possible. A maximum of 11,000 iterations per simulation scenario were performed.

4.4.5 Data generated from the simulation models for the comparison of MI approaches

Each simulation is programmed to store information on the sample size used, the level at which imputations are undertaken (composite score, subscale and item), as well as the percentage of missing data imposed and the number of imputations used.

In addition, the following data is saved to be used in the assessment of the different approaches:

4.4.5.1 Data generated from the full dataset

For each full dataset i.e. the dataset of the relevant sample size prior to missing data being simulated, the mean and standard error of the composite scores are calculated. In addition, the full dataset is also used to calculate the treatment effect and associated standard error in the regression model using the composite scores as the independent variable, adjusting for the baseline measure of the outcome variable as well as a pre-specified set of covariates (i.e. randomised treatment, age, gender and relevant baseline PROM).

4.4.5.2 Data generated from the dataset with imposed missing data

Means and standard errors of the composite PROMs are calculated using a CC approach. The composite scores with the imposed missing data are used to generate the treatment effects and associated standard errors based on a CCA regression model; the number of observations used in the CCA is also stored.

4.4.5.3 Data generated from the multiply imputed PROMs data

Rubin's adjusted estimates for the mean composite scores and treatment effects with corresponding standard errors are calculated for the PROMs scores from the multiply

imputed data for the different imputation approaches, using Stata's *mi estimate* command. To do this in the scenarios where the composite scores are calculated after MI of the items or subscales, these variables need to be declared to be multiply imputed data (using Stata's mi import command).

Instances of non-convergence were also collected for the imputations at item level.

4.4.5.4 Data generated to assess each model's performance

Based on the data generated from each simulation, the following estimates are generated to assess the performance of each MI approach under the different scenarios, in accordance with recommendations by Burton et al[113]:

- Root mean square error (RMSE): $\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{\theta_i} - \theta\right)^2}$
- Mean absolute error (MAE): $\frac{1}{N}\sum_{i=1}^{N}\left|\widehat{\theta_i} - \theta\right|$

  Where N denotes the number of simulations run (here 1,000 where feasible), θ the true value for the estimate of interest and θ_i is the estimate of interest obtained from the ith simulation.

The RMSE, which 'penalises variance as it gives errors with larger absolute values more weight than errors with small values'[121], is presented in the main text of the chapter; results for the MAE are provided in Appendix 8.

4.4.6 Simulation scenarios

Separate simulation programmes were written for each of the different imputation approaches. However, consistent seeds (dictating the use of random numbers within the

simulations) ensure that the same underlying data is used for the different simulations and that results are reproducible and comparable across the different scenarios.

4.4.6.1 Missing data patterns simulated

Initially, the patterns of missing data follow those observed for the five-year follow-up in Chapter 3. For the simulations considering missing data in the OKS, the patterns of missing data as shown in Table 4-1 are imposed on the complete cases subset of the KAT trial as described in section 4.4.3.2. Only the eight most commonly observed patterns are used; the other combinations of missing items relate to a total of 29 participants, with each pattern only observed in one or two participants. Therefore, these patterns are not commonly observed within this trial and not replicated in the simulation study. The percentages used in the simulations are hence based on the subset of participants falling into one of the eight relevant categories of missing data patterns (the denominator for the calculations is 334 instead of 363 participants).

Additionally, unit-nonresponse and a scenario where 70% of missing OKS data is due to item-nonresponse are also considered.

The above described simulation scenarios are based on the base case dataset, which reflects the data as observed in the KAT study, which did not show a significant difference between the two trial arms[28]. One additional scenario involves the introduction of a five point treatment effect through adding up to three points to the outcome scores in the patellar resurfacing arm, and subtracting up to three points from the OKS outcome scores in the no patellar resurfacing arm, as appropriate, i.e. such that the upper and lower range of the OKS was not exceeded.

Table 4-1: Missing data patterns imposed on the OKS in the complete cases subset of the KAT trial

Missingness patterns	Total	True %	% used in simulation	Cumulative %
Unit non-response	244	67.22%	73.05%	73.05%
Only item 7 missing	52	14.33%	15.57%	88.62%
Only item 4 missing	11	3.03%	3.29%	91.92%
Only item 6 missing	9	2.48%	2.69%	94.61%
Only item 9 missing	7	1.93%	2.10%	96.71%
Only item 10 missing	5	1.38%	1.50%	98.20%
Only item 1 missing	3	0.83%	0.90%	99.10%
Only item 12 missing	3	0.83%	0.90%	100.00%

Total number of observations: 1526. Of those, 363 (23.79%%) have at least one missing item (the above patterns cover 363 cases referring to the eight most commonly observed missing data patterns, covering 92% of participants with any missing data - other missing data patterns occurred too infrequently to be included in the simulation work)

Using a similar approach described above, Table 4-2 and Table 4-3 show the missing data patterns observed for the EQ-5D-3L, for which only the six most commonly observed patterns are used in the simulation, and the SF-12 at the five year follow-up:

Table 4-2: Missing data patterns imposed on the EQ-5D-3L in the complete cases subset of the KAT trial

Missingness patterns	Total	True %	% used in simulation	Cumulative %
Unit non-response	240	87.27%	87.91%	87.91%
Only item 5 missing	14	5.09%	5.13%	93.04%
Only item 1 missing	7	2.55%	2.56%	95.60%
Only item 4 missing	5	1.82%	1.83%	97.44%
Only item 3 missing	4	1.45%	1.47%	98.90%
Only item 2 missing	3	1.09%	1.10%	100.00%
Items 2 and 5 missing	1	0.36%	0.00%	
Items 4 and 5 missing	1	0.36%	0.00%	

Total number of observations: 1526. Of those, 275 (18.02%) have at least one missing item (only the above eight MD patterns observed in dataset – of those, the top six are used in the simulation, as the other patterns were observed too infrequently and could not be used in the logistic regression models)

Table 4-3: Missing data patterns imposed on the SF-12 in the complete cases subset of the KAT trial data

Missingness patterns	Total	True %	% used in simulation	Cumulative %
Unit non-response	224	47.26%	56.14%	56.14%
Only item 2b missing	81	17.09%	20.30%	76.44%
Only item 4b missing	26	5.49%	6.52%	82.96%
Items 2b and 3b missing	18	3.80%	4.51%	87.47%
Only item 3b missing	16	3.38%	4.01%	91.48%
Items 2b, 3b and 4b missing	14	2.95%	3.51%	94.99%
Items 2b and 4b missing	13	2.74%	3.26%	98.25%
Only item 6c missing	7	1.48%	1.75%	100.00%

Total number of observations: 1422. Of those, 474 (33.33%) have at least one missing item (the above patterns cover 399 cases referring to the eight most commonly observed missing data patterns, covering 84% of participants with any missing data - other missing data patterns occurred too infrequently to be included in the simulation work)

4.4.6.2 MAR mechanisms simulated

The MAR mechanisms are simulated using the algorithm by van Buuren[116], as discussed above. In summary, logistic regression models are fitted to a binary variable indicating whether or not a participant fell into each of the above listed missing data patterns, using the covariates described in section 4.4.4. The regression coefficients estimated from these logistic regression models are then used to calculate the linear scores within the algorithm to generate missing data.

4.4.6.3 Comments on the generation of missing data – MAR checks

Checks were performed to ensure the simulated missing data followed a MAR mechanism. Using the above described approach, missing data was simulated for the OKS data.

Table 4-4 below shows how baseline characteristics and subsequently the OKS at five years differ within the cases assigned to the odds subgroups. It can be concluded that the algorithm works as expected, and that MAR data with respect to the specified variables has been generated. Additional checks showed that the proportion of participants with missing data was in line with the specified level of overall missingness as requested within the algorithm.

Table 4-4: Baseline characteristics, and subsequently the OKS at 5 years, within odds subgroups (observed missing data patterns, OKS)

Odds subgroup	Allocated to patellar resurfacing	Age (years)	Baseline OKS	Height	Proportion of patients with ASA grade 1	Proportion of patients with ASA Grade 3	Proportion of patients at medium site	Proportion of patients at large site	OKS at 5 years
1	0.553	66.472	25.201	172.083	0.304	0.051	0.267	0.555	37.796
2	0.521	69.054	18.074	166.012	0.210	0.090	0.390	0.567	34.853
3	0.464	71.739	12.868	158.807	0.115	0.308	0.404	0.457	31.556

4.4.6.4 Steps to reduce convergence problems for MI at the item level

Problems with convergence for the imputation at the item level of the EQ-5D-3L have been observed previously in the work by Simons et al[112], especially for smaller sample sizes and higher percentages of missing data. Therefore, it was expected that the MI models applied at the item level for the OKS and SF-12, i.e. questionnaires with an additional seven items and additional levels within the items, would result in a higher probability of non-convergence.

Non-convergence is more likely to occur during the imputations at item level because these models are more complex than the imputation models at the composite score or subscale level, as up to 12 items are imputed simultaneously. Thus, the ordinal logit models contain large numbers of categorical variables, leading to model estimates becoming less stable where the categories contain small numbers; this problem is amplified when imputations are performed separately by treatment arm or on small sample sizes.

The following steps were taken to improve the chances of convergence within the simulation models looking at MI at the item level, but are applicable more widely in the implementation of logistic regression models:

- **Allowing for occasional instances of non-convergence:**

  In MI models, missing values are initially replaced by random draws (with replacement) from the observed values for this variable. The use of random draws means that sometimes an iteration may not converge as outlying starting values have been selected by chance in a low number of instances. Also, a complex model may converge in the majority of cases, but fail in a low proportion of the required imputations. In the current Stata code (*mi impute*, Stata 14.1), all valid imputations would be discarded at the first instance of non-convergence. If, as in the base case simulations, as many as 50 imputations are requested, this could lead to a large number of simulations being labelled as non-convergent even if some valid imputations could be obtained.

  To circumvent the loss of valid imputation data, the Stata code has been altered in the simulations used here. Instead of using the 'add(50)' option within the *mi impute* command to request 50 sets of imputed values, the imputation command is run requesting only one new imputation at a time (i.e. add(1)). This command is embedded in a loop that runs until the required number of valid imputations is obtained. If the imputations fail for half of the iterations requested, the imputation model for this iteration of the simulation is classed as failed.

  This approach was found to generate valid imputed datasets for complex imputation models where the conventional *mi impute* command would have failed to produce valid results.

- **Reducing the number of (categorical) covariates in the imputation model:** As discussed above, non-convergence may be caused by too many categorical covariates, especially where there are small numbers in some of the categories. The current literature suggests that categories can be combined, but this was felt to be inappropriate if the PROMs item themselves were to be imputed.

  An alternative approach is to simplify the imputation model by reducing the amount of covariates used. Details of the simplifications are provided in the results section (section 4.5).

- **Running imputations jointly for both treatment arms**

  Imputations were performed separately by treatment arm where feasible. However, this effectively reduces the sample size for the imputation models, which again can lead to convergence issues due to the inclusion of a large number of categorical variables in the models. Therefore, not all imputations were performed separately by treatment arm; details of the simplifications are provided in the results section (section 4.5).

- **Re-categorising variables to enhance the robustness**

  In all statistical models, the certainty around estimates increases with increasing sample size. Similarly, the estimates of categorical variables, which are always compared to a reference category are more robust if there are sufficient numbers within this reference category. Models may also fail if low counts in the reference category lead to perfect prediction, of if insufficient numbers are available to determine a regression coefficient. The *mi impute* command in Stata chooses the reference category by default, and there was no option to change this reference category within the command at the time this thesis was written. Therefore, a decision was made to change the order of the categories manually, where required. Of course, the categories need to maintain clear ranking, i.e. maintain the characteristics that define them as ordinal variables if they are used as the dependent variables in the imputation model based on an ordinal logit model (i.e. the PROMs items measured at follow-up). For this reason, the items could not simply be re-categorised to ensure the highest number of counts would be in the reference category; instead, data in the base case dataset was investigated and where the category with the lowest numeric code (i.e. the variable to be used as the default reference category) had lower counts than then category with the largest numeric code, the coding of the variable was inversed.
  For other categorical covariates, such as baseline items or ASA grade in the simulations considered within this chapter, the ordinal structure of the variables does not need to be maintained, and the data can be re-categorised to ensure the highest counts fall into the reference categories.

Some of the proposed steps to increase convergence are applicable to all imputation models (i.e. re-categorising variables and allowing occasional instances of non-convergence). However, other suggestions, such as running imputations jointly for both treatments arm, or reducing the amount of covariates in the model (thus possibly reducing its predictive power), may affect the performance of the imputation model. Consequently, a balance needs to be struck between making compromises in the set-up of the MI model while achieving suitable convergence of the model. The choice in approaches depends on the data available, as well as the researcher's opinion on which components are indispensable for the imputation of missing data.

4.4.6.5 Skipping simulations with insufficient missing data

The generation of missing data within these simulation models depends on the use of random numbers. While the overall percentage of missing data is, on average, equal to that specified in the simulation models, the actual percentage of cases with missing data differs slightly between simulations, and may fall slightly above or below the specified percentage, due to the use of random numbers.

In the analyses reported here, in some simulations for small sample sizes (n = 100) and low percentages of missing data (5%), no unit-nonresponse missing data was generated by chance. This caused problems with the simulation MI code especially for the item imputation, as the code expected some missing data in all variables to be imputed. Therefore, these scenarios were replaced with another draw.

4.5 Results from the simulation study

4.5.1 Number of imputations used for each simulation scenario

Table 4-5 provides an overview of the number of imputations used for each simulation. For the simulations of primary interests, i.e. those utilising the observed missing data patterns and observed data, 50 imputations were used for all MI models to impute composite scores and/or subscales, as well as for the base case item imputation models. Generally, item imputations for the lower sample sizes were in line with the percentage of data simulated to be missing. This principle was applied to all simulations considering a unit-nonresponse patterns for the OKS, and 10 imputations were used for the simulations considering the dataset with a simulated five point treatment effect in the OKS.

Table 4-5: Overview of number of imputations used in the different simulation scenarios

	Score imputation (all)	Subscale imputation (all)	Item imputation	
			Base cases	Lower sample sizes
OKS – observed missing data patterns	50	50	50	In line with % of data missing
OKS – unit-nonresponse	In line with % of data missing	In line with % of data missing	In line with % of data missing	In line with % of data missing
OKS – 70% item nonresponse	50	50	In line with % of data missing	In line with % of data missing
OKS – five point treatment effect	10	10	10	10
EQ-5D-3L – complex simulations	50	N/A	50	In line with % of data missing
EQ-5D-3L – simplified simulations	50	N/A	50	In line with % of data missing
SF-12 – complex simulations	N/A	50	50	In line with % of data missing
SF12 – simplified simulations	N/A	50	50	In line with % of data missing

4.5.2 Feasibility of the simulation models

Valid simulation results have been obtained for all simulation models imputing data at the composite score and subscale level. However, the complexity of the item MI models was magnified, as discussed above, and therefore increased levels of non-convergence were observed. To achieve improved convergence rates of the models, the following simplifications were applied:

- OKS: baseline PROMs items were excluded from the MI models, and the MI model was not run separately by treatment arm; instead, the treatment variable was added into the MI model as a covariate
- EQ-5D-3L: Two simulations were run, both separately by treatment arm; one simulation included baseline PROM items, the other excluded them. Results from the different models were compared.
- SF-12: baseline PROMs items were excluded from the MI model. Two simulations were run – one separately by treatment arm, one including the treatment variable in the MI model as covariate. Results from the different models were compared.

Table 4-6 to Table 4-13 show the percentage and number of simulations during which the imputation models failed to converge, together with the number of valid results obtained using the conditions lined out in section 4.4.6.4. A maximum of 11,000 simulations were run to replace simulations that did not converge.

It can be observed that increased instances of non-convergence are encountered both with decreasing sample size and increasing percentages of missing data for every given item MI model. These results indicate that complex MI at the item level may not be practically feasible in such scenarios. Simulations skipped due to insufficient missing data (only up to 0.45% for the OKS scenario of a sample size of 100 and 5% of missing data), are not shown separately.

In Table 4-6 to Table 4-9, the convergence rates for the item imputations of the different OKS scenarios can be seen. Table 4-7 shows the convergence rates for the OKS imputations, whereby the missing data follows the observed missing data patterns, i.e. approximately 73% of the missing data are due to unit-nonresponse. All missing data follows a unit-non-response mechanism in the data underlying the simulations presented in Table 4-6. It can be seen that non-convergence rates are lower in the latter scenario. It is possible that for the unit-nonresponse scenarios, the imputation model converges more easily as all missing data follow the same patterns and the ordinal logit models are therefore more predictive of the outcomes to be imputed. Table 4-8 shows the convergence rate under a simulation of 70% item nonresponse. Convergence rates are similar to those observed in the OKS simulations using the observed missing data patterns.

Table 4-9 shows the convergence rate when a five point treatment difference in terms of the OKS is simulated; the treatment effect for the patellar comparison in the KAT trial at five years was less than one point and statistically non-significant[93]. Non-convergence rates

here are higher than in the original simulations. The reason for this is likely to be related to the fact that by increasing five item scores by one point each, the counts within the lower item categories decrease, thus leading to convergence problems within the ordinal logit models.

Table 4-6: Percentage of imputation models that failed to converge (and number of valid results obtained) for the item imputations: OKS simulations

	Sample size			
% of cases with missing data	100	200	500	1030
5%	88.22% (1,000)	40.08% (1,000)	21.63% (1,000)	0.99% (1,000)
10%	95.62% (481)	50.67% (1,000)	24.87% (1,000)	3.94% (1,000)
20%	99.76%* (59)	64.48% (1,000)	27.80% (1,000)	12.51% (1,000)
40%	100%** (0)	99.65%* (38)	41.42% (1,000)	28.01% (1,000)

**The simulation was stopped after the initial 11000 iterations; for these scenarios, 1000 valid simulations are not available.*
***No valid results were obtained.*

Table 4-7: Percentage of imputation models that failed to converge (and number of valid results obtained) for the item imputations: OKS unit-nonresponse simulations

	Sample size			
% of cases with missing data	100	200	500	1030
5%	89.69% (1,000)	22.72% (1,000)	8.84% (1,000)	0% (1,000)
10%	95.32%* (515)	25.21% (1,000)	9.34% (1,000)	0.10% (1,000)
20%	99.55%* (50)	32.11% (1,000)	11.03% (1,000)	0.40% (1,000)
40%	100%** (0)	59.95% (1,000)	15.97% (1,000)	3.75% (1,000)

**The simulation was stopped after the initial 11000 iterations; for these scenarios, 1000 valid simulations are not available.*
***No valid results were obtained.*

Table 4-8: Percentage of imputation models that failed to converge (and number of valid results obtained) for the item imputations: OKS 70% item non-response simulations

	Sample size			
% of cases with missing data	100	200	500	1030
5%	87.95% (1,000)	35.44% (1,000)	20.00% (1,000)	0.70% (1,000)
10%	94.07%* (652)	45.62% (1,000)	23.37% (1,000)	1.57% (1,000)
20%	99.12%* (97)	56.95% (1,000)	27.48% (1,000)	5.21% (1,000)
40%	99.99%* (1)	78.38% (1,000)	28.57% (1,000)	10.39% (1,000)

**The simulation was stopped after the initial 11000 iterations; for these scenarios, 1000 valid simulations are not available.*
***No valid results were obtained.*

Table 4-9: Percentage of imputation models that failed to converge (and number of valid results obtained) for the item imputations: OKS simulations with five point treatment difference

	Sample size			
% of cases with missing data	100	200	500	1030
5%	82.50% (1,000)	54.15 (1,000)	42.46% (1,000)	24.30% (1,000)
10%	92.17%* (861)	59.89% (1,000)	43.57% (1,000)	30.60% (1,000)
20%	99.30%* (77)	65.75% (1,000)	45.12% (1,000)	40.97 (1,000)
40%	100%** (0)	95.34%* (513)	53.70% (1,000)	47.75% (1,000)

**The simulation was stopped after the initial 11000 iterations; for these scenarios, 1000 valid simulations are not available.*
***No valid results were obtained.*

Table 4-10 and Table 4-11 show the non-convergence rates for the EQ-5D-3L item level imputations. Initially, item level imputations included baseline items as explanatory variables, as they were thought to be important to predict items at the follow-up time point. However, convergence rates of the models were low, especially for sample sizes of 200 or less (see Table 4-10), and therefore further simulations were run where the baseline items were excluded from the imputation model (Table 4-11). Both simulations run imputations separately by treatment arm.

Table 4-10: Percentage of imputation models that failed to converge (and number of valid results obtained) for the item imputations: EQ-5D-3L simulations adjusting for baseline items and running simulations separately by treatment arm

	Sample size			
% of cases with missing data	100	200	500	1160
5%	99.56%* (96)	63.13% (1,000)	13.79% (1,000)	0.79% (1,000)
10%	99.93%* (8)	74.13% (1,000)	17.08% (1,000)	3.57% (1,000)
20%	100%** (0)	92.19% (1,000)	22.78% (1,000)	10.71% (1,000)
40%	100%** (0)	99.89%* (12)	40.12% (1,000)	18.96% (1,000)

**The simulation was stopped after the initial 11000 iterations; for these scenarios, 1000 valid simulations are not available.*
***No valid results were obtained.*

Table 4-11 Percentage of imputation models that failed to converge (and number of valid results obtained) for the item imputations: EQ-5D-3L simulations – simplified – no baseline items included

	Sample size			
% of cases with missing data	100	200	500	1160
5%	72.10% (1,000)	19.87% (1,000)	11.19% (1,000)	0.30% (1,000)
10%	83.75% (1,000)	25.48% (1,000)	14.24% (1,000)	1.86% (1,000)
20%	94.40%* (616)	45.30% (1,000)	17.29% (1,000)	6.80% (1,000)
40%	99.86%* (15)	85.56% (1,000)	23.55% (1,000)	17.36% (1,000)

**The simulation was stopped after the initial 11000 iterations; for these scenarios, 1000 valid simulations are not available.*

Table 4-12 and Table 4-13 show the non-convergence rates for the SF-12 item level imputations. Imputations including the baseline items were found to be infeasible and are not presented here. Table 4-12 shows results for the imputations when the imputation models were run separately by treatment arm. Due to issues with non-convergence, the imputation model was simplified to include the treatment variable as a categorical variable (Table 4-13), instead of running imputations separately by treatment arm. This improved convergence of the models; however, sufficient numbers of results could not be obtained for simulations with a sample size of 100, and the scenario with 40% of missing data for a sample size of 200.

Table 4-12: Percentage of imputation models that failed to converge (and number of valid results obtained) for the item imputations: SF-12 simulations

	Sample size			
% of cases with missing data	100	200	500	797
5%	100%** (0)	99.42%* (56)	51.53% (1,000)	4.58% (1,000)
10%	100%** (0)	99.94%* (7)	72.78% (1,000)	9.99% (1,000)
20%	100%** (0)	100%** (0)	90.84% (1,000)	29.58% (1,000)
40%	100%** (0)	100%** (0)	99.71%* (32)	81.08% (1,000)

**The simulation was stopped after the initial 11000 iterations; for these scenarios, 1000 valid simulations are not available.*
***No valid results were obtained.*

Table 4-13: Percentage of imputation models that failed to converge (and number of valid results obtained) for the item imputations: simplified SF-12 simulations, i.e. imputations not run separately by treatment arm

	Sample size			
% of cases with missing data	100	200	500	797
5%	91.87%* (663)	52.06% (1,000)	3.47% (1,000)	0% (1,000)
10%	98.17%* (198)	69.58% (1,000)	6.28% (1,000)	0.20% (1,000)
20%	99.85%* (16)	87.19% (1,000)	12.82% (1,000)	1.09% (1,000)
40%	100%** (0)	99.28%* (79)	34.73% (1,000)	8.42% (1,000)

**The simulation was stopped after the initial 11000 iterations; for these scenarios, 1000 valid simulations are not available.*
***No valid results were obtained.*

Simulations yielding fewer than 1000 valid results are excluded from the following comparisons.

4.5.3 Comparison of the different MI approaches

In this section, the results of the different imputation approaches are presented. Mostly, the comparisons focus on the RMSE, and the presentation of the standard errors (SEs) of these estimates compared to those from the full dataset without any missing data. The graphs show the RMSE (or SEs) on the y-axis. The x-axis shows the different sample size scenarios simulated, and the differently shaded lines on the graphs represent results for the different approaches to handling missing data, i.e. MI at the composite score level, subscale level (where appropriate) and item level. For the RMSE of the treatment coefficients, the RMSEs for a CCA are also shown. Separate graphs present the different percentages of missing data that were simulated. Similar graphs showing summaries of the MAE for the different MI approaches are shown in Appendix 8.

A trend that can be observed in all of the results is that the RMSE (and MAE) increases with increasing percentages of missing data, as well as with decreasing sample size. This again emphasises the importance of preventing the occurrence of missing data prospectively[18].

Results for the different PROMs measures and MI approaches are presented below.

4.5.3.1 Results for the OKS simulations – using the observed missing data patterns

In this section, the results for the different imputation approaches for missing OKS at five years are presented. Different simulation scenarios varying the sample size and overall percentage of missing data are considered, based on the missing data patterns shown in Table 4-1.

Considering the OKS composite score estimates

Table 4-14 shows the simulation results for the OKS scenario base cases, i.e. a sample size of 1030. For each scenario considering a different proportion of missing data, the smallest RMSE and MAE is highlighted in bold. Generally, for this sample size there is little difference in the RMSEs and MAEs between MI at the composite score, subscale and item level, with a very small advantage of MI at the subscale level for all scenarios. Differences between the different approaches are amplified as the proportion of missing data increases.

The OKS scoring manual recommends the mean item imputation for up to two missing items; scores cannot be calculated if more than two items are missing[42]. Similarly, one missing item for subscales can be substituted via mean imputation. In this simulation, the scoring manual was followed, but a variation of the simulations was also run whereby composite scores and subscales were considered missing when any items were unavailable.

It can be seen that marginally more accurate results are obtained for the composite score imputation for 5% and 10% of missing data when following the scoring manual. However, for scenarios of 20% of missing data or more, better results are achieved when not using item mean imputation. When imputing at the subscale level, not utilising mean imputation always results is slightly lower bias in terms of RMSE and MAE.

As mentioned above, the differences observed between the true OKS values and those obtained from the various imputation approaches are small. The maximum differences displayed in Table 4-14 are less than one point on the OKS, which ranges from 0 to 48. The observed differences are lower than the between group minimal important difference and the smallest possible detectable change, which are estimated to be five and four points,

respectively[12], and are therefore unlikely to be clinically relevant. The potential impact of these differences on the trial results is further discussed in section 4.6.1.

Table 4-14: Estimated OKS, RMSE and MAE for the OKS base case simulations (sample size = 1030): results and bias introduced for the PROMs composite score estimates

	Mean	SE	MAE	RMSE
True OKS	34.686	0.325		
5% missing data				
Score imputation - in line with scoring manual	34.707	0.333	0.0611	0.0756
Score imputation - no mean imputation	34.690	0.336	0.0667	0.0836
Subscale imputation - in line with scoring manual	34.710	0.333	0.0609	0.0760
Subscale imputation - no mean imputation	34.686	0.334	**0.0581**	**0.0724**
Item imputation	34.761	0.332	0.0853	0.1040
10% missing data				
Score imputation - in line with scoring manual	34.734	0.342	0.0927	0.1145
Score imputation - no mean imputation	34.697	0.347	0.0989	0.1227
Subscale imputation - in line with scoring manual	34.737	0.341	0.0929	0.1139
Subscale imputation - no mean imputation	34.687	0.342	**0.0831**	**0.1038**
Item imputation	34.824	0.339	0.1485	0.1750
20% missing data				
Score imputation - in line with scoring manual	34.799	0.360	0.1573	0.1958
Score imputation - no mean imputation	34.718	0.373	0.1546	0.1924
Subscale imputation - in line with scoring manual	34.810	0.359	0.1640	0.2021
Subscale imputation - no mean imputation	34.691	0.361	**0.1302**	**0.1628**
Item imputation	34.897	0.358	0.2427	0.2854
40% missing data				
Score imputation - in line with scoring manual	35.015	0.412	0.3649	0.4380
Score imputation - no mean imputation	34.813	0.456	0.3114	0.3902
Subscale imputation - in line with scoring manual	35.043	0.407	0.3833	0.4559
Subscale imputation - no mean imputation	34.695	0.412	**0.2379**	**0.2966**
Item imputation	34.812	0.423	0.3889	0.4793

Figure 4-2 shows the RMSE introduced into the estimate of the OKS composite score following the different MI approaches. It can be observed that imputation at the composite score and subscale level perform similarly in terms of RMSE. For larger proportions of

missing data, and smaller samples, imputation at the item level introduces more bias than the alternative MI approaches. Of note, a dip in the RMSE for the item imputation simulation scenario of 5% missing data and a sample size of 100 is shown in the graph. This is likely to be related to the high failure rate for these imputations (almost 90%); this resulted in a very selective sample being used, which differs from those for imputation at the composite score and item level by a lower mean OKS. Therefore, the observed dip is likely to be a consequence of the selective sampling, rather than a true benefit of MI at the item level. Similar features can be observed for the other simulation scenarios.

Insufficient results have been obtained for the item imputation for a sample size of 100, with 20% and 40% of missing data, as well as for the scenario considering a sample size of 200 with 40% of missing data.

The MAE, plots shown in Appendix 8.1 confirm these findings.

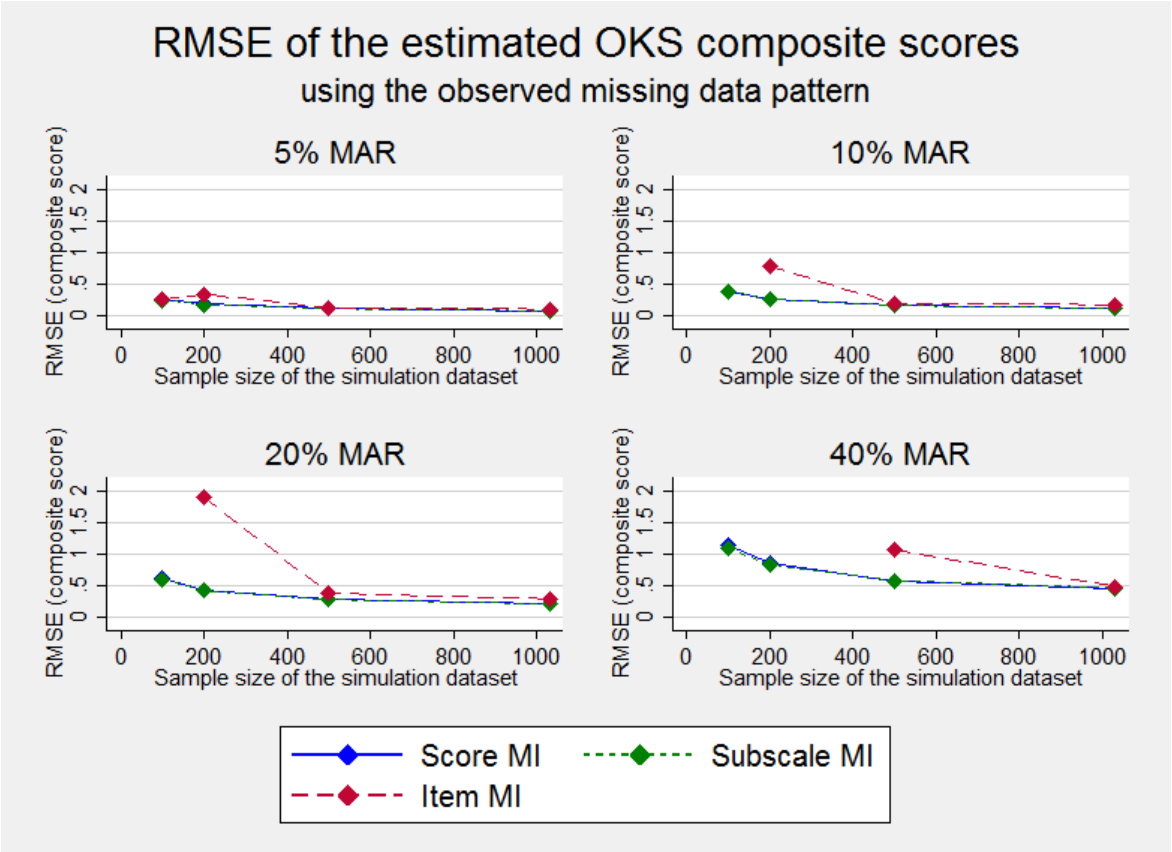


Figure 4-2: RMSE in the OKS composite score estimates

Figure 4-3 shows the SEs associated with the OKS composite score estimates following the different imputation approaches compared to the true SE, shown in black. SE estimates for the different imputation approaches are similar to the true estimates for 5% and 10% of missing data, but the MI estimates are larger compared to the true SE for 20% and 40% of missing, to account for the uncertainty around the imputed values. The SE estimates for the MI at the item level are slightly larger than those for the imputation at the composite score or subscale level for large proportions of missing data, where available. SEs for the subsequent simulation scenarios show similar patterns and are not shown separately.

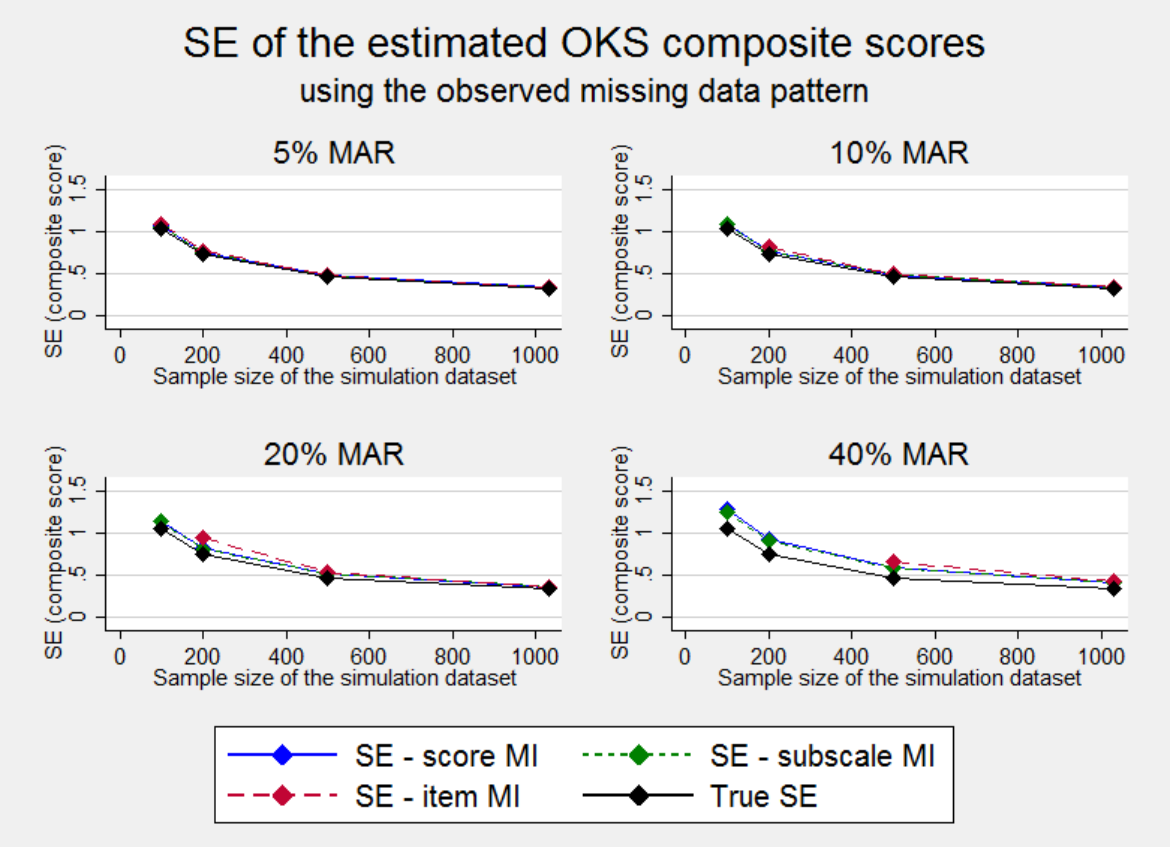


Figure 4-3: Standard errors of the estimated treatment coefficient

Considering the treatment coefficient for the regression on the OKS composite score estimates

Table 4-15 shows the base case simulation results for the estimates of the treatment effect obtained from the different MI approaches of handling missing data in the simulation dataset. Again, differences in terms of the RMSE and MAE are small for this sample size. Here, however, the treatment coefficients are marginally more accurately estimated when using item imputation. Also, in these scenarios, using mean imputation by following the scoring manual offers marginally better results than imputing at the composite score and subscale level. CCA performs similarly to the other approaches.

As for the composite score simulation results, the biases introduced in the estimates for the treatment coefficients (i.e. treatment effects) produced by the various approaches to handling missing data are small. The biggest bias observed is 0.2 (CCA for 40% missing data). Again, this difference is unlikely to be clinically relevant, as it much lower than the between group minimal important difference and the smallest possible detectable change, which are estimated to be five and four points, respectively[12]. The potential impact of these differences on the trial results is further discussed in section 4.6.1.

Table 4-15: Estimated treatment coefficients, RMSE and MAE for the OKS base case simulations (sample size = 1030): results and bias introduced for the treatment coefficient in the linear regression mode

	Mean	SE	MAE	RMSE
True treatment coefficient	0.456	0.619		
5% missing data				
Score imputation - in line with scoring manual	0.452	0.635	0.1233	0.1551
Score imputation - no mean imputation	0.455	0.641	0.1414	0.1793
Subscale imputation - in line with scoring manual	0.447	0.635	0.1214	0.1526
Subscale imputation - no mean imputation	0.457	0.636	0.1236	0.1550
Item imputation	0.466	0.632	**0.1189**	**0.1501**
Complete case analysis	0.468	0.634	0.1207	0.1521
10% missing data				
Score imputation - in line with scoring manual	0.455	0.653	0.1739	0.2208
Score imputation - no mean imputation	0.463	0.665	0.2014	0.2522
Subscale imputation - in line with scoring manual	0.446	0.652	0.1724	0.2167
Subscale imputation - no mean imputation	0.464	0.653	0.1750	0.2208
Item imputation	0.482	0.647	**0.1665**	**0.2117**
Complete case analysis	0.489	0.651	0.1707	0.2180
20% missing data				
Score imputation - in line with scoring manual	0.459	0.692	0.2709	0.3360
Score imputation - no mean imputation	0.466	0.720	0.3128	0.3898
Subscale imputation - in line with scoring manual	0.443	0.691	0.2651	0.3270
Subscale imputation - no mean imputation	0.476	0.693	0.2721	0.3374
Item imputation	0.511	0.684	**0.2610**	**0.3211**
Complete case analysis	0.539	0.690	0.2644	0.3287
40% missing data				
Score imputation - in line with scoring manual	0.458	0.801	0.4691	0.5800
Score imputation - no mean imputation	0.464	0.895	0.5910	0.7365
Subscale imputation - in line with scoring manual	0.415	0.792	0.4576	0.5703
Subscale imputation - no mean imputation	0.477	0.800	0.4904	0.6061
Item imputation	0.625	0.786	**0.4356**	**0.5394**
Complete case analysis	0.653	0.793	0.4476	0.5472

Figure 4-4 shows the RMSE introduced into the treatment coefficients in the regression model on the imputed OKS composite scores, as well as from a CCA. All approaches to handling missing data performed similarly, with a very subtle advantage of CCA in terms of RMSE for 40% of missing data and small samples. The MAE, plots shown in Appendix 8.1 confirm these findings.

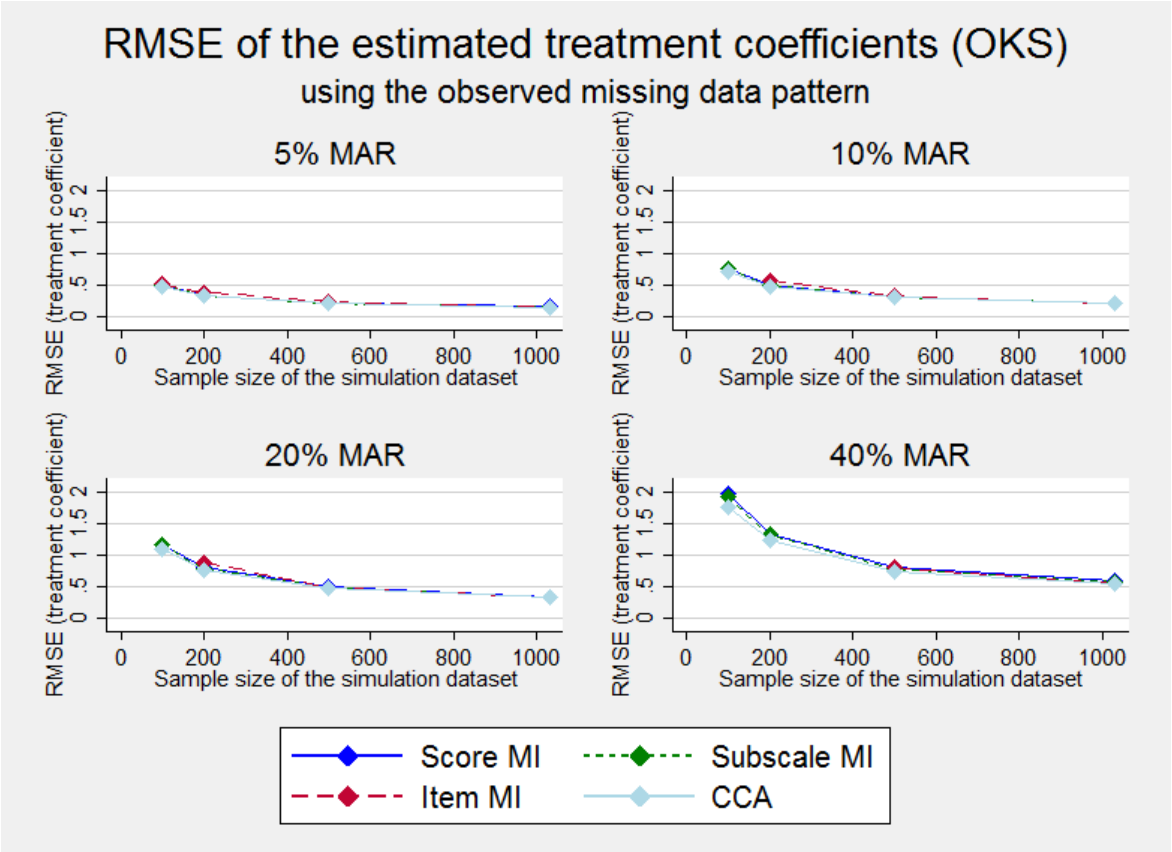


Figure 4-4: RMSE in the treatment coefficient estimates using the imputed OKS composite scores as the outcome variable in the regression model

Figure 4-5 shows the SEs associated with the treatment coefficient estimates for the different approaches to handling missing data, with the true SEs shown in black. Departures from the true SE estimates can be observed for larger proportions of missing data, and the SEs observed for the MI approaches are similar to those obtained from CCA.

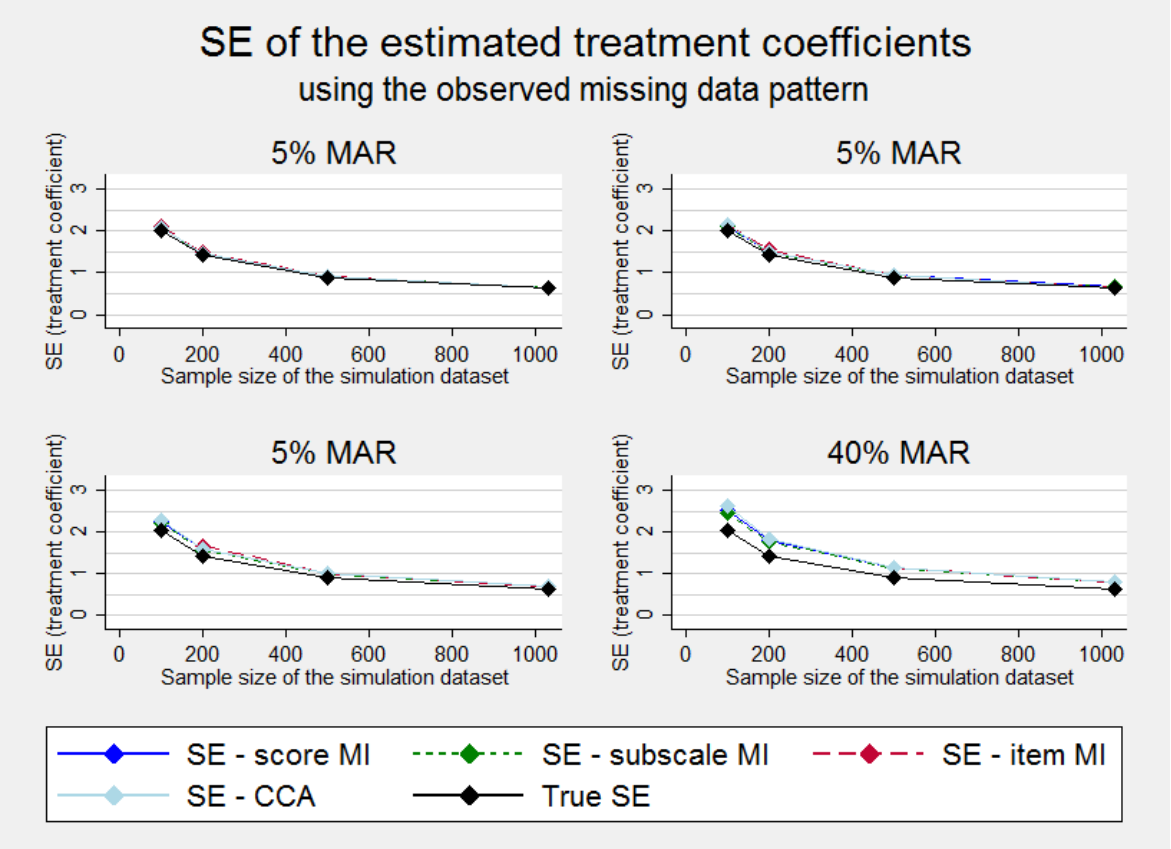


Figure 4-5: SE of the treatment coefficient using the imputed OKS composite scores as the outcome variable in the regression model

4.5.3.2 Results for the OKS simulations – simulating a unit nonresponse MAR mechanism

This section shows results for the different imputation approaches applied to the OKS under a unit-nonresponse MAR mechanism. For this scenario, convergence rates are improved compared to the simulations discussed in section 4.5.3.1; yet results for the item imputation for the simulations of a sample size of 100 and 20%, as well as 40% of missing data are not available due to non-convergence.

Considering the OKS composite score estimates

Figure 4-6 shows the RMSE introduced into the estimate of the OKS composite score under a unit nonresponse MAR mechanism. The different imputation approaches yield very similar results throughout almost all of the different simulation scenarios. However, more bias is introduced into the results when item imputation is used in small sample sizes and large proportions of missing data, i.e. in the scenario of 40% of missing data and a sample size of 200. The MAE plots shown in Appendix 8.1 confirm these findings.

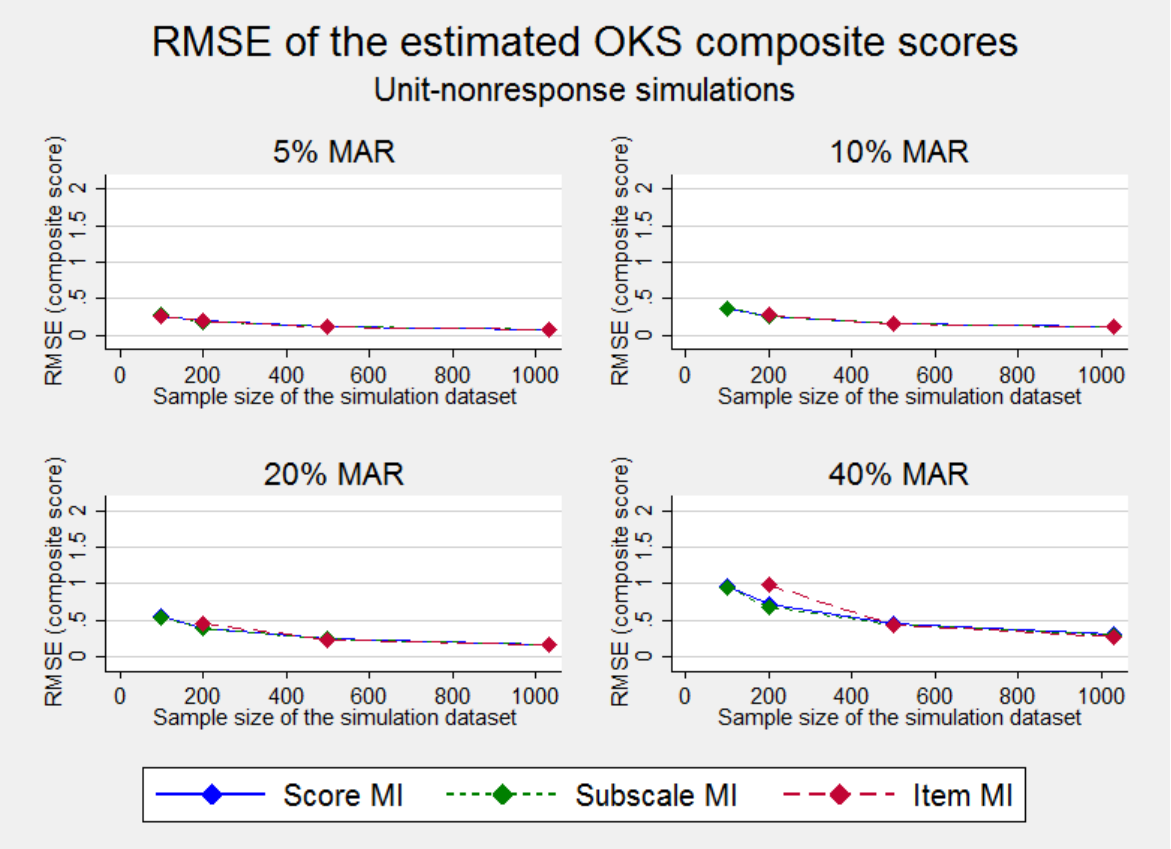


Figure 4-6: RMSE in the OKS composite score estimates (unit-nonresponse simulations)

Considering the treatment coefficient for the regression on the OKS estimates

Figure 4-7 shows the RMSE introduced into the treatment coefficient through the different approaches of handling missing data. All approaches yield very similar results for all sample size scenarios and up to 20% of missing data. Under a unit-nonresponse scenario, there is a small benefit of item imputation and CCA over imputation at the composite score and subscale level for sample sizes of 500 or less with 40% missing data. The MAE plots shown in Appendix 8.1 confirm these findings.

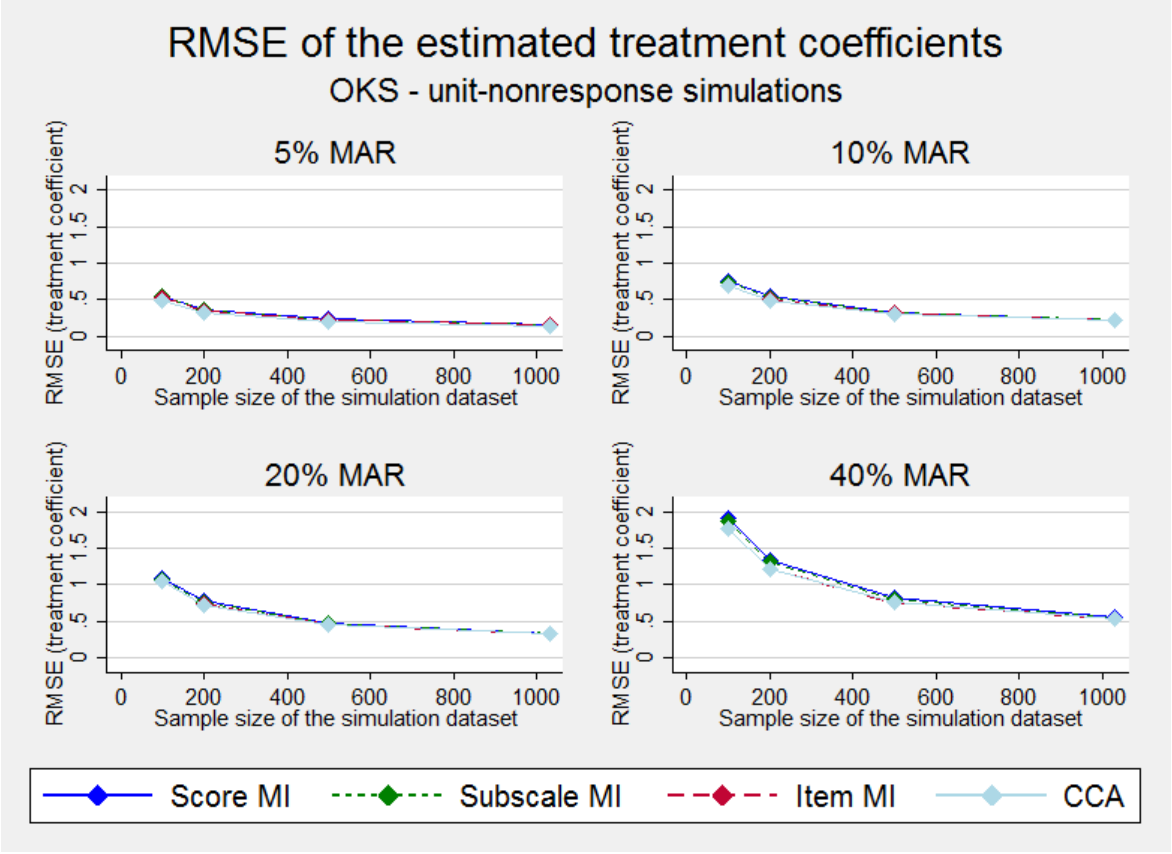


Figure 4-7: RMSE in the treatment coefficient estimates using the imputed OKS as the outcome variable in the regression model (unit-nonresponse simulations)

4.5.3.3 Results for the OKS simulations – simulating 70% item missingness

This section shows results for the different imputation approaches applied to the OKS when 70% of missing data is due to item-nonresponse. For this scenario, results for the item imputation for the simulations combining a sample size of 100 with 20% and 40% of missing data are not available due to non-convergence.

Considering the OKS composite score estimates

Figure 4-8 shows the RMSE introduced into the estimate of the OKS composite score in the presence of 70% item-nonresponse. The different approaches perform similarly for large sample sizes and small amounts of missing data. However, for larger proportions of missing data, there appears to be an advantage of imputation at the subscale level over imputation at the composite score level, possibly because imputation at the subscale can utilise information that is discarded in MI at the composite score level. Item imputation appears to perform worse than the two alternative methods for combinations of large proportions of missing data and small sample sizes. The MAE plots shown in Appendix 8.1 confirm these findings.

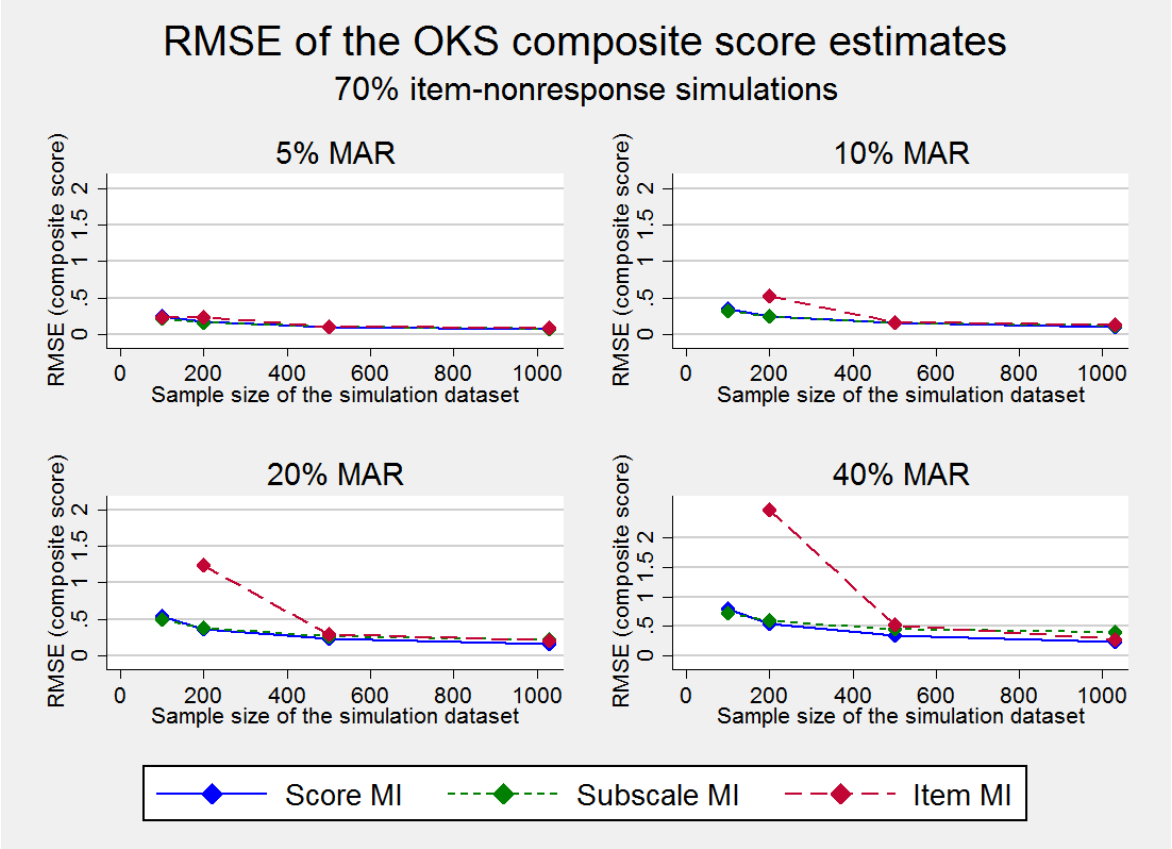


Figure 4-8: RMSE in the OKS composite score estimates (70% item non-response simulations)

Considering the treatment coefficient for the regression on the OKS estimates

Figure 4-9 shows the RMSE introduced into the treatment coefficient through the different approaches of handling missing data. Here, both imputation at the item and subscale level yield less bias in terms of RMSE compared to MI at the composite score level and CCA. The MAE plots shown in Appendix 8.1 confirm these findings.

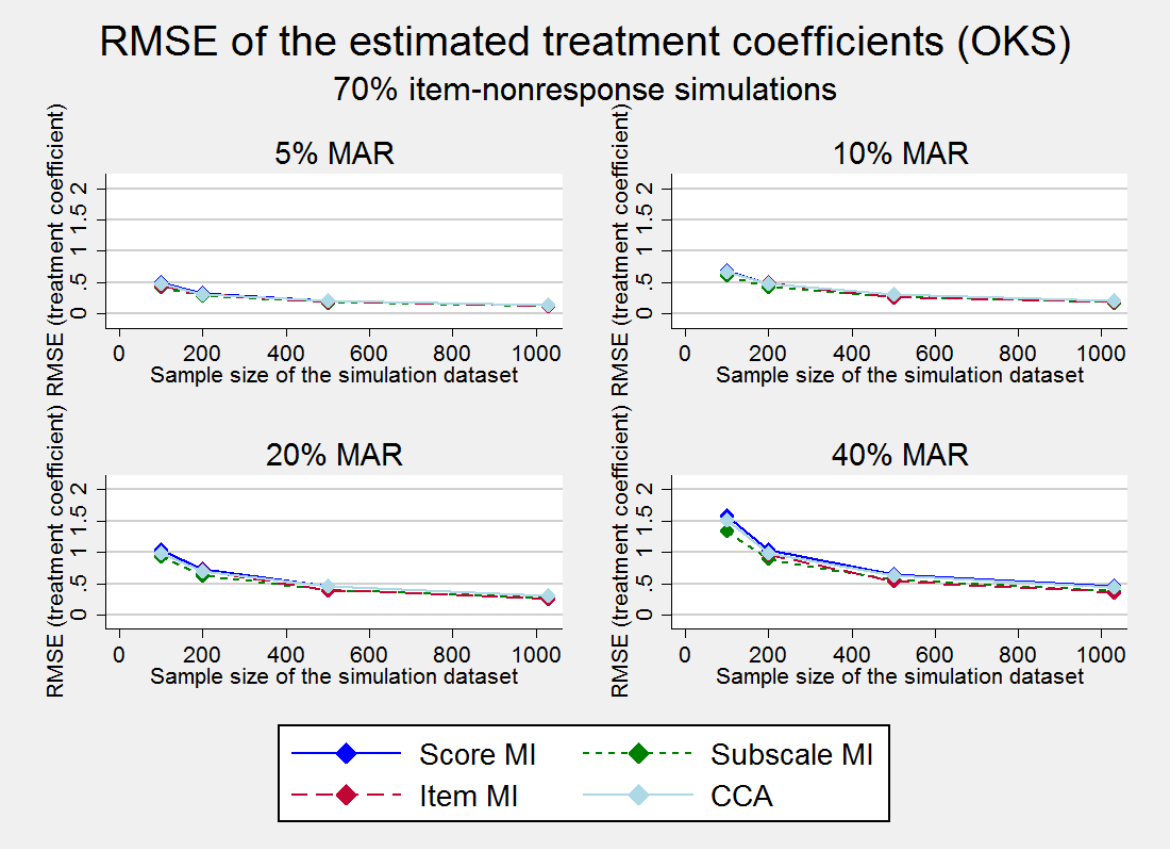


Figure 4-9: RMSE in the treatment coefficient estimates using the imputed OKS as the outcome variable in the regression model (70% item missingness simulations)

4.5.3.4 Results for the OKS simulations – introducing a five point treatment effect

This section shows results for the different imputation approaches applied to the OKS after simulating a five point treatment effect in the base case dataset. For this scenario, insufficient results for the item imputation are available for all simulations combining 10% of missing data or more and a sample size of 100, as well as the scenario considering 40% of missing data in a sample of 200.

Considering the OKS composite score estimates

Figure 4-10 shows the RMSE introduced into the estimate of the OKS composite score when the treatment difference between the groups has been increased to five points. Here, imputation at the item level performs notably worse than the other approaches in terms of RMSE for sample sizes of 500 of less. The MAE plots shown in Appendix 8.1 confirm these findings.

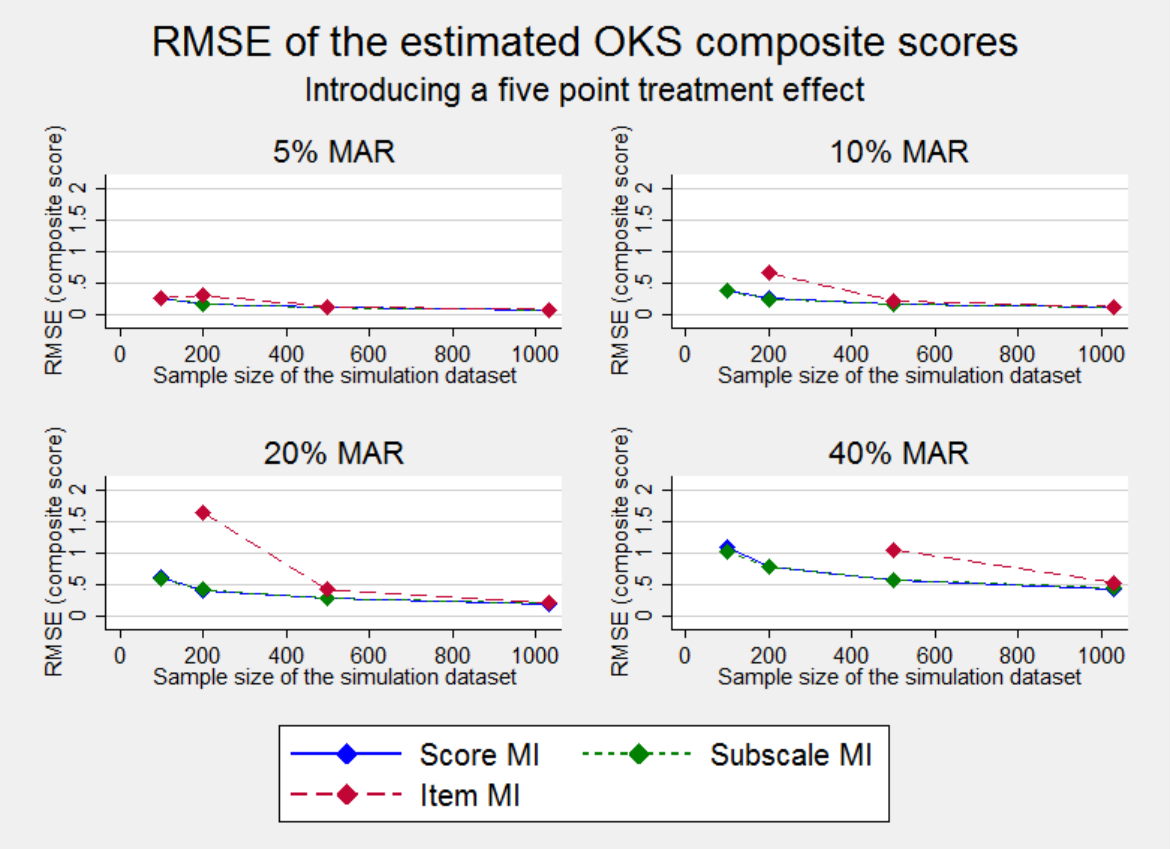


Figure 4-10: RMSE in the OKS composite score estimates (introducing a five point treatment effect)

Considering the treatment coefficient for the regression on the OKS estimates

Figure 4-11 shows the RMSE introduced into the treatment coefficient through the different approaches of handling missing data. For 10% of missing data or more, there is a disadvantage when imputing at the item level compared to imputation at the composite score or subscale level and CCA. However, these differences in bias decrease with decreasing sample size. The MAE plots shown in Appendix 8.1 confirm these findings.

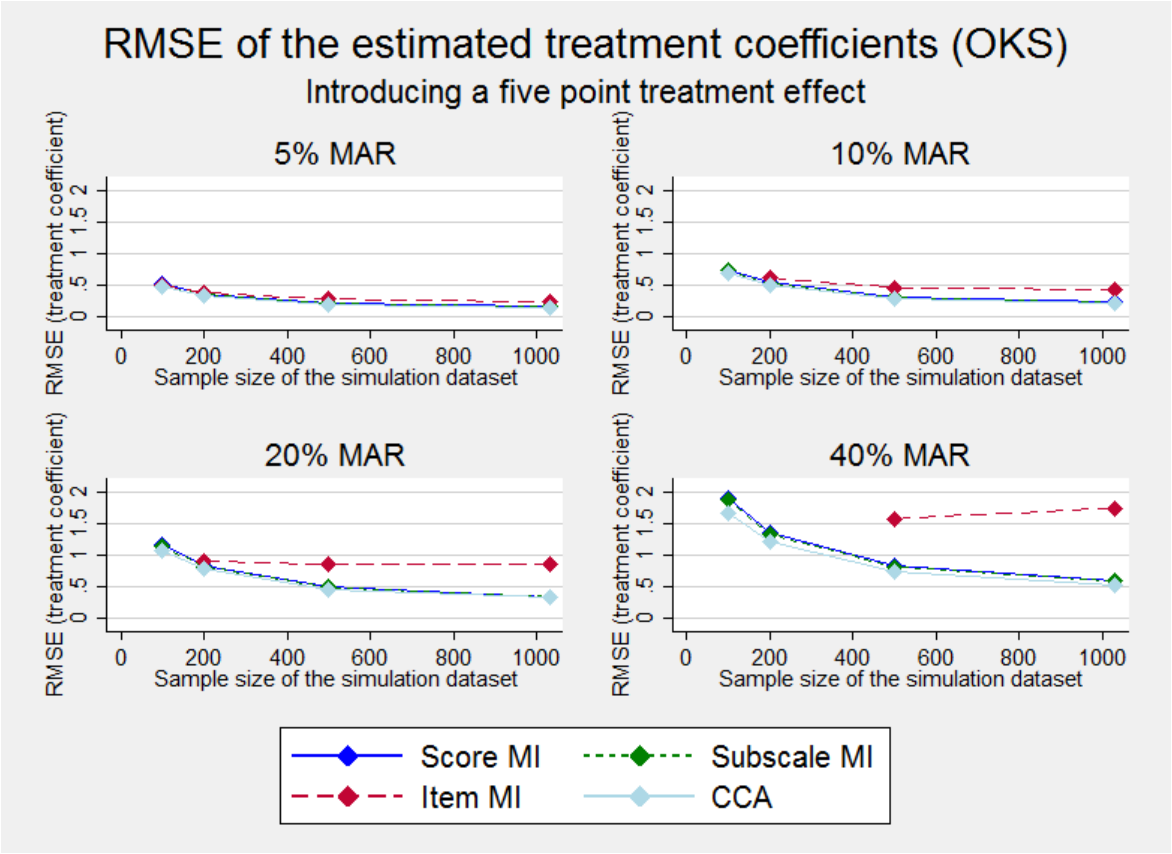


Figure 4-11: RMSE in the treatment coefficient estimates using the imputed OKS as the outcome variable in the regression model (introducing a five point treatment effect)

4.5.3.5 Results for the OKS simulations – Comparing scoring in line with the scoring manual versus no mean imputations

This section compares the bias introduced into the results when either following the OKS scoring manual, whereby up to two items can be replaced by the mean of the available items vs. not using mean imputation (i.e. classifying composite scores as missing when one or two missing items are unavailable).

Considering the OKS composite score estimates

In terms of RMSE, there is little difference between following the scoring manual versus not applying any mean imputation when multiply imputing at the composite score level, perhaps with the exception of 20% and 40% of missing data in a sample size of 100, where following the scoring manual provides marginally better results (Figure 4-12). Considering imputations at the subscale level, avoiding mean imputations offers a benefit for combinations of 40% of missing data and sample sizes of 200 and above (Figure 4-13).

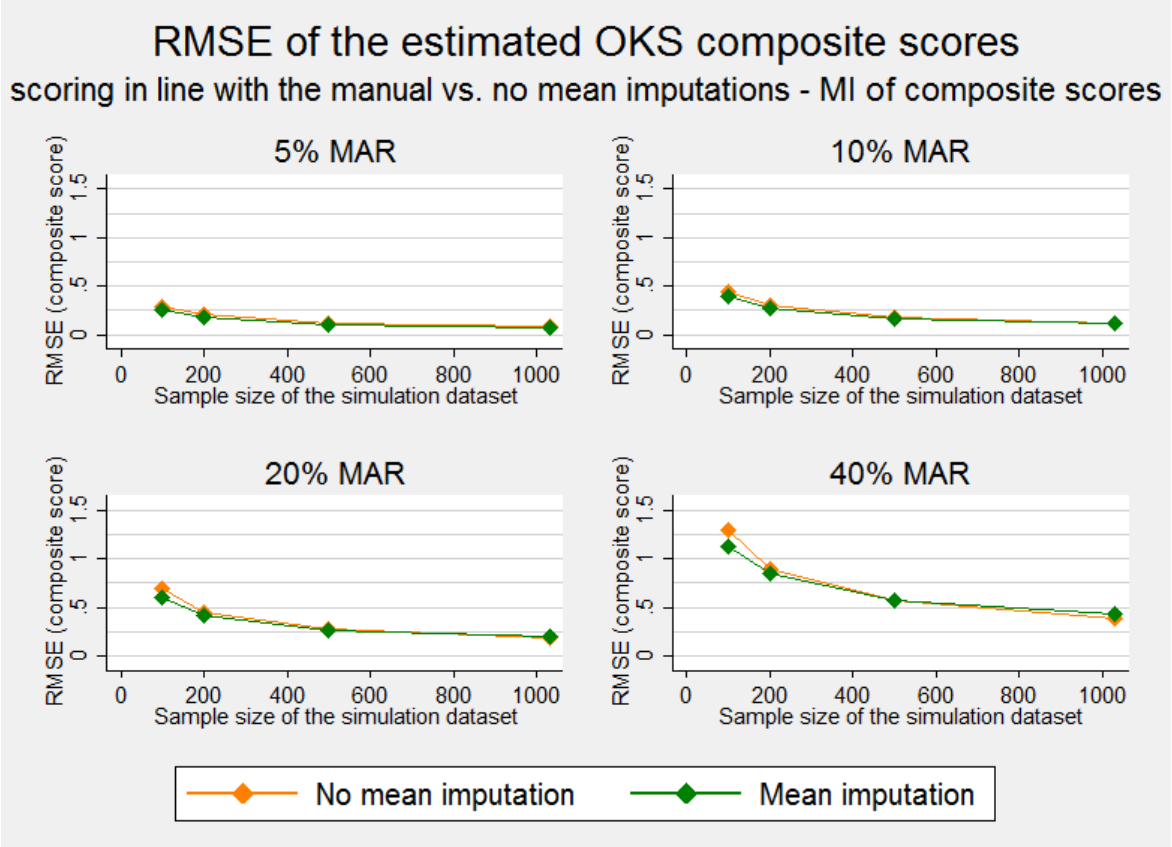


Figure 4-12: RMSE in the OKS composite score estimates comparing the effect of following the scoring manual vs. not using mean imputation for MI at the composite score level (using the observed missing data patterns)

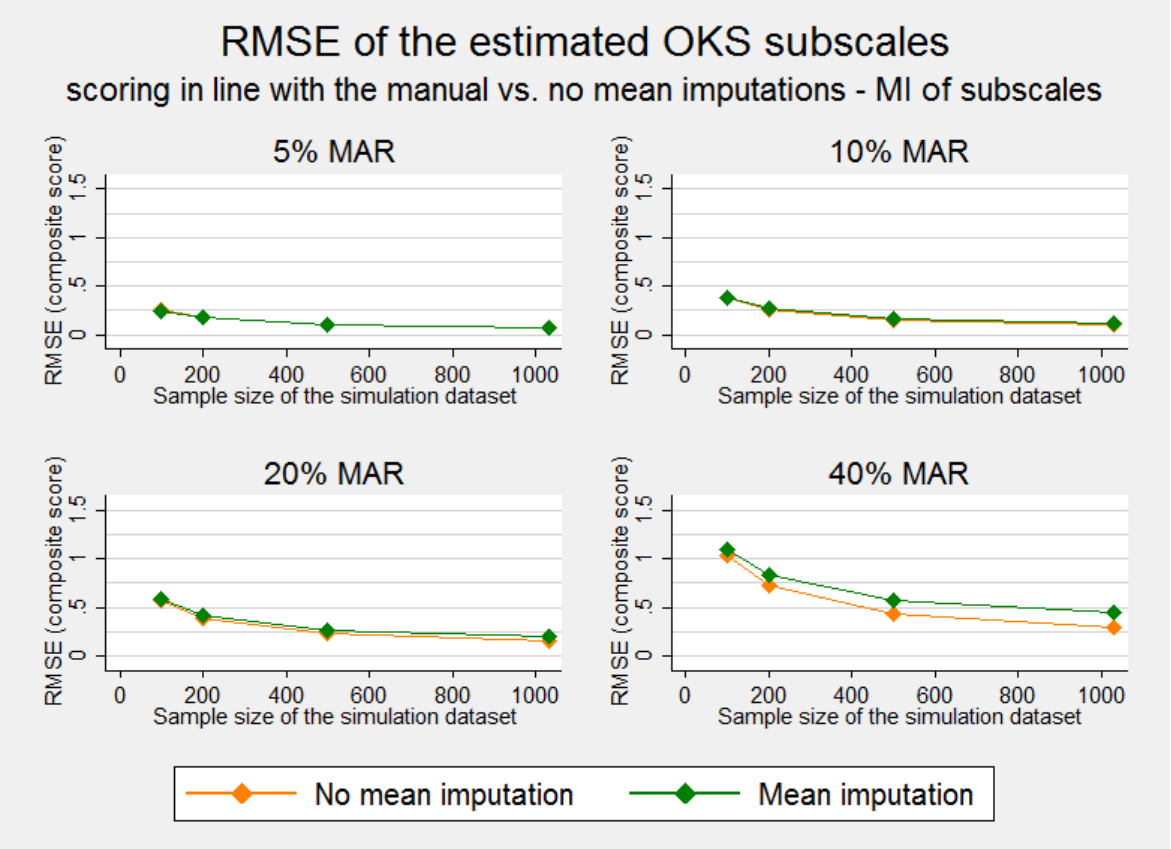


Figure 4-13: RMSE in the OKS composite score estimates comparing the effect of following the scoring manual vs. not using mean imputation for MI at the subscale level (using the observed missing data patterns)

Considering the treatment coefficient for the regression on the OKS estimates

The above observed effect is reversed, with the treatment coefficients for 40%, as well as for combinations of 20% of missing data and small sample sizes being more accurately estimated when the scoring manual is followed when MI was performed at the composite score level (Figure 4-14), but little difference being observed when MI is applied to the subscale level (Figure 4-15).

The graphs showing the MAE are presented in Appendix 8.1, and confirm these findings.

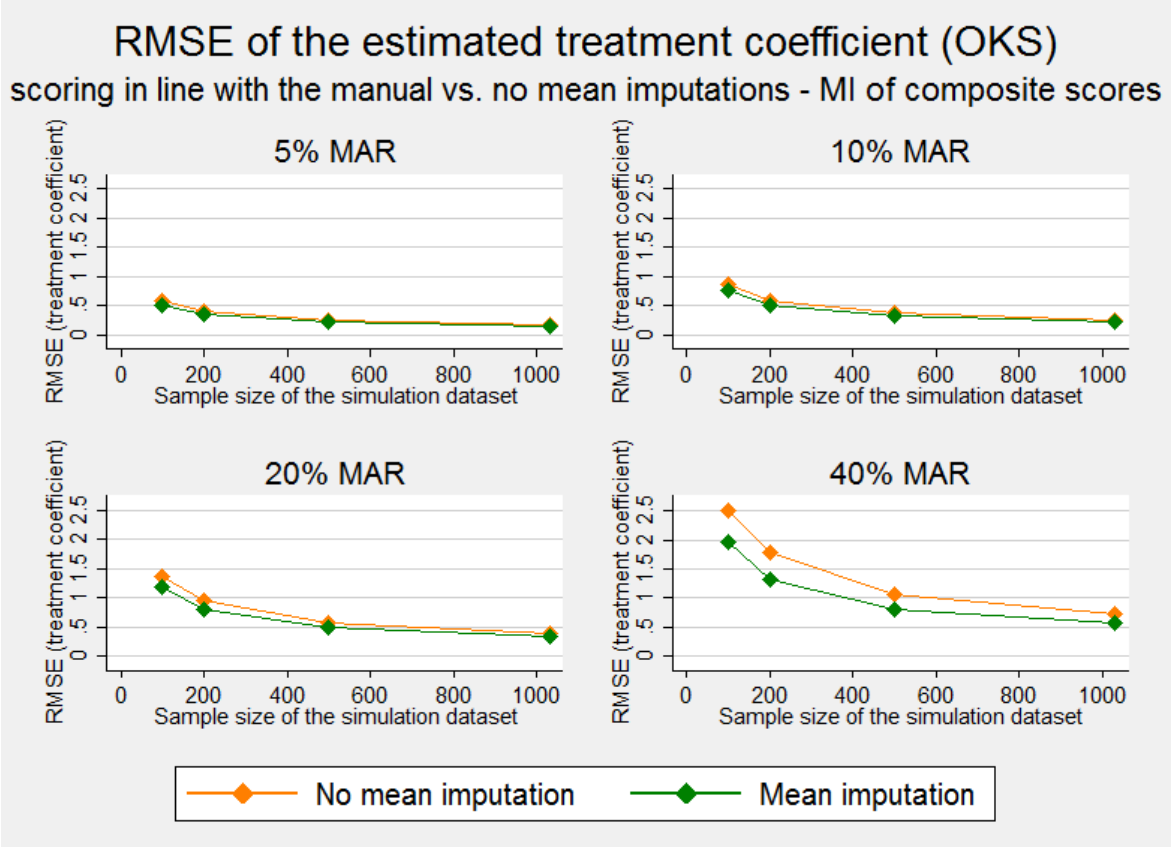


Figure 4-14: RMSE in the treatment coefficient estimates comparing the effect of following the scoring manual vs. not using mean imputation for MI at the composite score level (using the observed missing data patterns)

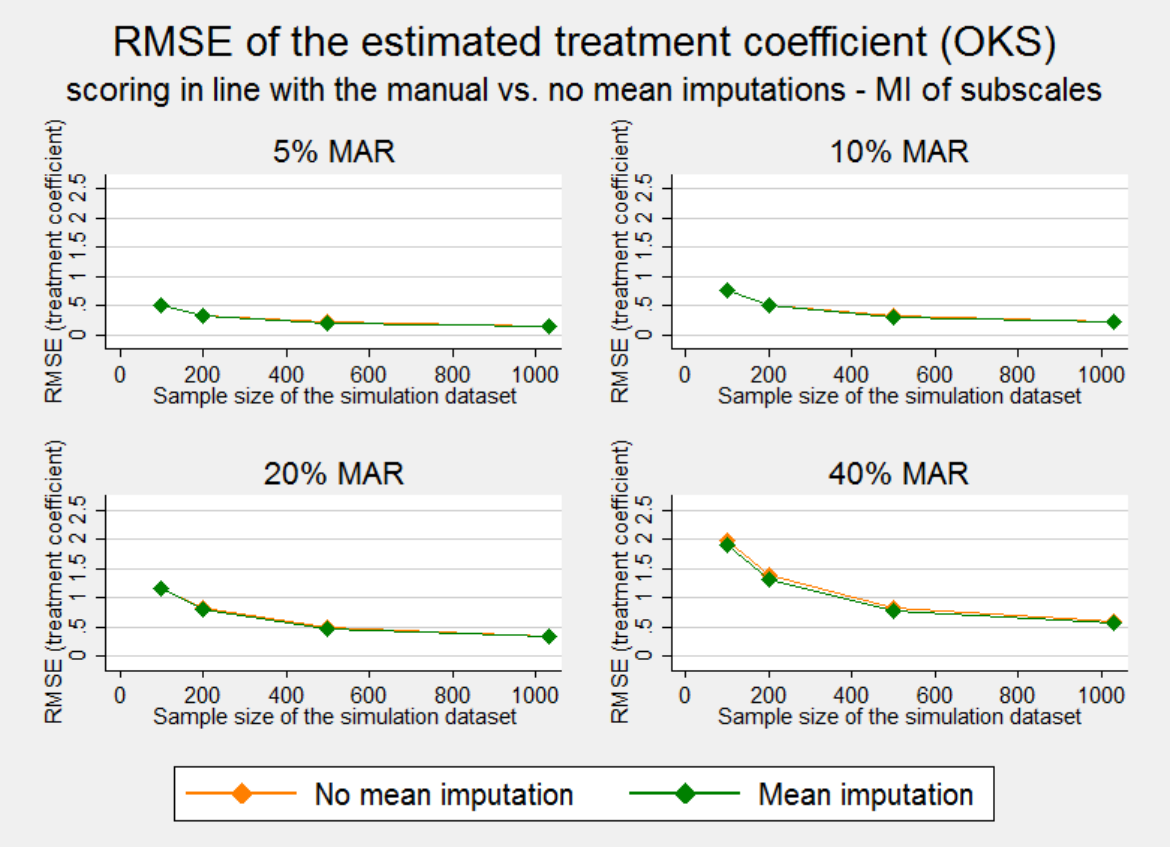


Figure 4-15: RMSE in the treatment coefficient estimates comparing the effect of following the scoring manual vs. not using mean imputation for MI at the subscale level (using the observed missing data patterns)

4.5.3.6 Results for the EQ-5D-3L simulations

This section examines the comparative performance of MI at the composite score vs. item level when applied to the EQ-5D-3L. The main portion of this section focuses on the results obtained from the simplified EQ-5D-3L item imputation model.

Considering the EQ-5D-3L composite score estimates

Table 4-16 shows the mean EQ-5D-3L composite scores, their SEs, together with the MAR and RMSE for the base case simulations. For 10% of missing data, the composite scores are slightly overestimated for imputations at the composite score and underestimated for imputations at the item level, with differences becoming more pronounced as the proportion of missing data increases. Comparable standard errors are produced by both imputation approaches.

Here, imputation at the composite score level consistently performs better than imputation at the item level in terms of MAE and RMSE. The differences in bias are very small overall, but become more pronounced with increasing amounts of missing data within the dataset.

Table 4-16: Estimated EQ-5D-3L composite score, RMSE and MAE for the EQ-5D-3L base case simulations (sample size = 1160): results and bias introduced for the composite score estimates

EQ-5D-3L base cases	Mean	SE	MAE	RMSE
True EQ-5D-3L	0.709	0.008		
5% missing data				
Score imputation	0.709	0.008	**0.0014**	**0.0017**
Item imputation	0.708	0.008	0.0015	0.0019
10% missing data				
Score imputation	0.710	0.008	**0.0020**	**0.0025**
Item imputation	0.707	0.008	0.0026	0.0032
20% missing data				
Score imputation	0.710	0.008	**0.0031**	**0.0039**
Item imputation	0.705	0.009	0.0050	0.0059
40% missing data				
Score imputation	0.711	0.010	**0.0054**	**0.0068**
Item imputation	0.698	0.011	0.0111	0.0126

Following on from the base cases presented in Table 4-16, Figure 4-16 shows the RMSE introduced into the estimate of the EQ-5D-3L composite score following the two imputation approaches. Very little difference can be observed for the performance of either imputation approach for 5% of missing data for any sample size. It can be observed that the lines indicating the RMSE drift further apart both with decreasing sample size and increasing proportions of missing data, showing that the comparative benefits of imputation at the composite score level increase in these scenarios. The MAE, plots shown in Appendix 8.2 confirm these findings.

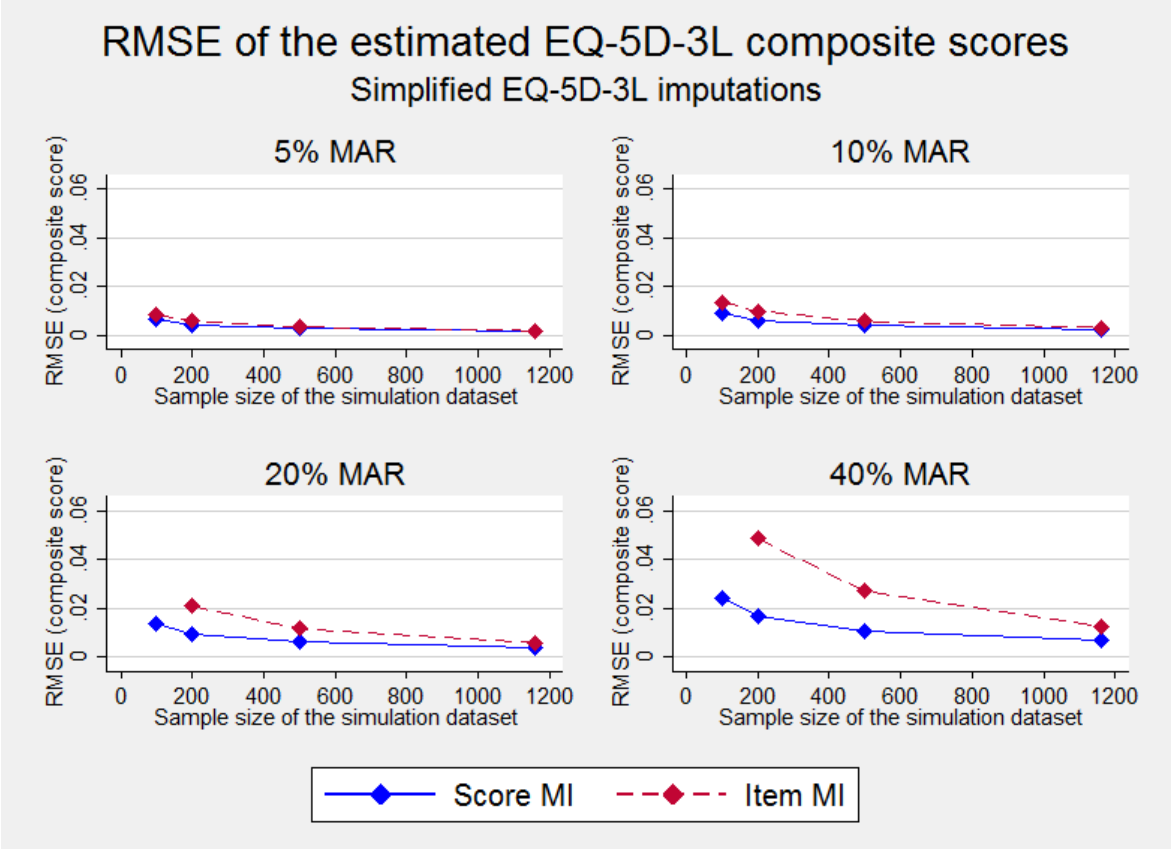


Figure 4-16: RMSE in the EQ-5D-3L composite score estimates

Figure 4-17 shows the SEs associated with the EQ-5D-3L estimates for the different simulation scenarios, with the true SEs shown in black. SEs for all sample sizes appear similar for low proportions of missing data. However, the SEs for the imputation approaches is larger than the true SE for larger proportions of missing data, to account for the uncertainty around the imputed data. Imputation at the item level produces larger SEs than imputation at the composite score level.

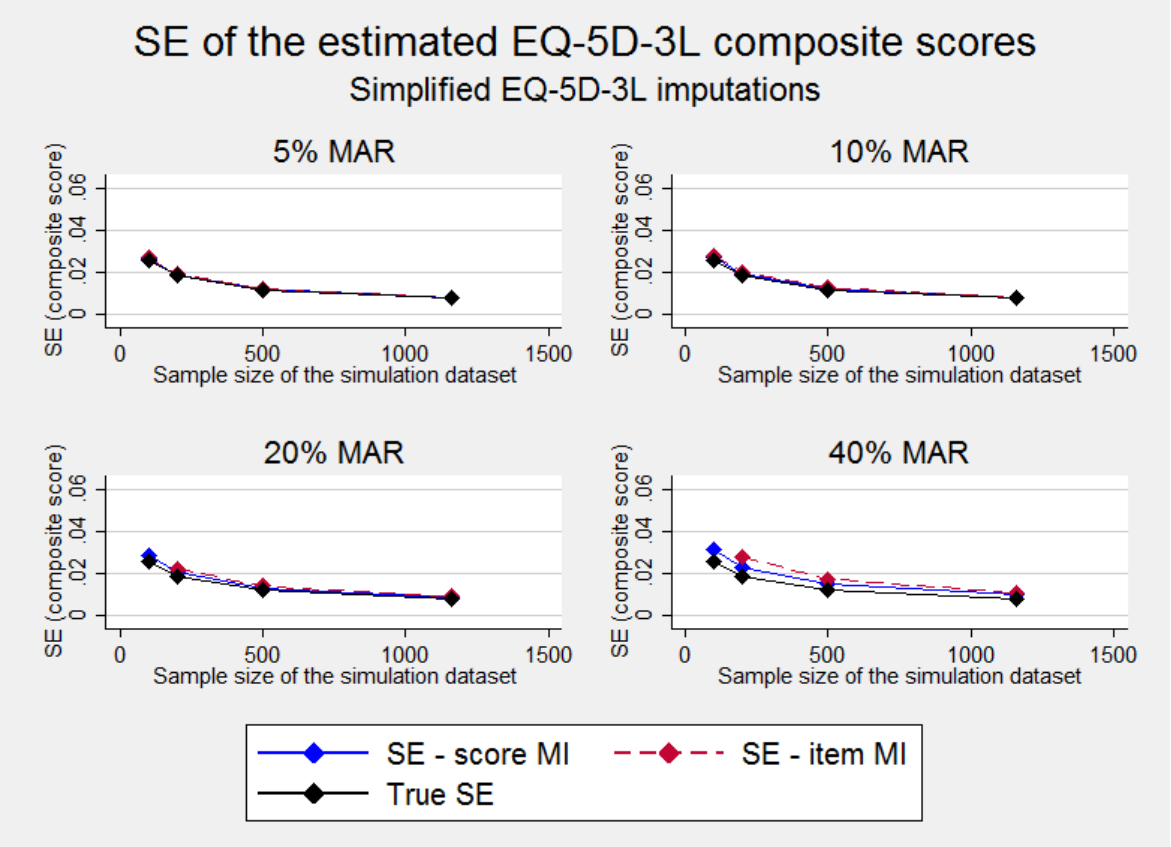


Figure 4-17: SE of the EQ-5D-3L composite score estimates

Considering the treatment coefficient for the regression on the EQ-5D-3L estimates

Table 4-17 shows the mean regression coefficients, their SEs as well as the MAR and RMSE associated with the estimates for the base case simulations. The mean coefficient estimates for the imputation at the composite score level averaged over the 1,000 simulations is closest to the actual EQ-5D-3L, however, individual results from each imputation at the composite score level are slightly more biased that those based on imputation at the item level or CCA, as indicated by the RMSEs and MAEs. Bias introduced into the treatment coefficient in terms of the MAR and RMSE indicate that for up to 10% of missing data, the different imputation approaches perform similarly, but CCA provides a small benefit over the imputation approaches for larger proportions of missing data in this scenario.

Table 4-17: Estimated treatment coefficients, MAE and RMSE for the EQ-5D-3L base case simulations (sample size = 1160): results and bias introduced for the treatment coefficient in the linear regression model

EQ-5D-3L base cases	Mean	SE	MAE	RMSE
True EQ-5D-3L treatment coefficient	0.014	0.015		
5% missing data				
Score imputation	0.014	0.015	0.0028	0.0035
Item imputation	0.014	0.015	**0.0027**	**0.0034**
Complete case analysis	0.014	0.015	**0.0027**	**0.0034**
10% missing data				
Score imputation	0.014	0.016	0.0041	0.0051
Item imputation	0.013	0.016	**0.0040**	**0.0049**
Complete cases	0.015	0.016	0.0040	0.0050
20% missing data				
Score imputation	0.014	0.017	0.0060	0.0075
Item imputation	0.012	0.017	0.0062	0.0077
Complete case analysis	0.015	0.016	**0.0057**	**0.0071**
40% missing data				
Score imputation	0.014	0.019	0.0100	0.0125
Item imputation	0.011	0.021	0.0104	0.0131
Complete case analysis	0.017	0.019	**0.0093**	**0.0115**

Figure 4-18 shows the RMSE introduced into the treatment coefficient in the regression model on the EQ-5D-3L scores obtained via the different imputation approaches, as well as from a CCA. This figure shows that all three approaches provide similar results for the scenarios with 5% and 10% of missing data. For 20% of missing data, an increased level of bias in the RMSE can be observed for a sample size of 200 for the item imputation, and MI at the item level performs worse than the other two approaches for a sample size of 200 and 500 when 40% of MAR data are simulated. A number of insufficient valid results were obtained for the item level imputation simulations for a sample size of 100 for 20% or more missing data. The MAE, plots shown in Appendix 8.2 confirm these findings.

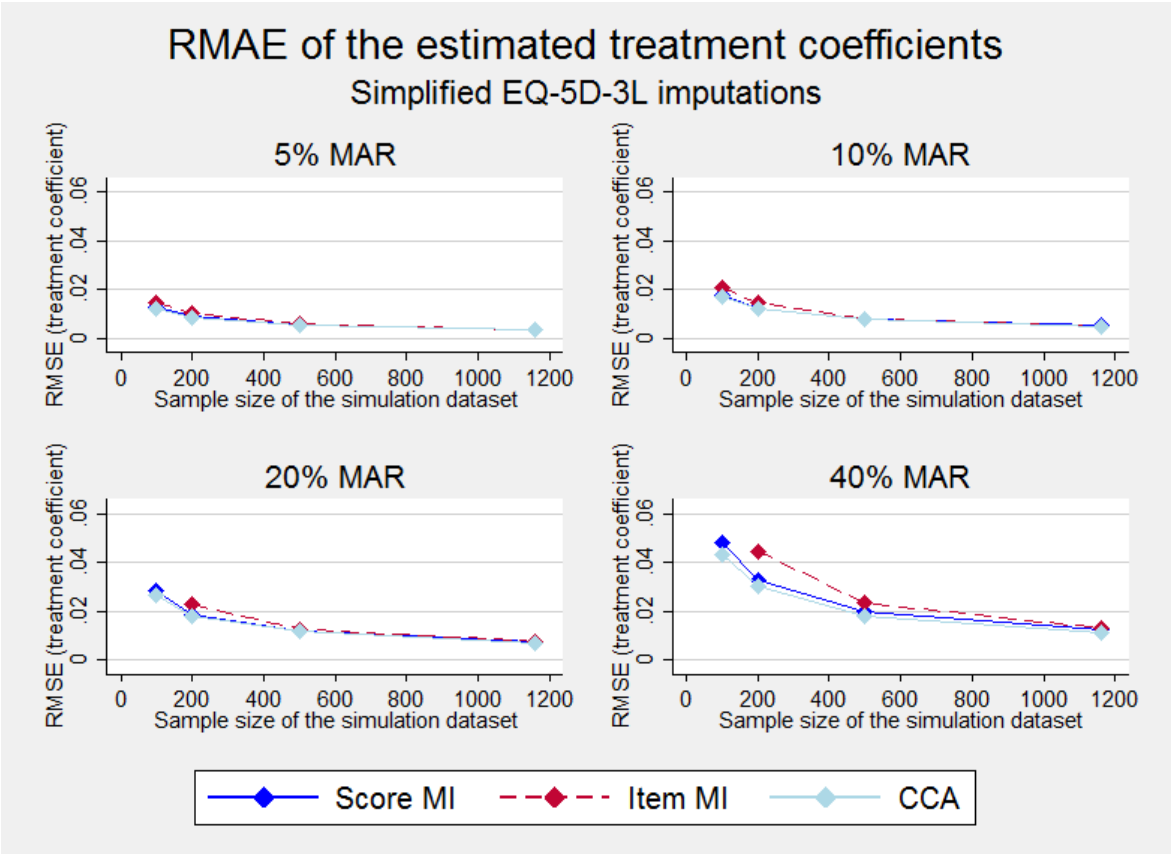


Figure 4-18: RMSE in the treatment coefficient estimates using the imputed EQ-5D-3L as the outcome variable in the regression model

Figure 4-5 shows the SEs associated with the treatment coefficient estimates for the difference simulation scenarios, with the true SEs shown in black. Larger SEs are produced

for larger proportions of missing data in both imputation approaches, as well as the CCA. Imputation at the composite score level and CCA produce similar SEs, while the SEs for the imputation at the item level are slightly larger.

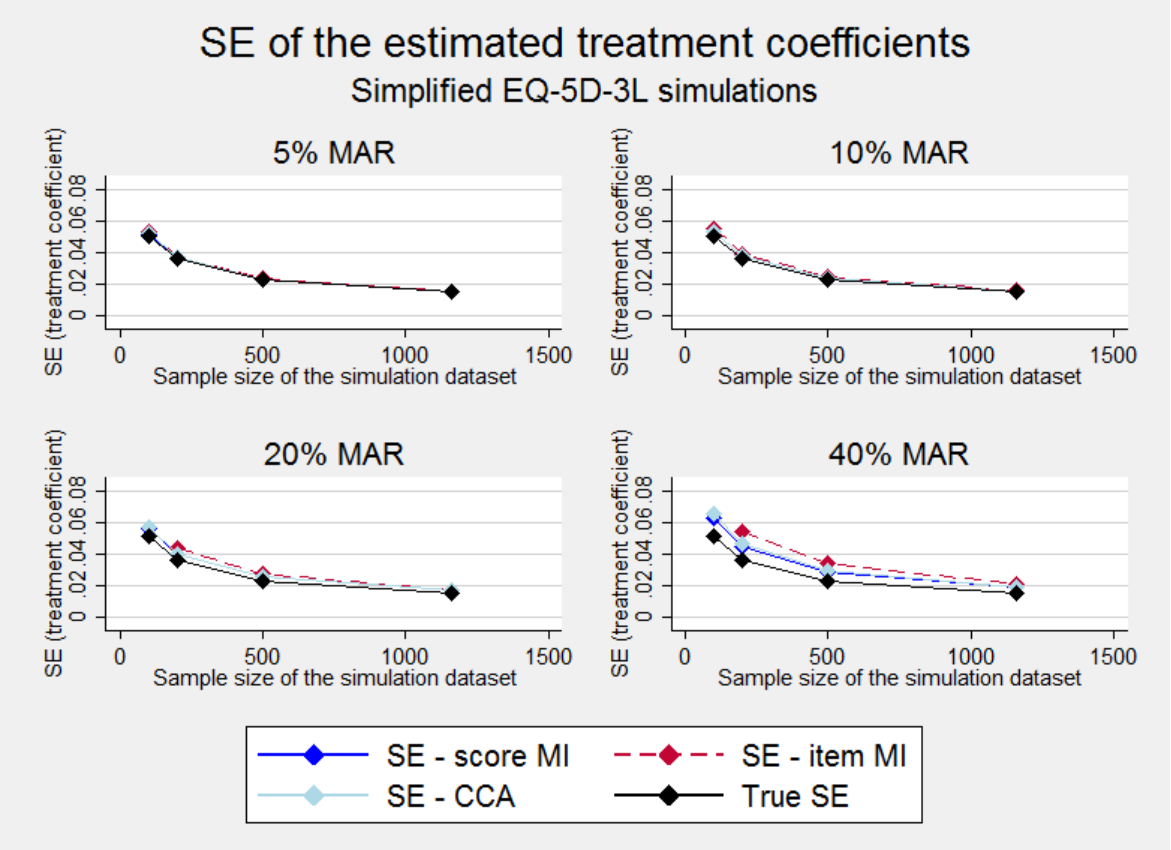


Figure 4-19: SE of the treatment coefficient using the imputed EQ-5D-3L as the outcome variable in the regression model

4.5.3.7 Results for the EQ-5D-3L simulations – comparison of the complex and simplified item-level imputations

The item imputation simulation for the EQ-5D-3L were run using a complex model adjusting also for baseline items, as well as due to issues with non-convergence, their simplified versions not adjusting for baseline items. Both imputation models were adjusted for baseline EQ-5D-3L composite score as well as relevant baseline characteristics, and run separately by treatment arm. Here, the bias introduced into both models is compared where 1,000 valid results were obtained.

Figure 4-20 shows the different RMSE in the EQ-5D-3L composite scores when using the complex and simplified imputation model. It can be observed that the more complex model introduces a higher level of bias.

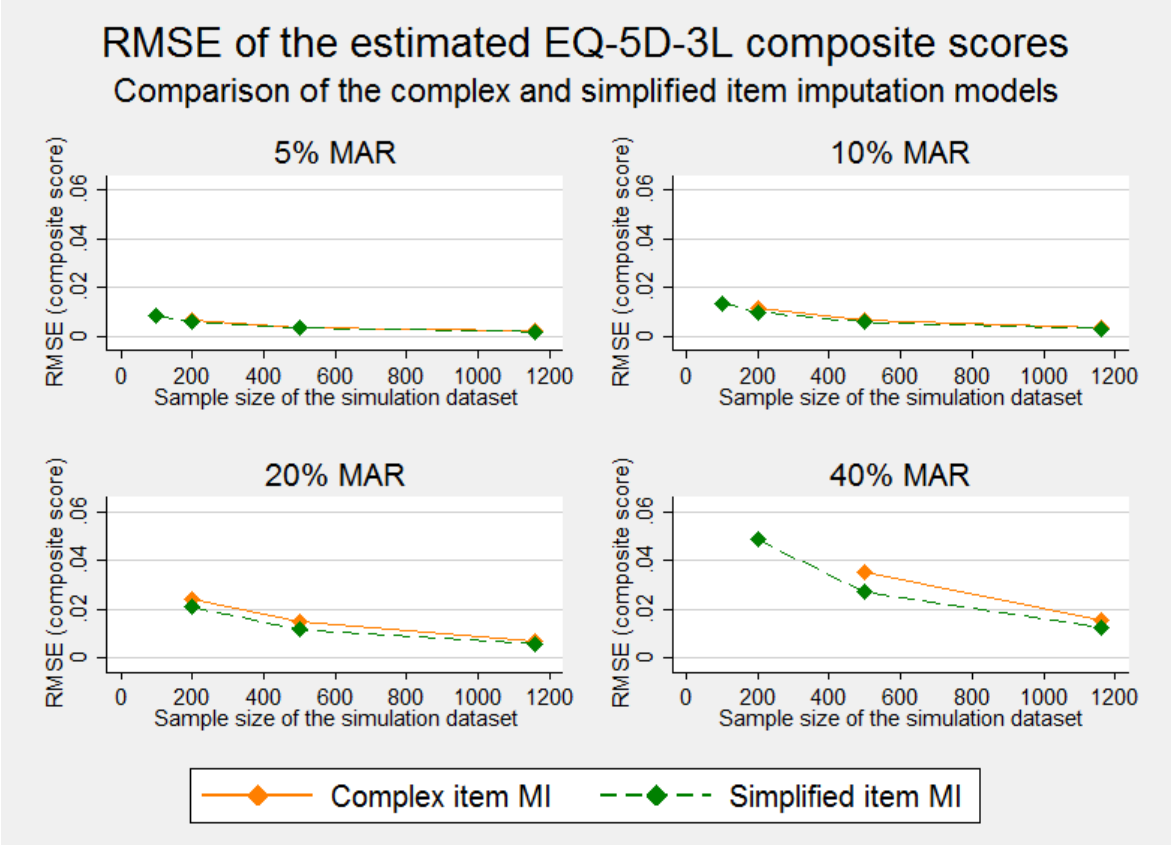


Figure 4-20: RMSE in the EQ-5D-3L composite score estimates – comparing the complex and simplified item imputation model

Figure 4-21 shows the RMSE in the treatment coefficient estimates. Here, very little difference can be observed between the two approaches; the more complex model does not offer an advantage over the simplified item MI model, but yields more convergence problems. The graphs for the MAE, shown in Appendix 8.2 support these findings.

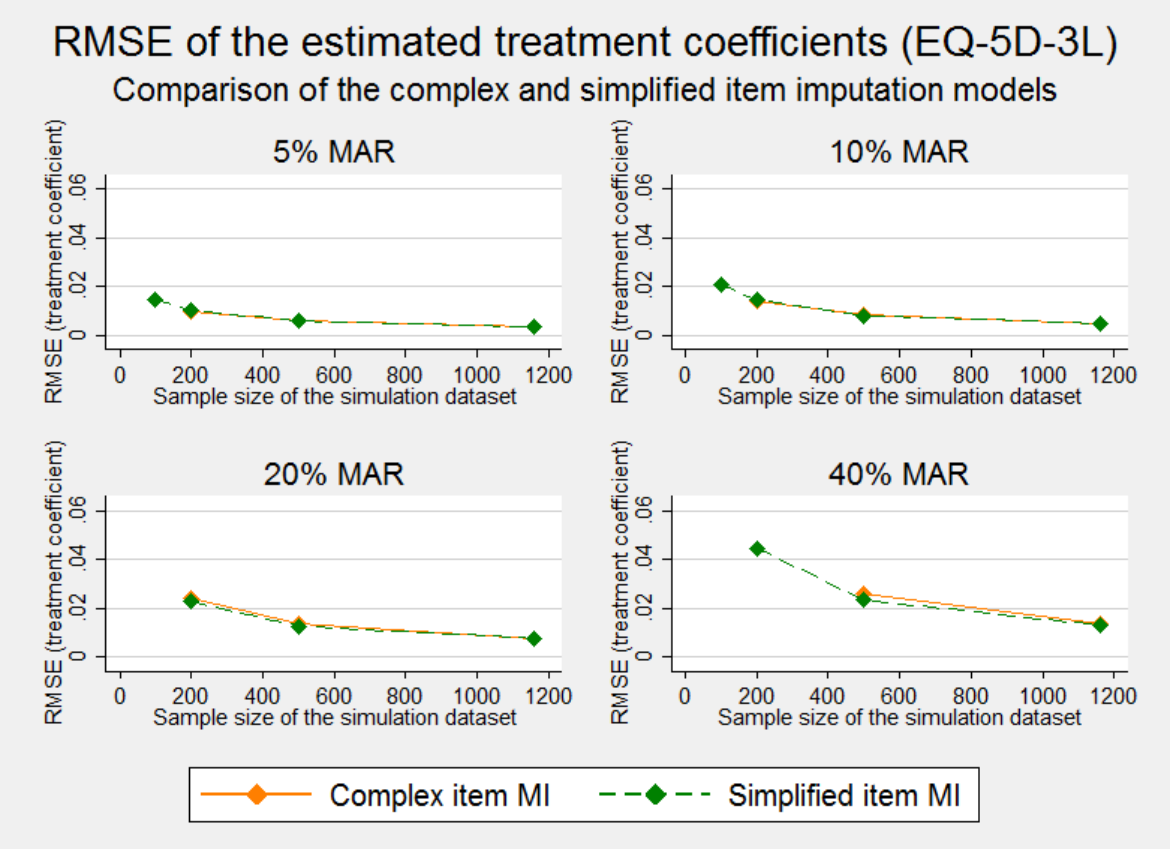


Figure 4-21: RMSE in the treatment coefficient estimates using the imputed EQ-5D-3L as the outcome variable in the regression model – comparing the complex and simplified item imputation model

4.5.3.8 Results for the SF-12 simulations

This section examines the comparative performance of MI at the subscales vs. item level in terms of bias introduced into composite scores and treatment coefficients when applied to the SF-12. The main portion of this section focuses on the results obtained from the simplified SF-12 item imputation model, i.e. the imputations not performed separately by treatment arm.

Considering the SF-12 composite score estimates

Table 4-18 and Table 4-19 show the mean SF-12 composite scores, their SEs, together with the MAR and RMSE for the base case simulations. For both the MCS and PCS scores, MI at the composite score level offers a small benefit in terms of RMSE and MAE when at least 10% of data are MAR.

Table 4-18: Estimated SF-12 MCS score, RMSE and MAE for the SF-12 base case simulations (sample size = 797): results and bias introduced for the PROMs composite score estimates

SF-12 base cases: MCS score	Mean	SE	MAE	RMSE
True MCS score	50.801	0.366		
5% missing data				
Score imputation	50.800	0.374	0.0617	0.0779
Item imputation	50.834	0.372	**0.0616**	**0.0776**
10% missing data				
Score imputation	50.806	0.382	**0.0883**	**0.1114**
Item imputation	50.877	0.379	0.0990	0.1234
20% missing data				
Score imputation	50.815	0.400	**0.1343**	**0.1691**
Item imputation	50.944	0.394	0.1730	0.2106
40% missing data				
Score imputation	50.853	0.444	**0.2332**	**0.2900**
Item imputation	51.091	0.432	0.3373	0.4002

Table 4-19: Estimated SF-12 PCS score, RMSE and MAE for the SF-12 base case simulations (sample size = 797): results and bias introduced for the PROMs composite score estimates

SF-12 base cases: PCS score	Mean	SE	MAE	RMSE
True PCS score	39.409	0.386		
5% missing data				
Score imputation	39.414	0.395	0.0695	0.0872
Item imputation	39.448	0.393	**0.0682**	**0.0852**
10% missing data				
Score imputation	39.427	0.404	**0.1009**	**0.1239**
Item imputation	39.491	0.400	0.1073	0.1322
20% missing data				
Score imputation	39.448	0.424	**0.1529**	**0.1894**
Item imputation	39.579	0.416	0.1956	0.2354
40% missing data				
Score imputation	39.522	0.474	**0.2549**	**0.3222**
Item imputation	39.811	0.457	0.4190	0.4781

Figure 4-22 and Figure 4-23 show the RMSE introduced in the SF-12 MCS and PCS scores for various sample size and missing data scenarios. For the SF-12 MCS score simulations, item imputation tends to perform worse for larger proportions of missing data and smaller sample sizes. For the SF-12 PCS score, very little differences can be detected between the imputation approaches, with the exception of the 40% missing data scenario, where the item MI approach yields more bias.

These results are confirmed by plots of the MAE, shown in Appendix 8.3.

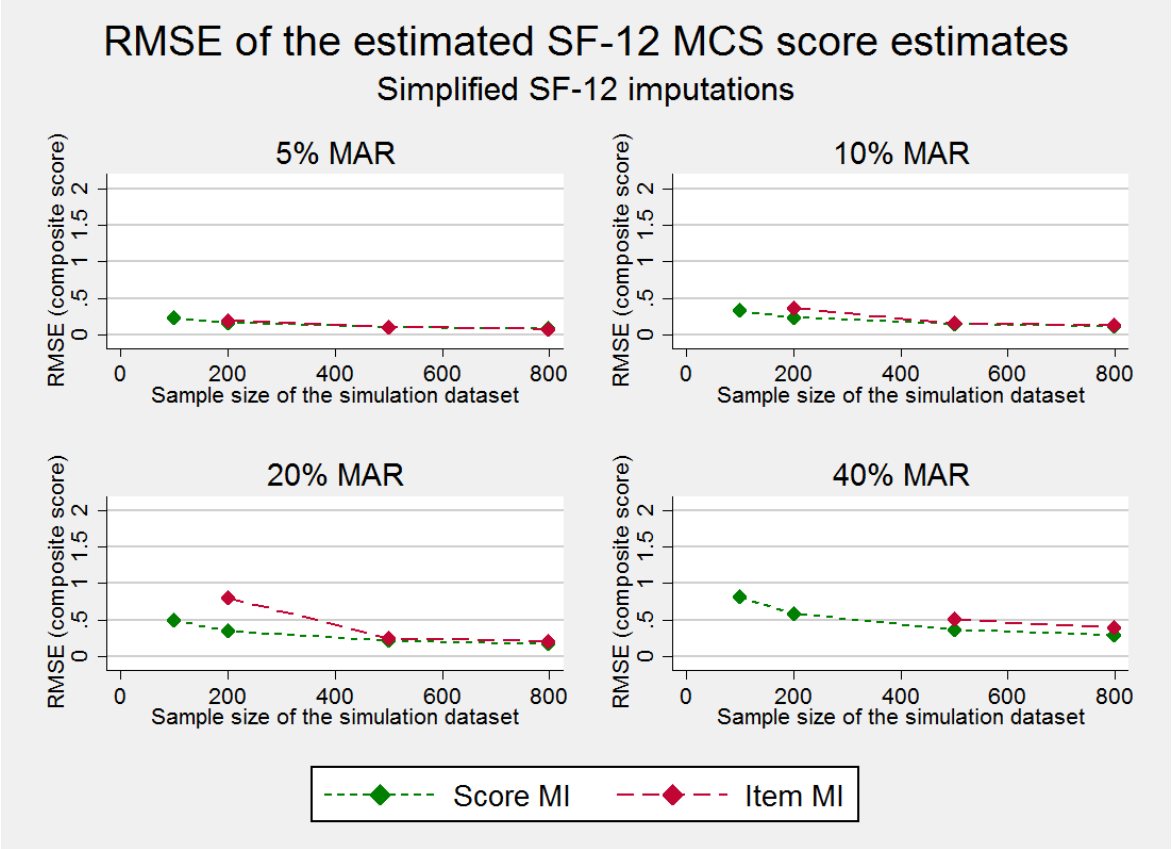


Figure 4-22: RMSE in the SF-12 MCS score estimates

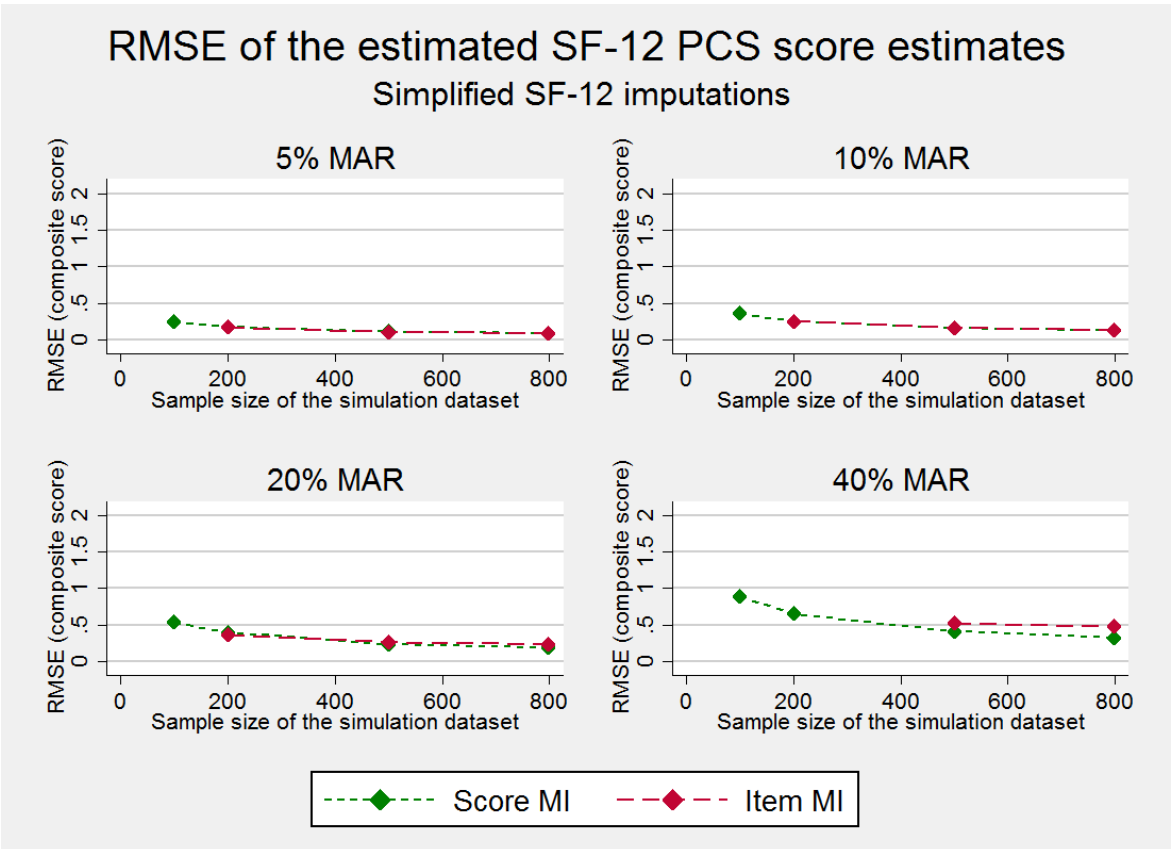


Figure 4-23: RMSE in the SF-12 PCS score estimates

Figure 4-24 and Figure 4-25 show the SEs associated with the SF-12 MCS and PCS scores. As before, the SEs for the two imputation approaches are larger than the true SEs for larger proportion of missing data, to account for uncertainty around the imputed data. Again, the SEs for the item imputation are slightly higher than those for the imputation at the composite score level, although few valid results were obtained for larger percentages of missing data and smaller sample sizes.

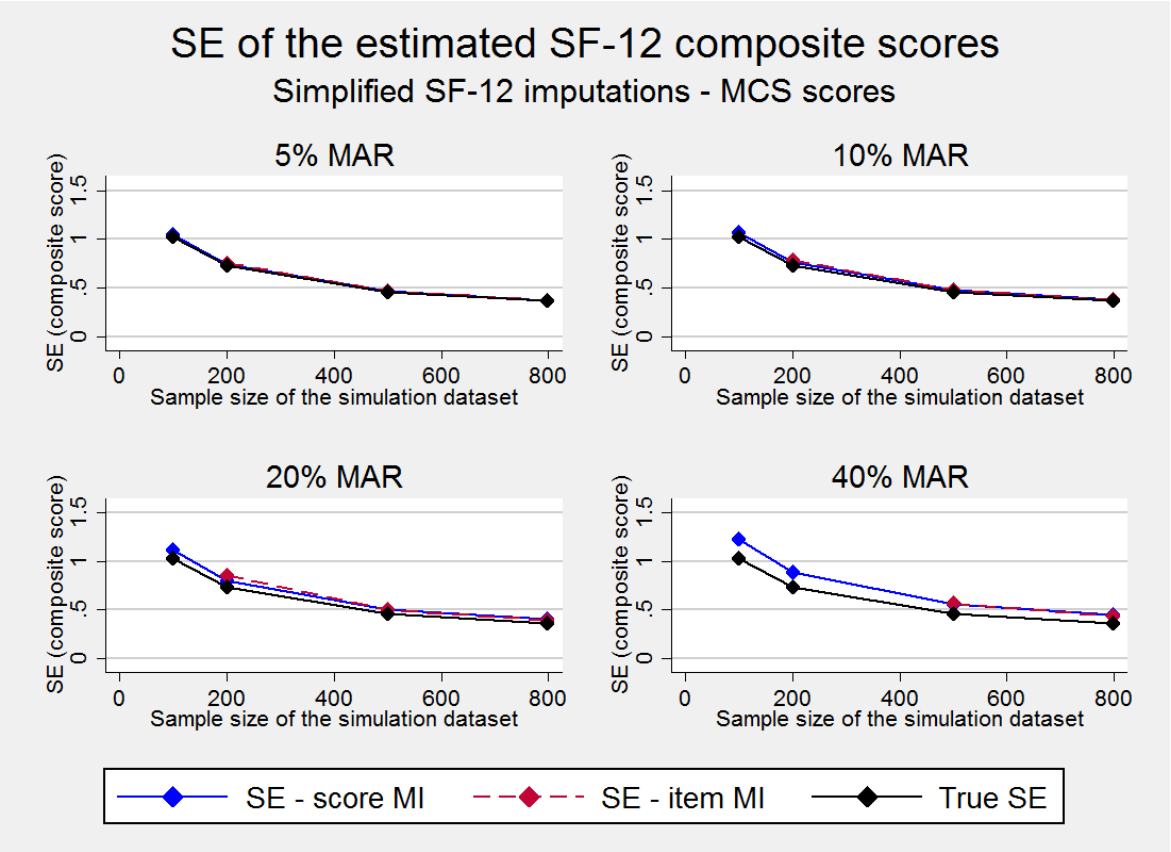


Figure 4-24: SE of the treatment coefficient using the imputed SF-12 MCS score as the outcome variable in the regression model

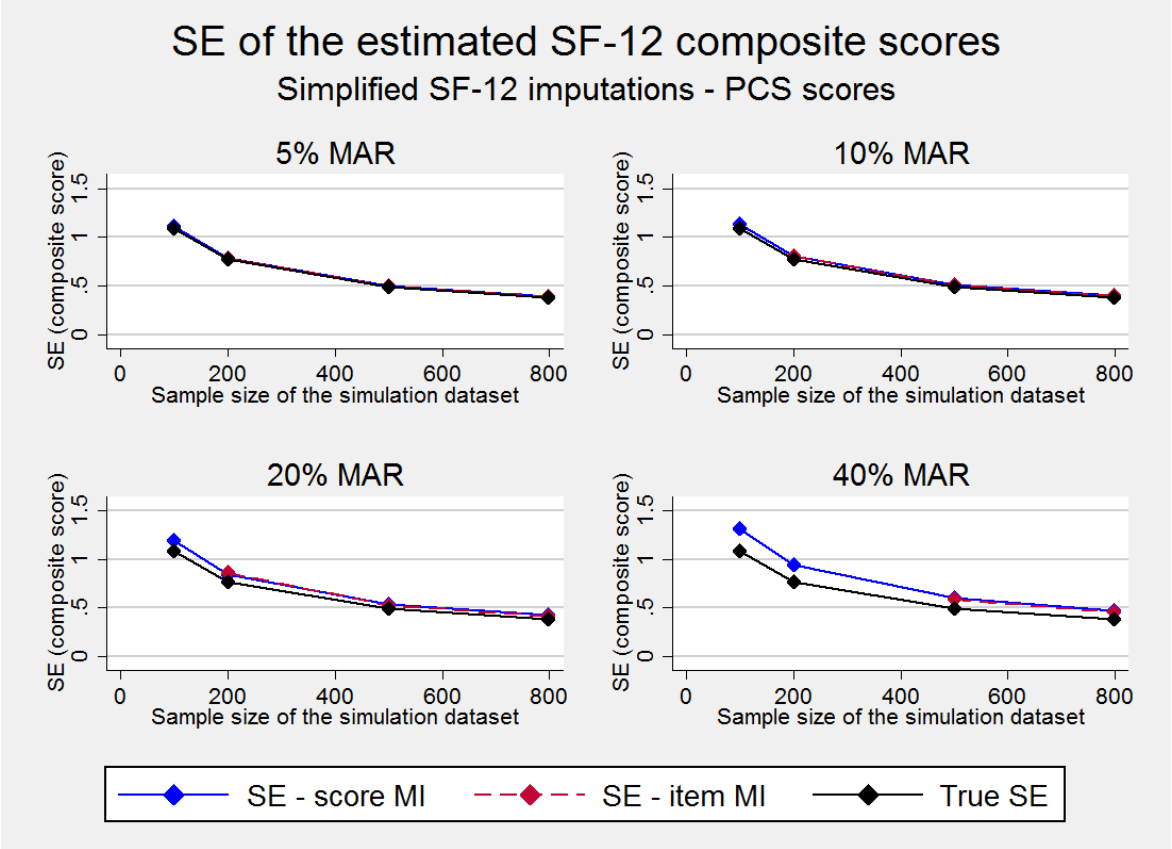


Figure 4-25: SE of the treatment coefficient using the imputed SF-12 PCS score as the outcome variable in the regression model

Considering the treatment coefficient for the regression on the SF-12 estimates

Table 4-20 and Table 4-21 show the mean regression coefficients, their SEs and the MAR and RMSE associated with the estimates. For the base cases, i.e. a sample size of 797, imputation at the item level consistently outperforms both imputation at the subscale level and CCA.

Table 4-20: Estimated treatment coefficients, MAE and RMSE for the SF-12 base case simulations (sample size = 797): results and bias introduced for the treatment coefficient in the linear regression model for the MCS score

SF-12 base cases: MCS score	Mean	SE	MAE	RMSE
True MCS treatment coefficient	0.459	0.665		
5% missing data				
Score imputation	0.454	0.683	0.1215	0.1523
Item imputation	0.465	0.679	**0.1074**	**0.1366**
Complete cases analysis	0.456	0.681	0.1208	0.1514
10% missing data				
Score imputation	0.452	0.701	0.1801	0.2239
Item imputation	0.474	0.694	**0.1519**	**0.1920**
Complete cases analysis	0.461	0.699	0.1796	0.2226
20% missing data				
Score imputation	0.452	0.743	0.2702	0.3352
Item imputation	0.489	0.726	**0.2249**	**0.2852**
Complete cases analysis	0.471	0.740	0.2703	0.3356
40% missing data				
Score imputation	0.436	0.854	0.4157	0.5215
Item imputation	0.530	0.805	**0.3394**	**0.4305**
Complete cases analysis	0.469	0.851	0.4221	0.5291

Table 4-21: Estimated treatment coefficients, MAE and RMSE for the SF-12 base case simulations (sample size = 797): results and bias introduced for the treatment coefficient in the linear regression model for the PCS score

SF-12 base cases: PCS score	Mean	SE	MAE	RMSE
True PCS treatment coefficient	0.556	0.730		
5% missing data				
Score imputation	0.537	0.752	0.1356	0.1751
Item imputation	0.533	0.746	**0.1208**	**0.1500**
Complete cases analysis	0.547	0.749	0.1317	0.1688
10% missing data				
Score imputation	0.508	0.776	0.2006	0.2540
Item imputation	0.509	0.763	**0.1727**	**0.2199**
Complete cases analysis	0.539	0.769	0.1909	0.2397
20% missing data				
Score imputation	0.469	0.827	0.2947	0.3697
Item imputation	0.438	0.799	**0.2610**	**0.3264**
Complete cases analysis	0.530	0.817	0.2819	0.3567
40% missing data				
Score imputation	0.363	0.952	0.5086	0.6321
Item imputation	0.321	0.884	**0.4103**	**0.5183**
Complete cases analysis	0.482	0.945	0.4968	0.6227

Figure 4-26 and Figure 4-27 show the RMSE introduced into the treatment coefficient in the regression model on the SF-12 subscales obtained via the different imputation approaches, as well as from a CCA. From the figure, it can be seen that all three approaches provide similar results for the scenarios with 5% and 10% of missing data. For 20% of missing data or more, a small benefit of imputation at the item level can be observed where these imputations are feasible. The MAE plots shown in Appendix 8.3 confirm these findings.

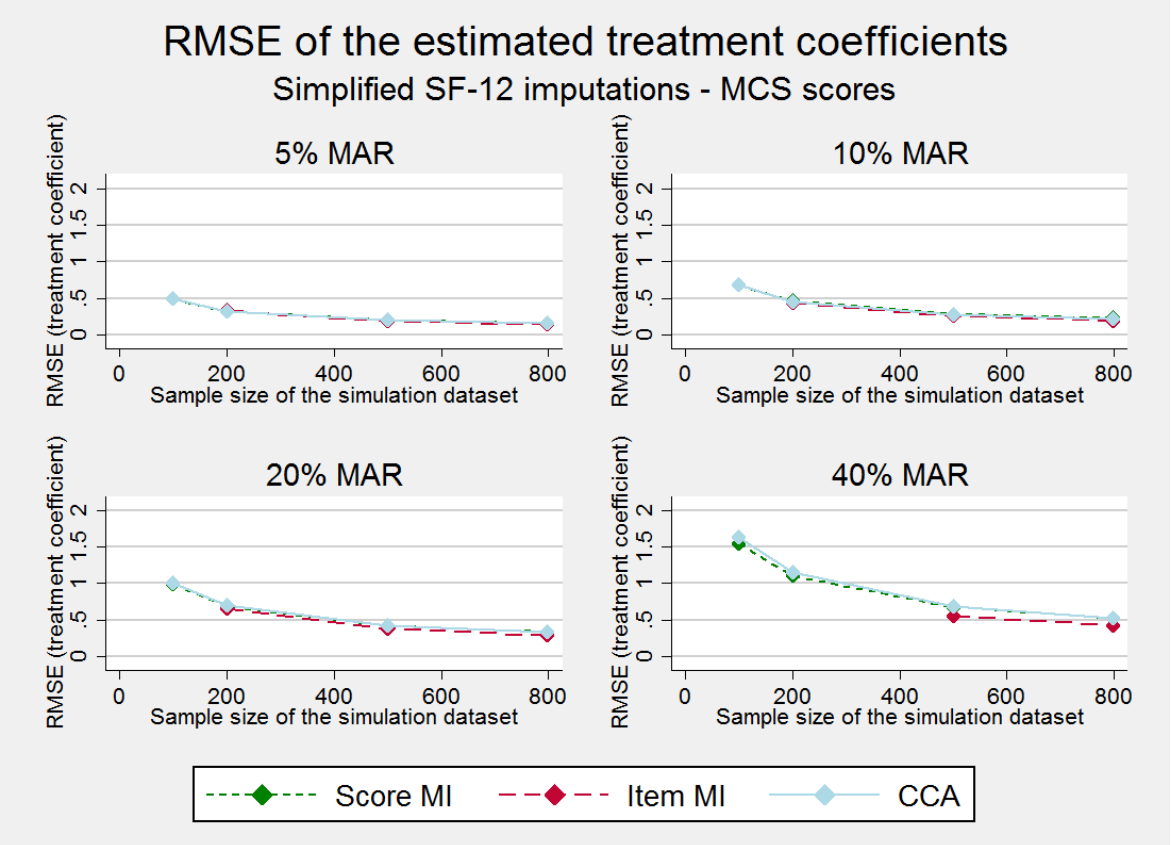


Figure 4-26: RMSE in the treatment coefficient estimates using the imputed SF-12 MCS score as the outcome variable in the regression model

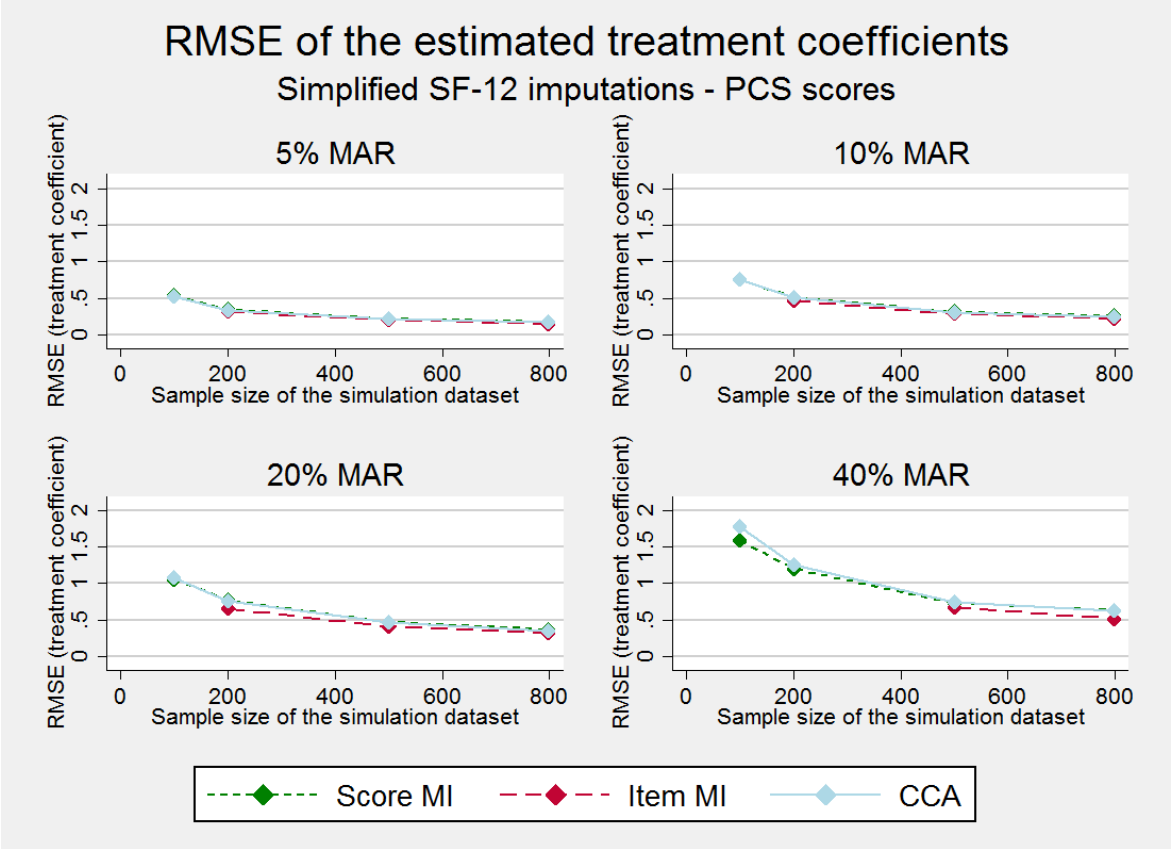


Figure 4-27: RMSE in the treatment coefficient estimates using the imputed SF-12 PCS score as the outcome variable in the regression model

Figure 4-28 and Figure 4-29 show the SEs around the treatment coefficients. Similar patterns to those described previously can be observed, with CCA and MI producing somewhat larger errors compared to the true SEs.

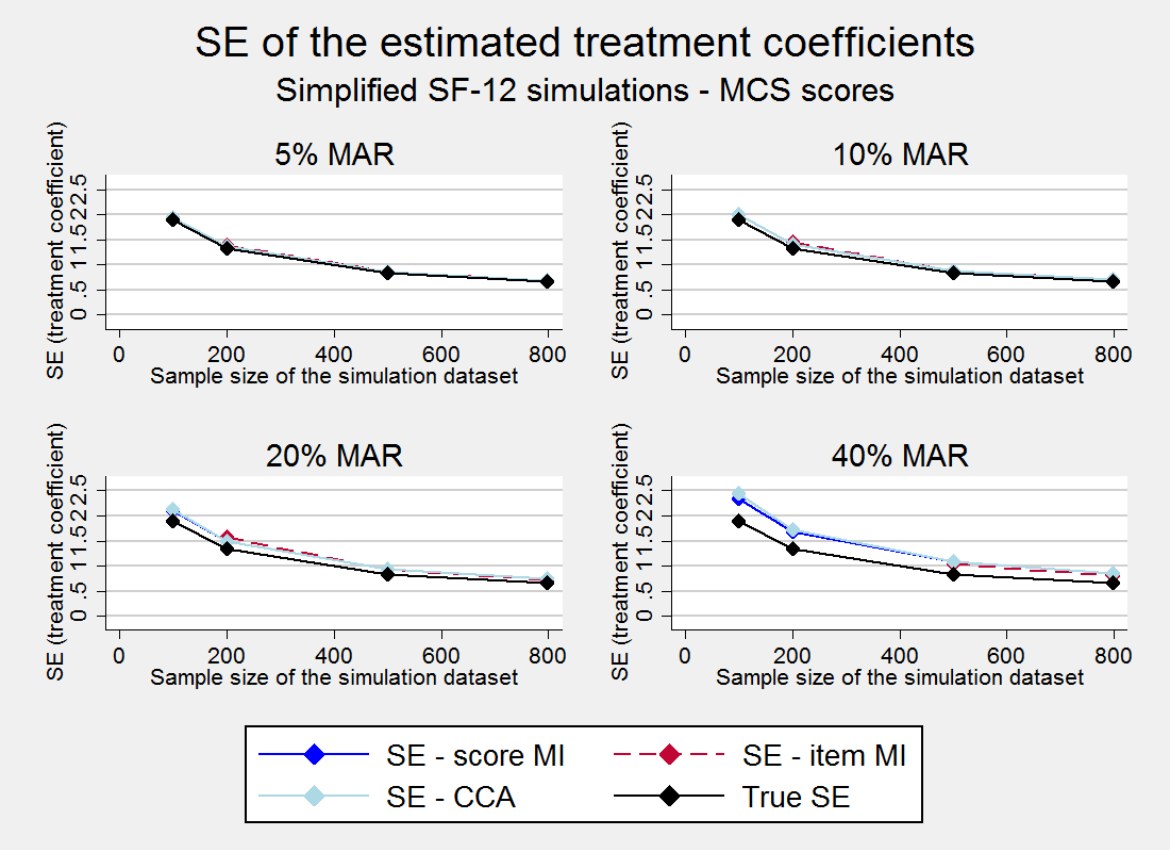


Figure 4-28: SE of the treatment coefficient using the imputed SF-12 MCS score as the outcome variable in the regression model

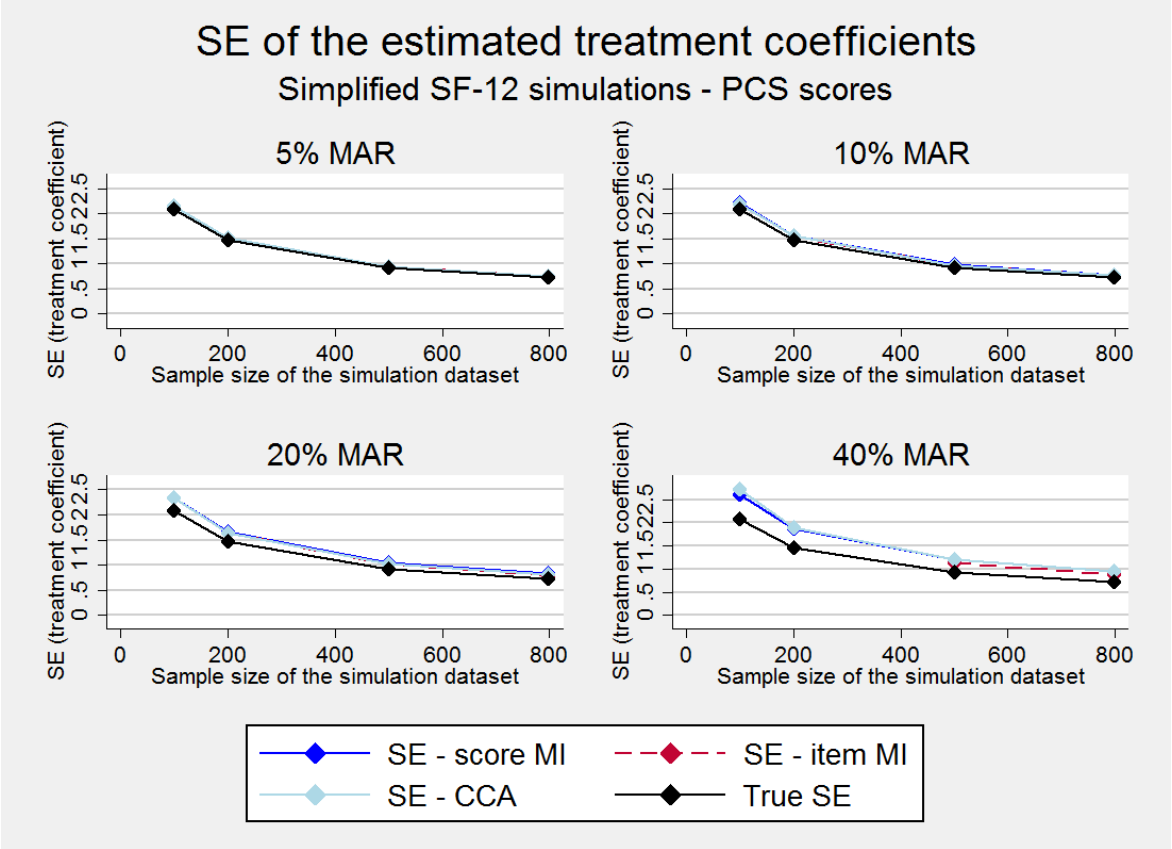


Figure 4-29: SE of the treatment coefficient using the imputed SF-12 PCS score as the outcome variable in the regression model

Comparison of the complex and simplified item imputation model

As for the EQ-5D-3L, the item imputation simulation for the SF-12 were run using a complex model and, due to issues with non-convergence, as a simplified versions. For the SF-12, the complex model refers to the imputation models being run separately by treatment arm, and the simplified version included treatment as a categorical covariate into the imputation model instead. Here, the bias introduced into both models is compared where 1,000 valid results were obtained.

Figure 4-30 and Figure 4-31 show the RMSEs in the SF-12 composite scores when using the complex and simplified item imputation model. Very little difference between two imputation models can be observed for the PCS score, however, the estimation of the MCS score appears lightly more biased for larger proportions of missing data and sample sizes of 500 when the more complex model is used. The corresponding MAE plots, presented in Appendix 8.3, confirm these findings.

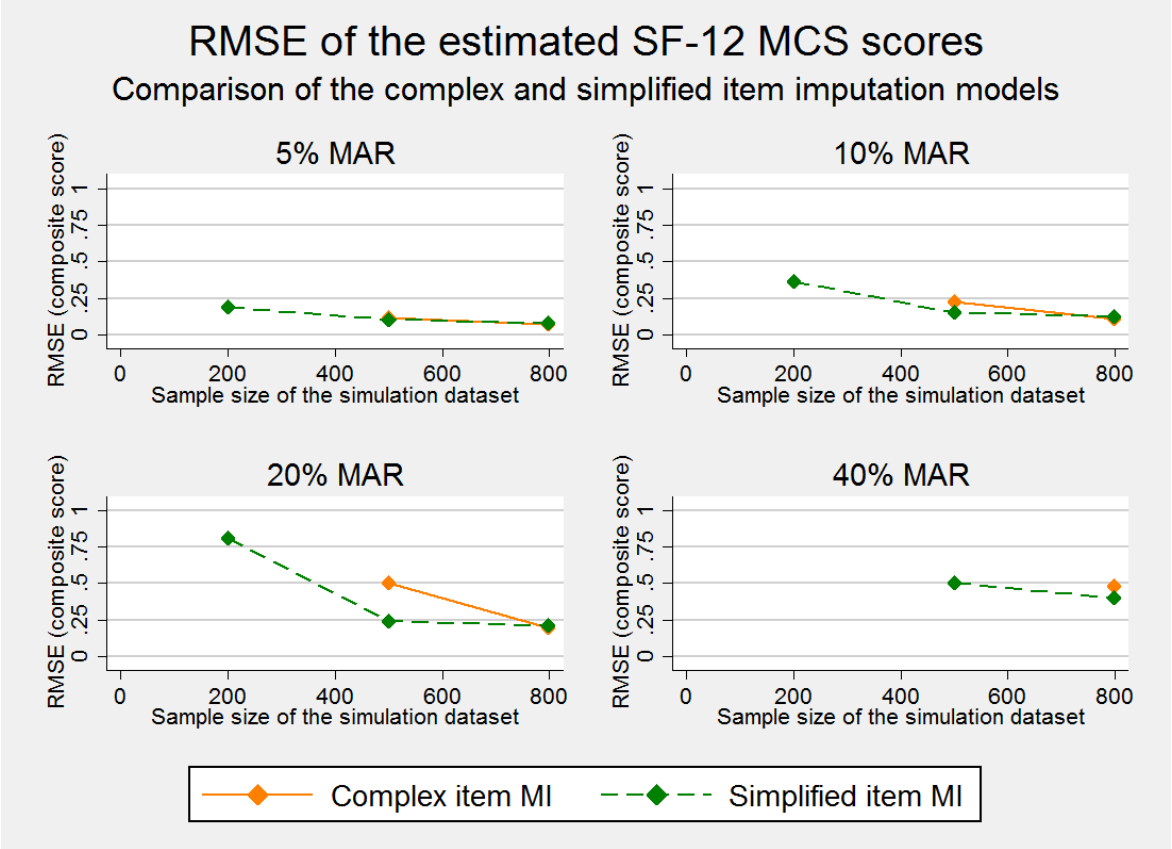


Figure 4-30: RMSE in the SF-12 MCS score estimates – comparing the complex and simplified item imputation model

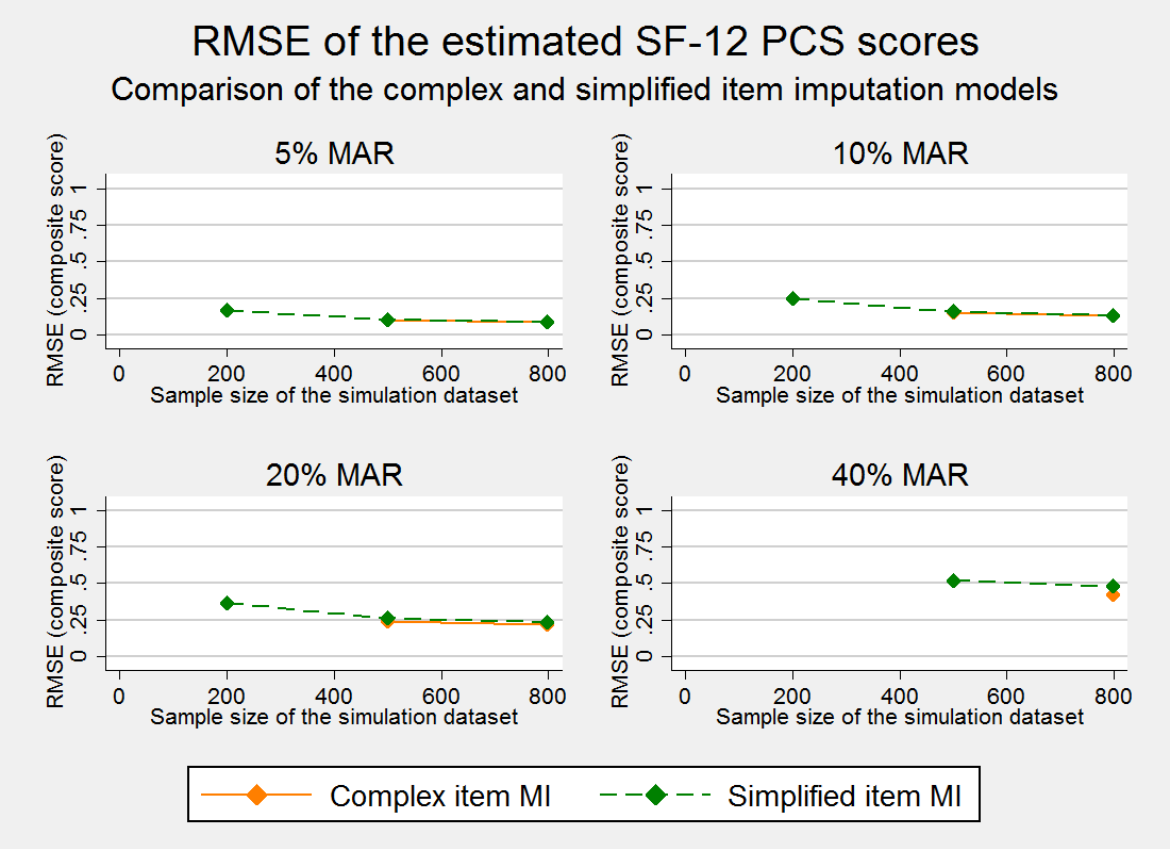


Figure 4-31: RMSE in the SF-12 PCS score estimates – comparing the complex and simplified item imputation model

Figure 4-32 and Figure 4-33 show the RMSE in the treatment coefficient estimates for MI at the item level. Here, some small difference can be observed between the two approaches; the more complex models induce slightly more bias for large proportions of missing data, while also incurring more convergence problems. The graphs for the MAE, shown in Appendix 8.3 support these findings.

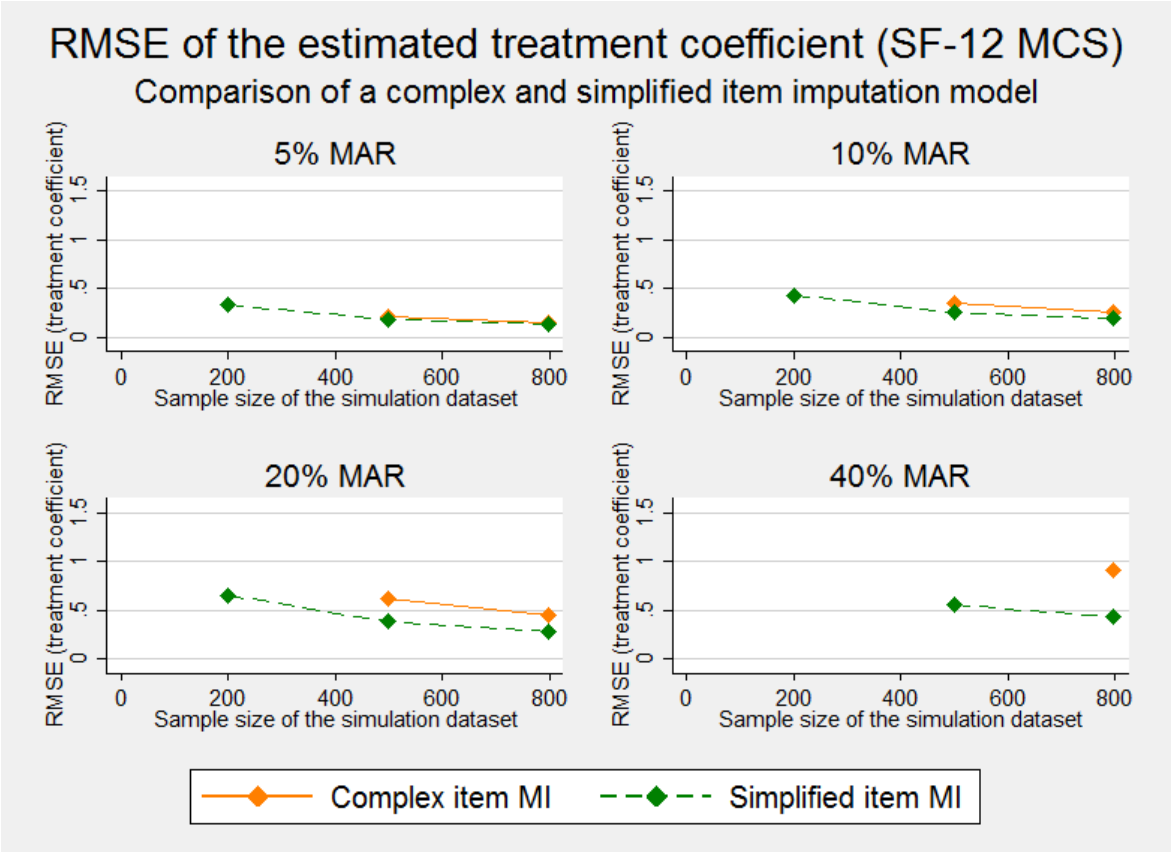


Figure 4-32: RMSE in the treatment coefficient estimates using the imputed SF-12 MCS scores as the outcome variable in the regression model – comparing the complex and simplified item imputation model

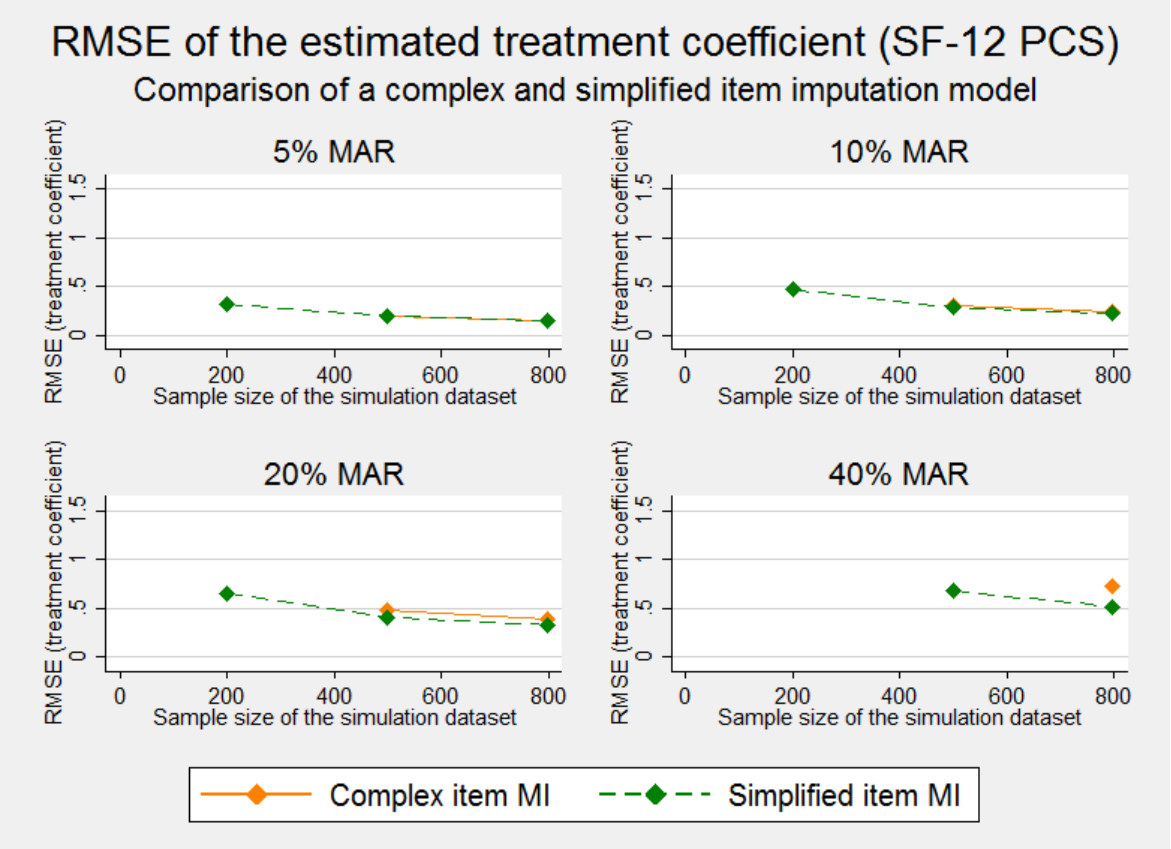


Figure 4-33: RMSE in the treatment coefficient estimates using the imputed SF-12 PCS scores as the outcome variable in the regression model – comparing the complex and simplified item imputation model

4.5.3.9 Assessments of the results compared to simulations using a larger number of iterations

Additional numbers of iterations can add to the robustness of simulation results. In the existing research[52, 112], 500 to 1,000 simulations were used. To maximise the robustness of the simulation results, while balancing the time available to run the simulations, 1,000 simulations were used throughout this chapter, as described in the methods section, 4.4.2. To assess the possible limitations of not opting for a higher number of simulations, the EQ-5D-3L simulation for a sample size of 500 with 10% of missing data were re-run for the MI at the composite score level, this time generating 5,000 simulations. Very subtle differences were observed in the simulation results. However, the key results used in the assessment of the different approaches to handle missing data, namely the RMSEs, MAEs and SEs, figures differed only at the third decimal place for the RMSE of the treatment coefficient for the item imputation and CCA, and at the fourth decimal place for all other estimates.

Therefore, it is believed that the results presented here are robust, and that the number of iterations used in the simulation study was sufficient.

4.5.4 Length of time needed for the simulation work to complete

For the EQ-5D-3L and SF-12 simulations, running time was measured for each simulation iteration, including the generation of missing data, the MI process and the generation of results. Of note, the timings are not entirely consistent, as they are also influenced by unrelated simulations running simultaneously, other computer usage and the computer they were run on.

Generally, the running time for the different imputation approaches varied, with imputations at the composite score and subscale levels running much faster. On average, the simulations for the SF-12 subscale level imputations took 11 seconds to run for a sample size of 797 and 40% of missing data, with running time decreasing for smaller proportions of missing data and smaller samples. The maximum average running time was estimated to be around 30 minutes for one simulation iteration for MI at the item level. As such, complex MI imputations at the item level in an analysis of an RCT with comparable characteristics in terms of sample size and proportion of missing data, and a similar PROM are likely to be performed within 30 minutes. Therefore, the additional running time of these simulations is not considered prohibitive in scenarios where these imputation approaches are considered appropriate.

Additional details on the running times of selected simulations are provided in Figure 4-22 to 4-25.

Table 4-22: Average running time (in seconds) for the EQ-5D-3L simulations – imputation at the composite score level

		Sample sizes of simulated datasets			
		100	200	500	1160
% of missing data in simulated dataset	5%	10.1 (4.0, 30.24)	8.3 (4.0, 12.3)	5.2 (5.0, 8.2)	4.0 (3.5, 7.1)
	10%	12.2 (4.6, 17.0)	6.8. (4.6, 15.2)	6.1 (4.9, 9.0)	4.6 (4.2, 7.6)
	20%	11.5 (4.5, 44.0)	4.8 (4.5, 11.1)	5.7 (5.0, 20.5)	5.4 (4.1, 12.0)
	40%	9.0 (4.4, 12.4)	5.0 (4.6, 19.5)	5.2 (5.0, 12.7)	8.4 (5.2, 14.4)

Table 4-23: Average running time (in seconds) for the EQ-5D-3L simulations – imputation at the item level

		Sample sizes of simulated datasets			
		100	200	500	1160
% of missing data in simulated dataset	5%	n/a	3.1 (0.6, 3.5224)	12.2 (1.4, 41.	Not recorded
	10%	n/a	5.9 (1.0, 63.2)	39.6 (2.4, 139.3)	Not recorded
	20%	n/a	17.4 (1.9, 222.9)	82.2 (4.3, 464.9)	Not recorded
	40%	n/a	n/a	428.9 (6.6, 938, 4)	Not recorded

Table 4-24: Average running time (in seconds) for the SF-12 simplified simulations – imputation at the composite score level

		Sample sizes of simulated datasets			
		100	200	500	1160
% of missing data in simulated dataset	5%	2.8 (0.1, 7.9)	6.4 (6.4, 6.4)	8.1 (8.2, 8.2)	7.2 (7.2, 7.2)
	10%	8.1 (8.1, 8.1)	7.7 (7.7, 7.7)	7.5 (7.5, 7.5)	7.6 (7.6, 7.6)
	20%	8.2 (8.2, 8.2)	6.9 (6.9, 6.9)	6.8 (6.8, 6.8)	10.7 (10.7, 10.7)
	40%	6.7 (6.7, 6.7)	7.3 (7.3, 7.3)	6.9 (6.9, 6.9)	10.6 (10.6, 10.6)

Table 4-25: Average running time (in seconds) for the SF-12 simplified simulations – imputation at the item level

		Sample sizes of simulated datasets			
		100	200	500	1160
% of missing data in simulated dataset	5%	8.2 (6.5, 33.8)	49.1 (7.9, 213.1)	125.4 (8.6, 234.0)	810.8 (56.6, 1201.2)
	10%	17.6 (13.9, 21.8)	180 (15.2, 422.8)	337.4 (20.2, 496.7)	1723.7 (117.6, 2528.8)
	20%	n/a	311.0 (19.2, 485.3)	654.4 (35.6, 1122.9)	843.5 (797.9, 1105.3)
	40%	n/a	n/a	1698.9 (980.1, 3515.7)	484.4 (755.9, 1251.1)

4.6 Discussion

This chapter aimed to compare different approaches to multiply imputing missing PROMs outcome data, i.e. imputing at either the composite score, subscale or item level, and to assess the benefits and disadvantages of these approaches.

While convergence issues have been observed for the more complex item imputation models for small samples, the models have been shown to be feasible for larger sample sizes. Steps to improve the convergence of these models, which can be easily applied with standard statistical software have been implemented and presented. Therefore, item imputation can be considered a possible alternative to imputation at the composite score level in many scenarios. However, the question remains whether it offers any benefits over MI applied to the composite scores; how imputation at the subscale level (where available) compares to both approaches; and how CCAs compare to imputations in the context of regression analysis. The results of this chapter shed some light on this question, but the choice of which approach should be used depends on a multitude of factors, which are discussed in more detail in the following section.

4.6.1 Choice of imputation approach

MI is considered to be one of the most appropriate methods to handle missing data. Therefore, it should be noted that whilst CCA was not shown to be performing considerably worse than the different MI approaches for the estimation of the treatment coefficients, MI may still be preferable. The reason for there being little difference between MI and CCA in this study is likely to be related to the generation of missing data, and the implementation of the MAR algorithm. Many of the variables included in this algorithm were also part of the analysis models, and were hence adjusted for in the CCA. In reality, it

is possible that the MAR mechanism is related to more variables outside of the analysis model. In these scenarios, MI is preferable to CCA, as it offers the opportunity to adjust the imputed data for many variables plausibly both related to the probability of data being missing, and predictive of the missing data values themselves.

As expected, the different MI approaches yielded similar results in certain scenarios; however, differences were observed in some settings. These differences were generally small, i.e. in the treatment coefficient in the OKS regression models, differences of up to one point were observed. The OKS score ranges from 0-48, and differences up to one point lie within the measurement error of the PROM, and also do not exceed the minimal important difference, which are estimated to be four points and five points, respectively[12]. Therefore, the potential for a reduction in bias for the different approaches for handling missing data may not be clinically relevant per se. However, RCTs are often powered to detect small differences in outcome, particularly when two established procedures are compared, and researchers would only expect to observe a small improvement. For instance, the KAT study was powered to detect a 1.5 point difference in the OKS for the patella resurfacing[92]. Thus, even the small differences introduced by choosing one imputation approach over the other could affect the trial conclusions. Therefore, it is important to carefully consider the most appropriate MI approach, taking the following considerations into account. Guidance is provided with regards to four areas of interest, namely the combination of sample size and proportion of missing data, missing data patterns, feasibility of the imputation approach and the planned analysis.

4.6.1.1 Sample size and proportion of missing data

Generally, all MI approaches yielded similar results for large sample sizes. Exceptions to this were the simulations of the SF-12, where item imputation yielded slightly less biased results for the treatment coefficients for both subscales, and also for the MCS score, particularly for larger percentages of missing data.

For combinations of smaller sample sizes (100 – 500 observations) and low percentages of missing data (up to 10%), all MI approaches performed similarly. However, when the amount of missing data was increased to 20% or more, imputation at the composite score and / or subscale level performed better for the OKS simulations using the observed missing data patterns, as well as for the scenarios considering a unit-nonresponse mechanism and a five point treatment difference, as well as for the EQ-5D-3L simulations, in line with previous research[112]. Exceptions to this were the OKS simulations with a simulated missing data patterns of 70% item non-response, where item imputation yielded best results in terms of the treatment coefficient, and the subscale imputation produced the least bias when considering the PROM composite score. Similar results were obtained for the SF-12, for which a high proportion of item missingness was observed.

4.6.1.2 Observed missing data patterns

In line with the hypotheses, all imputation approaches performed similarly for the unit-nonresponse scenario, as well as the scenarios where the missing data patterns followed primarily a unit-nonresponse missing data pattern, i.e. the observed missing data patterns within the OKS (73% simulated unit-nonresponse) and the observed missing data patterns for the EQ-5D-3L (approximately 88% simulated unit-nonresponse).

When the amount of item missingness was increased, item imputation was more likely to be beneficial, especially for smaller sample sizes and larger amounts of missing data. Here, this was shown for the OKS simulations using 70% of item missingness, as well as some of the SF-12 results, where the observed missing data patterns yielded around 44% item missingness.

4.6.1.3 Feasibility of the relevant imputation models

The results of this chapter have shown that imputation at the item level may not be feasible if the MI model is too complex, and the sample size is small, and/or the proportion of missing data is high. However, the results for the EQ-5D-3L and SF-12 show that more complex models do not necessarily yield less biased results. Therefore, simplification of the models to allow imputations at either the subscale or item level are possible. However, simplification should be considered carefully. The exclusion of baseline PROM items may be acceptable where baseline composite scores and/or subscales are included. Similarly, the SF-12 simulations showed no benefit in terms of accuracy when running simulations separately by treatment arm. However, this is likely to be related to the fact that the KAT study showed no statistically significant treatment effect with regards to the patellar resurfacing randomisation[93]. However, where such an effect may exist, where the variance around the imputed variable or the relationship between the imputed variable and covariates may vary between treatments, or where a subgroup analysis including the treatment allocation is important, imputations should be performed separately by treatment arm[58, 120].

The running time of the imputation models may also be a consideration. Imputation at the item level can be much more time consuming, depending on the complexity of the model,

as well as the number of imputations required. However, the maximum time to run a single imputation model, as opposed to a large number of imputation within a simulation context, is unlikely to be considered prohibitive considering the computer power available.

4.6.1.4 Planned analyses

While imputation at the item or subscale level does often not offer a distinct benefit in reducing the bias in the estimates obtained, there may still be cases where imputation at the item or subscale level, as opposed to the score level, is beneficial. This applies to cases where the planned analysis, for RCTs often in the form of a pre-specified statistical analysis plan, includes not only the analysis of the composite scores, but also of the subscales (where applicable) or even the PROM items. If feasible, imputation at the item or subscale level ensures that a common imputation dataset can be used for all analyses related to the relevant PROM. This means that all key analyses are based on the same dataset and make the same assumption about missing data.

While the best choice of imputation approach limits the amount of bias introduced in the study results, increasing amounts of missing data, as well as small sample sizes open up any study to the introduction of considerable bias. Therefore, even the best imputation approach cannot eliminate bias from the study results, and the prospective limitation of missing data within any RCT remains vital, as recommended in the current literature[17, 18, 40, 51]. Sensitivity analyses remain imperative to understand how the trial results change with varying assumptions about the underlying missing data mechanism. The use of sensitivity analysis to assess the robustness of RCT results following analyses adjusting for missing data are discussed further in Chapter 6.

4.6.1.5 Recommendations for choice of MI approach

The following recommendations are based on the above listed considerations:

- Where the estimate of the treatment coefficient is of interest: Imputation at the item level should be used in scenarios with a high proportion of missing data due to item-nonresponse and missing data in 20% of participants or more. Similarly, where applicable, imputation at the subscale level should be used instead of imputation at the composite score level where the missing data pattern in a primarily item-nonresponse scenario allows the calculation of some of the subscales.
- Where the unadjusted estimates of the composite PROMs scores are of interest: item imputation should be avoided for combinations of large proportions of participants with missing data and small sample sizes if missing data is primarily due to unit-nonresponse.
- The choice of the imputation approach in scenarios not covered above should depend on practical considerations. Where little difference exists in the performance of the MI approaches, imputation of the item or subscale level should be performed if subsequent analyses require such data. Otherwise, imputation at the composite score level should be performed as they are more efficient, i.e. their completion time is faster and the risk of non-convergence is lower.
- As per recommended best practice, all variables included in the analysis model, including treatment allocation, and those related to the missing data mechanism and to predict missing outcomes are to be included in the imputation models. To

improve feasibility of the models, baseline PROMs items should not be included in the imputation models.

4.6.2 Novel aspects and limitations of this research

4.6.2.1 Contribution to the literature

The work presented in this chapter validates and expands existing research. Simons et al[112] concluded that imputation at the EQ-5D-3L items and composite scores perform similarly in most scenarios. However, they found that composite score MI was beneficial for small samples and larger percentages of missing data. Also, in scenarios with increased levels of item missingness as opposed to missing data patterns dominated by unit-nonresponse, item imputation was preferable. This chapter replicated these findings for additional PROMs and in a separate dataset. In addition, the benefits and disadvantages of imputation at the subscale level for the OKS were also evaluated.

This research was then extended to investigating the effect of the different imputation approaches, as well as CCA, have on the treatment coefficient (i.e. treatment effect), which is of primary importance in the analysis of RCTs. Hereby, it was shown that imputation at the item level was more beneficial, particularly when higher proportions of data were missing due to item nonresponse. This chapter adds to the findings by Eekhout et al[52], who investigated the effect of imputation on the treatment coefficient of the PROM used as a covariate in a regression model. Therefore, this chapter offers additional information and guidance to researchers faced with missing PROMs data in RCTs.

It was also shown that when MI is applied at the composite score level, using item mean imputation for up to two missing items does not offer benefits in terms of the RMSE over

not using item imputation for the estimation of composite scores, while the treatment coefficients are slightly less biased for large proportions of missing data when mean imputation is avoided. For MI at the subscale level and large percentages of missing data, estimates of the composite scores benefit from avoiding mean imputation, while treatment coefficients are less biased when the scoring manual is followed.

4.6.2.2 Limitations of this research

Although every effort was made to conduct this simulation study as thoroughly and completely as possible, it is not without limitations.

Firstly, a limited amount of missing data scenarios are considered. Maximum sample sizes were restricted to the number of observations in the base cases; however, sample sizes ranging from 100 to over 1,000 participants are thought to be representative of the vast majority of RCT. Future work will consider larger sample sizes, which may increase generalisability of these findings to larger-scale epidemiological research. Arguably, different missing data scenarios, particularly other missing data patterns could also have been investigated. However, the missing data patterns used are based on those observed in the RCTs investigated in Chapter 3. It is believed that these are realistic missing data patterns, which included variations in the amount of unit-nonresponse for the OKS simulations. Additionally, the PROMs used in this chapter are carefully developed and validated questionnaires and are commonly used in clinical research. They generally tend to be completed well, and it was therefore felt that additional item missingness patterns would not contribute much to the practical guidance this chapter aims to provide, as these patterns would be unlikely to be observed in practice.

The simulation models are limited to the KAT dataset. Validation in additional datasets would have been a valuable addition to this chapter, however, access to suitably large RCT or observational datasets utilising similar PROMs could not be obtained within the remit of this thesis. Furthermore, the use of additional PROMs in this validation study might have further confirmed the generalisability of these findings. However, findings are believed to be generalisable to comparable PROMs, based on the fact that findings by Simons et al[112] were replicated, and also confirmed these in additional questionnaires.

The findings are limited to PROMs with up to 12 items, and the PDQ-39, which was considered in Chapter 3 was not used in this simulation study. This is because item imputation was not considered feasible for such a large number of items. Arguably, uncertainty still exists as to the maximum number of items within a PROM for which item imputation would still be considered feasible, which is likely to be related to both the construct of the PROM, as well as the sample size.

MNAR mechanism and misspecification of MI models were not considered in this chapter, although Simons et al[112] reported benefits of MI at the item level over MI at the composite score level for the latter scenario. However, it was felt that MI levels could be misspecified in a number of ways, and that the results from selected misspecifications may not be generalisable to misspecification in general. This is because some variables are much more predictive of the missing data than others. The same principle applies to MNAR scenarios, which could be considered as misspecified MI models, as they are unable to account for important factors that are predictive of data being missing as well as the missing observations themselves. MNAR analysis are best addressed as part of a sensitivity analysis[122-124], details of which are further discussed in Chapter 6.

In addition, it was aimed to keep the simulation scenarios comparable across the different approaches by using consistent seeds. This means that the n^{th} imputation for each scenario is based on the same underlying dataset, using the same random numbers. Since some simulations were skipped due to non-convergence of the item imputations, this intra-comparability does not apply any longer. However, as each simulation estimates the true results, as well as the results obtained from the imputation approaches, the RMSE and MAE directly compare the true results with the imputation estimates, this is not anticipated to have affected the results. Some bias may have been introduced into the results due to the fact that non-convergence may be more likely to occur in datasets with certain characteristics, thus introducing a systematic selection bias into the results for the item imputation. However, this is thought to have an effect only for simulations with very high non-convergence rates (approximately 90% or higher), as discussed in section 4.5.3.1. For these cases, it could be seen that the 'true' OKS scores for the simulations with valid item results are slightly smaller than those obtained for the simulations that did not converge for the item imputations. The reverse was observed for the 'true' treatment coefficients, suggesting some difference in the datasets underlying the composite score and subscale imputations vs. the item imputations. Simulations with higher convergence rates are not thought to be affected.

Some of the non-convergence rates observed in the results are very high. This is likely to be a result of the fact that the MI models were constructed using the full base case datasets, as opposed to smaller sample sizes, for which they may be too complex. The MI models were kept consistent to allow a direct comparison of performance between the scenarios considering different sample sizes and proportions of missing data. In reality, MI

models should be generated so that they are relevant to the dataset under consideration, and take adjust their complexity based on the type of data collected, as well as the amount of data available.

Finally, there is a school of thought that recommends larger numbers of simulations are beneficial. However, as shown in section 4.5.3.9, only very small differences were observed between the simulations run 1,000 times vs. 5,000 times. Therefore, the benefit of running additional simulations was small, and it was not thought that the use of additional simulations would have changed the results of this chapter.

4.7 Conclusions

This chapter compared the performance of MI applied at either the composite score, subscale (where appropriate) or item level for handling missing PROMs outcome data. Advantages and disadvantages of these approaches were also discussed. A number of factors need to be taken into account when deciding on an imputation approach. These include the missing data patterns, sample size and overall percentage of participants with missing data, but also elements such as the planned analysis and PROMs characteristics.

The choice of the most appropriate MI approach helps to limit the bias introduced into the trial results, but does not lessen the need to take steps to limit missing data occurrence within RCTs. Appropriate sensitivity analysis to assess the impact of missing data on the trial results when changing the underlying assumptions about the missing data mechanism remains imperative.

Chapter 5 : A comparison of statistical approaches for analysing missing longitudinal patient reported outcome data in randomised controlled trials

5.1 Introduction

Most RCTs assess participants' outcomes at multiple follow-up assessments after baseline, as also shown in Chapter 2[39]. Appropriate statistical analysis, such as mixed-effects linear regression, can handle such longitudinal data. These statistical models adequately account for the fact that multiple observations are obtained from each of the participants, and are likely to be more correlated than the values between different individuals[125-127].

For some RCTs, it is appropriate to select a primary analysis time point, and perform a cross-sectional analysis of the outcomes at this time point, similar to the analyses considered in Chapter 4. However, even where the primary analysis of an RCT requires a cross-sectional analysis, often the secondary analyses includes longitudinal analyses.

Longitudinal data analyses are therefore an important tool in medical research, however they can also be affected by missing data. Indeed, longitudinal follow-up data can be subject to both monotone missingness, where no additional observations are available for a participant after a specific follow-up time point (e.g. arising from drop-out or withdrawal from the trial), and intermittent missingness, whereby some observations are unavailable followed by subsequent obtained data[20]. These factors mean that often only a small subset from a longitudinal dataset can be used in a CCA. Therefore, it is important to understand which statistical approaches are most appropriate for the analysis of longitudinal RCT data, particularly with a focus on PROMs, and this chapter aims to further explore this issue. Similarly to any analysis with missing data, approaches such as listwise/ pairwise deletion

and simple imputation methods, including LOCF, are known to be likely to introduce bias into the study results, and are therefore discouraged[20, 128-130]. This chapter explores the benefits and disadvantages of common approaches to handling missing data[58, 131], namely multiple imputation, maximum likelihood methods and inverse probability weighting. Bayesian approaches to handling missing data are beyond the scope of this project and therefore not discussed further.

This chapter is structured as follows: the statistical approaches to handle missing outcome data in longitudinal settings used in this chapter, i.e. maximum likelihood, multiple imputation and inverse probability weighting are introduced in section 5.2. Also, an overview of the existing literature reviewing and assessing the comparative performance of these methods is presented. Following on from this background of the topic area, the hypotheses and objectives of this chapter are detailed in section 5.3.

Section 5.4 presents a 'motivating example', i.e. a case study in which the different statistical approaches are applied to an example dataset to demonstrate how results are affected by the choice of different statistical methods to address missing data within a longitudinal follow-up.

The methodology and set-up of the simulation study are described in section 5.5, and results presented in section 5.6. The impact of this research, together with its strength and limitations is discussed in section 5.7, and the conclusions of this chapter are summarised in section 5.8.

5.2 Statistical methods for analysing longitudinal RCT data and their comparative performance

As stated previously in this thesis, many RCTs collect outcome information repeatedly over the follow-up period. A review by Bell et al[25] suggests this number is close to 80%, and the review performed for Chapter 2[39] confirms this, reporting that 82.3% of reviewed RCTs measured outcomes repeatedly. However, the primary analysis for many RCTs commonly only uses a single follow-up measure. Indeed, Bell at el[25] estimate that only 18% of RCTs account for longitudinal outcomes in their analysis.

Bell et al[25], Ibrahim et al[132], Enders et al[128] and others[20, 62, 67, 133] provide useful overviews of statistical methods available for the analysis of longitudinal data with missing observations. They strongly discourage ad hoc missing data methods such as deletion methods, where observations with missing data are discarded, or single imputation methods, which are likely to cause bias and generate overly precise standard errors. However, reviews of the literature indicate that these methods are commonly used in the analysis of RCTs in general, and also for those with longitudinal follow-up data[23, 25, 39, 58, 62, 99, 128]. Instead, the authors propose the use of model based approaches, assuming MAR data, including maximum likelihood estimation (ML), multiple imputation (MI), and inverse probability weighting (IPW), which are considered further in this chapter.

5.2.1 Review the statistical methods covered in this chapter

This section aims to introduce the different approaches to handling missing longitudinal data which are to be considered within this chapter, namely ML, MI and IPW. Notably, each of these approaches could be applied in a number of different ways. Therefore, the general ideas underlying each approach are introduced, together with the implementations considered within this chapter, whereby emphasis is put on implementation using established commands in standard statistical software. Extensions to the approaches considered in this chapter have been discussed in the literature, but they are not available in standard statistical software[134].

Following the overview of the different methods, their perceived benefits and disadvantages, as referred to in the current literature, are discussed, together with their comparative performance based on previous studies, where available.

The statistical analysis of longitudinal data needs to take into account the correlation between follow-up measures within individuals. The follow-up measures are likely to be more closely correlated within than across individuals, and not taking into account this model structure results in the derivation of misleading standard errors, and hence study results[125, 127].

In the literature, these models are referred to by various names, including repeated measures models, multilevel mixed-effects linear regression models, population average models or growth curve models[20], depending on context and different terminology in the research field. Within this chapter, the term multilevel mixed-effects linear regression model (i.e. multilevel mixed-effects model) is used to refer to the analysis of longitudinal data. Stata, as well as other statistical software packages, offer a range of commands to appropriately analyse longitudinal data.

The statistical approaches to address missing data investigated in this chapter are ML, MI and IPW. These approaches were chosen due to their robust underlying methodology, ease of implementation in standard statistical software (such as Stata), and their use in RCT analyses to date, as shown in past reviews[25, 39, 99]. Therefore, it is assumed that researchers and statisticians are more familiar with these approaches and more comfortable using them, but would benefit from a direct comparison of those methods specifically within a PROMs and RCT context.

5.2.1.1 Maximum likelihood

Within a ML framework, parameter estimates are obtained through an iterative process so as to maximise the likelihood of producing the sample data[128]. In the presence of missing observations, the information from the available data points are utilised to make inferences about the missing data, i.e. to 'implicitly impute the unobserved data'[20] in order to estimate the model parameters in the maximum likelihood approach[21, 59, 128]. As such, this approach assumes that the missing data is following a MAR mechanism, taking into account the covariates included in the multilevel mixed-effects model. The analysis model and the model for addressing the missing data are implemented simultaneously. This has been described as one of the advantages of ML methods over other statistical approaches for handling longitudinal missing data. The simultaneous modelling also means that ML can have increased efficiency over MI, which imputes data and fits the analysis model in two separate steps and IPW, which firstly models the missingness model and then applies the weighted analysis model[135]. Also, the simultaneous modelling means that the information contained in the data is not overestimated, and therefore true estimates for the standard errors are obtained[122]. However, this process also means that the modelling of the missing

data is restricted to the variables included into the analysis model. For RCTs, the analysis models are typically limited to baseline PROMs measurements and key baseline data, possibly also used in the randomisation procedure, which restricts the MAR assumption to those variables. It is possible to take into account auxiliary variables outside the analysis model to better model the missing data mechanism[60, 134]; however, these approaches are not commonly implemented in standard software[59], and therefore not considered within this chapter.

In Stata's 'mixed' command for multilevel mixed-effects linear regression analysis, models are fitted by default using the ML approach, and the ML approach to handle missing data in a longitudinal setting is therefore straightforward to implement using standard statistical software.

The maximum likelihood estimation assumes a normal distribution when estimating parameters from incomplete data[128] and performs best when this condition is met.

5.2.1.2 Multiple imputation

The methodology and advantages of MI[50, 58], especially its ability to take into account auxiliary variables that may affect the probability of data being missing, but which are not included in the analysis model, have been described in the previous chapters.

Following on from Chapter 4, the MI approach could easily be extended from the cross-sectional setting to simultaneously impute missing longitudinal outcome data, thus incorporating the information from other follow-up points into the estimation of missing data where available.

There has been some research on utilising MI to handle missing longitudinal data, and the key findings and opinions are summarised here. Generally, it is found that MI is a feasible

approach for handling missing longitudinal data, and has been applied in a number of case studies and comparative methodological work[99, 132, 136]. It is described as one of the most commonly applied modern missing data methods[137], and has been found to be able to handle non-monotone missing data patterns well[138]. As discussed previously, MI allows, where appropriate, for a large number of auxiliary variables to be included into the imputation model[28]. Hence, in contrast to the ML model, modelling the missing data mechanism is not restricted to only those variables included in the analysis model, meaning the MAR assumption is more likely to be met[58]. However, some caution is recommended when applying multiple imputation to longitudinal missing data, as this approach can be very computationally intensive and may 'break down because of co-linearity and over-fitting'[139]. Consequently, some researchers propose utilising only adjacent follow-up time points in their imputation models[139, 140]. However, this may be less prohibitive for the use of MI in RCTs, which often have a limited number of follow-up time points compared with studies consisting of routinely collected clinical data or large cohort studies.

A potential disadvantage of applying MI to longitudinal missing data is that standard statistical software does not currently allow MI to formally account for a multilevel structure (i.e. longitudinal data clustered within individuals). However, such imputations are possible in specialist statistical software, such as MLWin[141]. Within this chapter, multiple imputation by chained equation (MICE) is used, as in Chapter 4. By simultaneously imputing the missing data for the different follow-up time points, the correlation between these longitudinal measurements can be taken into account. It is anticipated that this approach adequately models the multilevel structure of the data. SEs derived from this approach will be considered in subsequent simulation work.

The findings from Chapter 4 show that multiple imputation at the item level can be advantageous within a cross-sectional analysis in certain scenarios. However, this previous work also showed that multiple imputation models can be problematic when too many values are imputed simultaneously, and that convergence may not be achieved when categorical data to be imputed has a low number of counts in some of the categories. Therefore, it was deemed infeasible to attempt imputation at the item level in a longitudinal context with several follow-up times points, and this chapter implements on MI at the composite score level for the relevant PROMs.

5.2.1.3 Inverse probability weighting

Traditionally, inverse probability weighting has been used in survey studies. Observations with a low probability of being included in the survey are given a higher weight in the analysis model to mitigate against bias introduced by the sampling design[142, 143].

However, the technique has also gained popularity in the handling of missing data[135, 144, 145]. The rationale for using inverse probability weighting (IPW) in the presence of missing data is that the subset of participants with complete data may not be representative of the full dataset. Under IPW, a complete cases subset of the data is analysed with cases weighted differentially so as to adjust for the bias introduced by a conventional CCA[21]. This is similar to the way survey data is adjusted for the probability of individuals with certain characteristics being included in the sample. As such, complete cases that have a low probability of being observed due to missing data in comparable participants are given a higher weight in the analysis compared to those with a high probability of being observed. This accounts for the participants that cannot be included into the analysis model due to missing data.

In implementing IPW models the full dataset is first used to estimate the missingness model, i.e. the probability of each case having complete data is estimated using a logistic regression model using appropriate data as covariates. These variables are chosen in line with the assumed MAR mechanism. The complete case analysis is then weighted by the inverse of these fitted values, i.e. cases which are less likely to provide complete data are given more weight in the analysis[135, 144]. Robust standard errors are recommended in order to account for the uncertainty around the weights[144]. It should be noted though that the use of the complete cases stipulates that some observed data are discarded for the analysis model, a potential disadvantage compared to ML and MI, which utilise all observed data. Through the inclusion of a separate missingness model, similarly to MI, the assumed MAR mechanism is made more realistic as auxiliary variables outside of the analysis model can be included. However, especially when many follow-up time points are included in the study, and a combination of monotone and intermittent missing data patterns are observed, the complete cases subset used for analysis may only include a small proportion of the original dataset. Additionally, if the missingness model differentiates only between those participants with all follow-up completed versus those with at least one incomplete follow-up time point, it is restricted to the baseline information only. Especially if the relationship between responsiveness and missingness changes over time, any such missingness model would be suboptimal. Also, the above described IPW does not account for potential differences in cases with different missing data patterns; i.e., it is feasible that those with intermittent missing data differ from those with monotone missing data, i.e. those who withdraw from the trial or are lost to follow-up. To address this limitation, stratification of the IPW approach have been suggested[135], but this is only thought to be appropriate if the number of missingness patterns are small, as sufficient numbers of

participants are needed within the strata to estimate the missingness model. Therefore, the stratification approach to IPW may be less feasible for RCTs than epidemiological studies, which often include larger patient populations or more auxiliary variables to adequately distinguish between individuals with different missingness patterns. Also, it is thought that IPW may produce biased estimates if a small amount of participants have very low probabilities of being observed, resulting in large weights being attributed to them[67]. For these reasons, stratified IPW are considered further in this chapter.

5.2.2 Comparative performance of the methods based on the literature

The literature offers few direct comparisons between ML, MI and IPW, and some papers which do compare them use a case study approach, applying the different approaches to an example dataset, rather than evaluating their comparative performance in relation to a known truth using a simulation study. This section aims to provide an overview of the literature on the comparative performance of ML, MI and IWP.

In the context of missing longitudinal PROMs data in RCT settings, Kang et al[146] compared repeated measures analysis of variance (whereby all participants with incomplete cases were dropped), generalised estimating equations (GEE), and a mixed-effect model repeated measures approach without using ML estimation. They concluded that the mixed effect model repeated measures approach appeared to provide the most appropriate results in terms of accuracy and SEs. However, their study is lacking a comparison of alternative methods for handling missing data, such as MI or IPW.

MI and ML methods were compared in their ability to predict missing QoL data that was initially missing but subsequently recovered through additional data chasing, by Fielding et al[136]. The authors reported that MI methods (particularly predictive mean matching) performed better in terms of bias than the ML methods in the majority of cases considered.

Ferro et al[147] compared MI and ML methods with listwise deletion in a simulation study. They raised concerns that longitudinal MI may not be feasible for studies with many follow-up time points, and found that ML methods and cross-sectional MI provided comparable results. However, their paper focussed on intermittently missing data and large sample sizes (N > 5,000) and it is unclear if these conclusions are appropriate for RCT analyses, which often consider much smaller sample sizes.

Twisk et al[148] compared the performance of mixed-effects linear regression models to MI, and found that MI offered no "obvious gains" over mixed-effects linear regression. However, their MI model only accounted for variables in the analysis model, and benefits may be observed for MI models taking into account additional variables.

Kadengye et al[149] compared direct likelihood methods to simple and multiple imputation approaches in a multilevel context. They reported that the direct likelihood methods produced almost unbiased estimates under MCAR and MAR, and concluded that they performed better than the imputation methods when item scores are missing in a multilevel setting. The authors also raised concerns about the ability of imputation methods to adequately account for multilevel or clustering characteristics of the dataset, a limitation that may also apply to longitudinal data. They commented on the ease of implementation of the likelihood approach, but conceded that imputation methods are beneficial in certain situations, specifically when confronted with a combination of missing outcome and explanatory data.

Carpenter et al[143] reviewed methods to increase the efficiency of IPW, namely doubly-robust methods, i.e. methods which are still valid for some misspecification of the missingness model under certain conditions, and compared these to multiple imputation in a simulation study. They concluded that MI and doubly-robust IPW perform similarly when both the missingness model and imputation model are specified correctly. However, IPW is less sensitive to some misspecification in the models (i.e. either the missingness model or the analysis model) and may therefore be more appropriate in some scenarios. The authors also concluded that both MI and IPW are likely to outperform ML approaches if not all variables needed for the MAR assumptions are included in the analysis model.

Kenward and Carpenter[28] argued in their discussion that MI may offer little advantage over ML based methods for longitudinal missing data under the MAR assumptions. However, they reiterated that MI has some distinct advantages over ML, in that it can handle very complex patterns of missingness, while also being able to take into account variables outside the analysis model. Finally, MI was also recognised to be a convenient tool for suitable MNAR sensitivity analyses.

Instead of considering their comparative performance, much of the existing literature has focussed on assessing the value of modifications or extensions to each methods in terms of its ability to handle missing (outcome) data in a longitudinal context. For instance, Plumton et al[150], who considered simplified MI models where the theoretical best MI model is computationally infeasible due to large numbers of variables or items to be included into the model. Eekhout et al[60] similarly proposed that while item imputation has been shown to be beneficial in some circumstances[52, 112], there are occasions when summary information is preferable so as to not overcomplicate the imputation model in a longitudinal context.

In their research, Biering et al[140] and Carpenter et al[26] focussed on the implementation of MI methods for longitudinal data. Biering et al[140] proposed the use of observations adjacent to the missing data to be included in the imputation model, and explored different approaches of handling death during the follow-up, an indicator for which can be added to the MI model. Carpenter et al[26] focused on the handling of protocol deviations, namely withdrawal and noncompliance, which resulted in missing data, considering treatment effects in a best-case scenario (de jure estimand) and pragmatic scenarios representing real practice (de facto estimand).

Donneau et al[138] compared different approaches to MI, namely joint modelling and fully conditional specification approaches, for binary and PROMs data. They concluded that modelling approaches based on a multivariate normal model can be used for the imputation of binary data, although the fully conditional specification approach performed better in some circumstances.

Reviewing the benefits and drawbacks of using IPW for handling missing data, Seaman et al[144] suggested weight stabilisation and augmentation (where some of the missing data is imputed) to improve efficiency of the IPW process. Seaman et al[145] extended the augmented IPW models further in an additional publication, in which they acknowledge MI and IPW as two commonly used approaches. While they consider MI to be more efficient, they acknowledged that in situations where many data points are missing for an individual, researchers may feel more confident using IPW. As an alternative, they suggested using a combination of MI and IPW. They found this approach useful in the presence of large proportions of missing data following different missing data pattern, as it offered advantages over the sole use of MI or IPW when the imputation models were misspecified. Doidge[135] compared and discussed different approaches to IPW, specifically stratified and stepped IPW, as also discussed in section 5.2.1.3.

The search strategy used to identify the publications discussed above is presented in Appendix 9.

5.3 Research hypothesis and objectives

Based on the methodology and literature discussed above, hypotheses and aims for this chapter are presented in this section.

5.3.1 Hypothesis for this chapter

As described above, ML approaches have been found to perform well for MAR data where the analysis model includes all variables required to meet the MAR assumption. However, the results of Chapter 3 have demonstrated that missing data mechanisms can be complex and difficult to disentangle. Therefore, it is likely that additional variables beyond the analysis model explain missingness, a fact that could introduce bias into the ML model. For this reason, it is hypothesised that the MI model, as well as the missingness model utilised in IPW have a better chance of adhering to the MAR assumption. However, the IPW models used in this chapter are set up such that they cannot distinguish between participants without any follow-up data, and those with monotone missing follow-up data, or with intermittently missing follow-up data. Therefore, they are at risk of not utilising some available information, rendering them less efficient. Also, the missingness model for the IPW is based on a logistic regression model, which may fail to converge, especially for small datasets and where categorical explanatory variables have low counts in some of their categories.

For the MI models, no concerns about convergence exist, as the PROMs data is imputed at the composite score level, and linear regression models are very likely to be fitted adequately[v]. For the reasons outlined above, it is hypothesised that the MI models to

[v] Alternatives to linear regression models can be implemented in MI, but are not the focus of this chapter.

address missing longitudinal data are least biased, as they are more likely to account for the missingness mechanism, and utilise all available data. It is hypothesised that the ML models yield more bias than the MI approaches, as they may not be able to fully account for the MAR assumptions. However, as they utilise all data available, the hypothesis is that they produce less biased estimates, particularly for larger proportions of missing data, than the IPW approach. While IPW is able to account more fully for the MAR assumptions than ML, it may discard large proportions of the available data, as its main analysis focusses on those trial participants with complete follow-up data.

Similarly, it is hypothesised that ML and MI produce unbiased standard errors, particularly for small to moderate amounts of missing data. Due to the loss of potentially large amounts of data in IPW, it is hypothesised that the IPW SEs are overestimated.

5.3.2 Objectives for this chapter

This chapter aims to compare the performance of the following approaches for handling missing longitudinal PROMs in RCT analyses:

- Maximum likelihood (ML)
- Multiple imputation (MI)
- Inverse probability weighting (IPW)

Performance is assessed by considering the amount of bias introduced into the treatment coefficients of the analysis model, presented as RMSE and MAE.

Within this research, performance of the different analysis approaches is considered for a wide range of sample sizes, changes in the prevalence of missing data, and differences in missing data mechanisms.

This chapter also aims to provide guidance for researchers on how to handle missing PROMs outcome data in longitudinal analyses based on simulation studies reflecting real-life RCTs.

5.4 Motivating example

In this motivating example, the three above described statistical approaches for handling longitudinal PROMs data with some missing follow-up are applied to a subset of the KAT study. The analysis approaches are expected to produce different results, thus reemphasising the need to further investigate the advantages and disadvantages of the different approaches.

The KAT trial has previously been introduced in Chapter 3. Briefly, this is an RCT assessing the clinical and cost effectiveness of new developments in knee replacement[92, 93], in which a number of PROMs, including the OKS, the EQ-5D-3L and SF-12, were collected at baseline, three months, one year and yearly thereafter. In this example, the above described approaches for handling missing longitudinal data are applied to a subset of the KAT trial including data for the initial five years of follow-up. The subset of data used is not representative of the full trial dataset and is not aiming to reproduce the trial analysis.

5.4.1 Analysis population

The KAT subset to be used in this exploratory work contains 1334 observations, and is the subset of participants with complete data for the baseline variables which are to be included in either the analysis models or the missingness models (i.e. the imputation model or the model to determine the weights to be used in IPW). These variables are baseline OKS, age, gender, BMI, height, ASA physical status (a system to assess patients' fitness prior to surgery)[101, 102] and size of recruiting centre.

Missing data patterns have been described previously in Chapter 3, and were found to be a mixture of monotone and intermittent missing data, with the percentage of missing PROMs scores at each follow-up time point increasing with time. In the full dataset,

approximately 30% of participants had some missing follow-up data. Within this subset, approximately 26% of participants have some missing follow-up data. Additional details can be found below.

5.4.2 Analysis model used

The multilevel mixed-effects linear regression model aims to compare the OKS over time between the treatment groups. It includes a random intercept and a random slope, thus allowing the intercept and slope to differ between individuals, which was a more appropriate fit than the random intercept model alone. The model is fitted using a maximum likelihood approach and use an unstructured covariance structure.

The model is adjusted for randomisation allocation, baseline OKS, gender and age, and follow-up time point (implemented as a dummy variable for each year, using the one-year follow-up as the reference category). Non-linear terms or interactions are not included into this model, although the KAT trial analysis included interactions between treatment and time[93]. This is because in this chapter, the main focus is on the bias introduced into the treatment coefficient by the different analyses.

This repeated measures model is used as the analysis model for all three approaches to handle missing data. In addition to the covariates in the analysis model, the MI and missingness model for the IPW also include for ASA physical status, BMI (instead of height, which was previously used), and size of the randomising centre. These variables were found to be associated with the probability of data being missing in the exploratory work in Chapter 3, as well as outcomes, and were used as part of the MAR mechanism and MI models in the simulation work in Chapter 4. The MI and IPW models additionally adjust for two post randomisation variables: whether or not the participants received their allocated

intervention (coded as yes vs. no/not specified) and whether the participants experienced complications during their operation (coded as yes vs. no/not specified). As the imputation uses a MICE approach, missing data for each follow-up OKS is also based on OKS values collected at the other follow-up time points. 25 imputations were used in this example for the MI model.

The weights for the IPW model were derived by firstly generating a binary variable indicating if participants either completed follow-up vs. having some missing data over the follow-up. Then, a logistic regression model adjusting for the above mentioned variables was then fitted to this indicator variable to establish the probability of complete follow-up being available for each participant. The inverse of these probabilities is then used as weights in the analysis model, which is fitted to the subset of complete cases (i.e. participants for whom all follow-up data is available). Robust standard errors are used for the analysis model for the IPW.

Stata's *mixed* and *mi impute* commands were used to implement the models. The code for each analysis is presented in Appendix 10.

5.4.3 Results from the different analyses

Table 5-1 shows the percentage of missing OKS data at each relevant follow-up time point overall and by treatment arm. In this subset of the trial data, all baseline data is non-missing.

Table 5-1: Missing OKS data by treatment arm in the subset used in this exploratory analysis

	OKS cannot be calculated/ is missing		
Time point	No patellar resurfacing (N = 660)	Patella resurfacing (N = 674)	Total (N = 1334)
Baseline	0 (0%)	0 (0%)	0 (0%)
3 month	33 (5%)	51 (7.57%)	84 (6.3%)
1 year	42 (6.36%)	43 (6.38%)	85 (6.37%)
2 years	65 (9.85%)	82 (12.17%)	147 (11.02%)
3 years	76 (11.52%)	81 (12.02%)	157 (11.77%)
4 years	75 (11.36%)	83 (12.31%)	158 (11.84%)
5 years	96 (14.55%)	92 (13.65%)	188 (14.09%)

As described above, Figure 5-1 reiterates that a wide range of missing data patterns have been observed within the KAT trial, covering both intermittent and monotone missing data patterns.

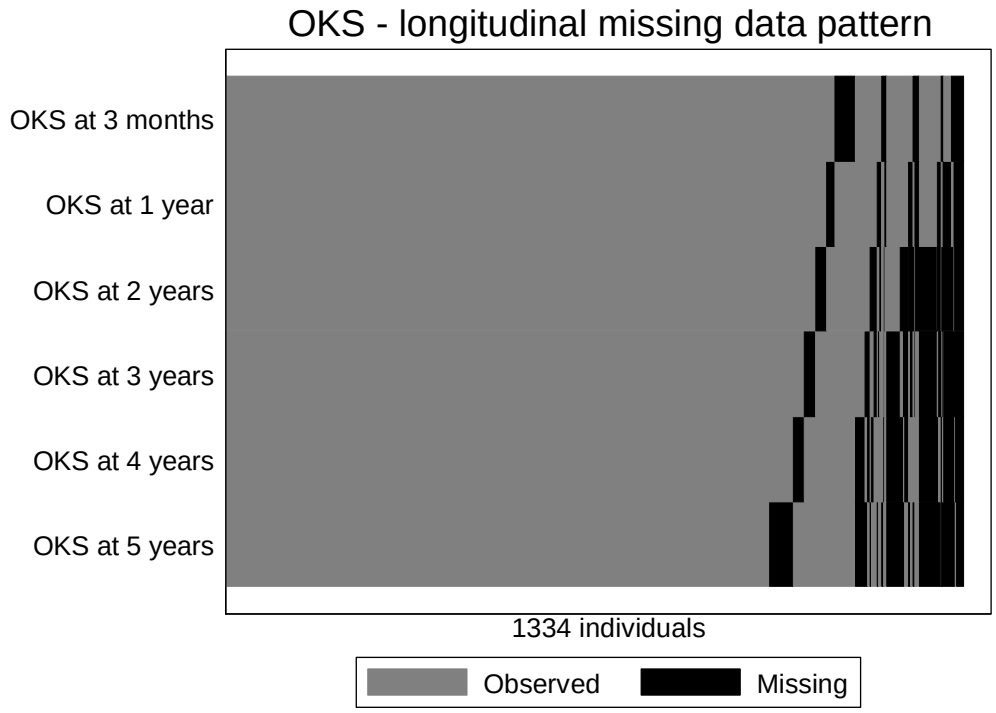


Figure 5-1: Longitudinal missing data pattern in the subset of the data used in the motivating example

Table 5-2 shows the OKS values observed in the KAT data subset over time by treatment arm and overall. A pronounced increase in the OKS is observed from baseline to the three month follow-up. The OKS increases further to one year; change thereafter is less marked, and a slight decrease in the OKS can be observed after year two. However, all the average changes after one year are small and unlikely to be clinically relevant. This data also suggest that the difference between the treatment arms is small.

Table 5-2: OKS over time in the subset of KAT data

	No patellar resurfacing [mean (95% CI)]	Patellar resurfacing [mean (95% CI)]	Total [mean (95% CI)]
Baseline	18.16 (17.57, 18.74)	18.54 (17.98, 19.10)	18.34 (17.95, 18.75)
3 month	30.92 (30.18, 31.67)	31.62 (30.85, 32.39)	31.27 (30.74, 31.80)
1 year	34.77 (33.97, 35.56)	35.02 (34.25, 35.78)	34.90 (34.34, 35.44)
2 years	35.60 (34.77, 36.43)	35.98 (35.17, 36.78)	35.79 (35.21, 36.37)
3 years	34.97 (34.12, 35.82)	35.67 (34.86, 36.48)	35.32 (34.74, 35.91)
4 years	34.52 (33.68, 35.36)	35.14 (34.28, 36.00)	34.83 (34.23, 35.43)
5 years	34.83 (33.80, 35.65)	35.15 (34.29, 36.00)	34.97 (34.37 ,35.58)

The results and parameter estimates from the three different models to handle the missing PROMs data in the longitudinal context of the KAT trial are shown in Table 5-3. The estimates of the treatment coefficients derived from the different analysis approaches, which is the key focus of any RCT analysis, did not differ statistically significantly from zero for any of the analyses, in line with the main trial analysis[91, 93]. The estimates from the ML and MI analysis appear similar in their magnitude, while that derived from the IPW analysis is smaller, although not significantly different. The number of participants, the total number of observations and observations per participant included in the model also depend on the modelling approach chosen. This reiterates the fact that MI includes all randomised participants into the model, ML includes all those for whom data has been collected for at least one follow-up time point, and IPW utilises only complete cases. Here, the IPW approach only includes 985 out of 1335 participants (73.78%), while also discarding some of the observed information.

Table 5-3: Model results for the different analysis approaches

	Coefficient, SE, (95% CI)		
	ML model	MI model	IPW model (robust standard errors)
Participants included	1320	1334	983
Observations included (per participants)	7185 (5.4 – average)	8004 (6)	5898 (6)
Randomised intervention (patella resurfacing)	0.369, 0.459, (-0.529, 1.268)	0.339, 0.461, (-0.563, 1.242)	0.087, 0.526, (-0.944, 1.119)
Baseline OKS	0.46, 0.032, (0.397, 0.523)	0.463, 0.032, (0.4, 0.526)	0.434, 0.041, (0.353, 0.514)
Time			
3 months vs. 1 year	-3.409, 0.188, (-3.776, -3.041)	-3.389, 0.193, (-3.767, -3.011)	-3.513, 0.223, (-3.95, -3.075)
2 years vs. 1 year	0.57, 0.191, (0.195, 0.946)	0.537, 0.191, (0.162, 0.911)	0.66, 0.183, (0.302, 1.018)
3 years vs. 1 year	0.102, 0.204, (-0.297, 0.502)	0.088, 0.208, (-0.32, 0.497)	0.243, 0.219, (-0.186, 0.671)
4 years vs. 1 year	-0.444, 0.222, (-0.88, -0.008)	-0.446, 0.219, (-0.876, -0.015)	-0.316, 0.247, (-0.799, 0.168)
5 years vs. 1 year	-0.502, 0.247, (-0.987, -0.018)	-0.423, 0.245, (-0.903, 0.057)	-0.347, 0.25, (-0.838, 0.144)
Gender (male)	0.267, 0.484, (-0.682, 1.215)	0.244, 0.486, (-0.709, 1.197)	0.161, 0.56, (-0.936, 1.259)
Age	0.054, 0.029, (-0.003, 0.112)	0.057, 0.029, (0, 0.115)	0.042, 0.037, (-0.031, 0.115)
Constant	22.177, 2.103, (18.056, 26.298)	21.907, 2.098, (17.794, 26.02)	24.353, 2.705, (19.052, 29.654)
Unstructured covariance matrix			
Standard deviation (time)	0.014, 0.001, (0.013, 0.015)	0.014, 0.001, (0.013, 0.015)	0.013, 0.001, (0.012, 0.014)
Standard deviation (Intercept)	8.109, 0.191, (7.743, 8.492)	8.145, 0.194, (7.773, 8.535)	8.014, 0.218, (7.598, 8.452)
Correlation (Time, intercept)	-0.186, 0.04, (-0.263, -0.107)	-0.193, 0.039, (-0.268, -0.115)	-0.199, 0.045, (-0.286, -0.11)
Standard deviation (Residuals)	4.588, 0.048, (4.495, 4.683)	4.635, 0.053, (4.533, 4.74)	4.505, 0.098, (4.316, 4.702)

The results between the ML and MI model are likely to differ because the MI includes variables beyond the analysis model, including post-operative complications, and whether or not the allocated procedure was received. These variables are related to both the probability of data being missing, as well as the OKS outcomes, and may therefore contribute to different model results being obtained. The IPW model uses considerably fewer observations than both of the other models. The inclusion of only participants who have completely observed follow-up data creates a biased subset, as these participants have higher OKS scores at all follow-up time points compared to those participants with some missing follow-up data. This is reflected in the increased values of the estimate of the constant in the model. The weights used in the IPW model may not be able to fully compensate for the selection bias introduced by focussing on the subset of participants without missing follow-up data.

In conclusion, the data presented in Table 5-3 demonstrate that the different statistical approaches produce different results, including different magnitudes for the treatment coefficient, on which decision making for health care pathways may be based. Although it is true that the differences in the treatment coefficients for this example are unlikely to be of clinical significance, other RCT scenarios may yield very different results. From this example alone, it is impossible to conclude which results are the most appropriate. It is unknown which of the parameter estimates are closest to the true parameter estimates. In addition, the true underlying missing data mechanism is unknown, and therefore it is impossible to determine if the assumptions made for each model are appropriate, or if bias has been introduced through inappropriate assumptions. Therefore, a simulation study further investigates the advantages and disadvantages of the different approaches.

5.5 Simulation methodology

5.5.1 Rationale for performing a simulation study

As discussed in Chapter 4, a simulation study is performed to assess the performance of the different statistical approaches in relation to a known truth[113], and investigate the benefits and drawbacks of the different statistical approaches to handling missing longitudinal PROMs data for different scenarios, varying the sample size and overall percentage of participants with missing data. The work in this chapter also follows relevant guidance on simulation studies[113], as described in Chapter 4, to produce reproducible and robust results.

5.5.2 General simulation procedures

All programming for this research is performed using the statistical software Stata version 14. The code for the relevant programmes can be found in the Appendix 11. 1000 iterations of the simulation are run for each scenario considered.

5.5.2.1 Simulation scenarios to be investigated

This simulation study uses data from the KAT study[93], focussing on participants who consented to the patellar resurfacing randomisation. Of the 1715 participants randomised to this comparison, 983 participants have complete data for the clinical baseline covariates utilised in this simulation study, as well as for the OKS at baseline and all follow-up visits to five years. The data for these 983 participants is considered to be the base case dataset, and forms the basis for this simulation study.

Sample sizes considered

The performance of the ML, MI and IPW are compared for a range of sample sizes, namely 100, 250, 500, 750 and 983 participants, whereby the latter is the total number of participants in the base case. For each simulation, datasets of the relevant size are subsampled from the base case dataset, as in Chapter 4.

Proportions of missing data considered

As well as different sample sizes, the simulation study also investigates model performance based on different proportions of missing data. MAR data are imposed on 10%, 20%, 30%, 40%, 50% and 60% of the population for each of the sample size scenarios.

Missing data patterns, mechanisms and treatment effects investigated

All missing data generated within this simulation study follows a MAR mechanism. For most simulations, the missing data pattern implemented follows the eight most frequently observed data patterns, as identified in section 4.4.3.2, which are a mixture of intermittent and monotone missingness. Additional scenarios consider monotone missingness only, as well as a 'stronger' MAR mechanism, which is simulated by making the variables outside the analysis model more predictive of the probability of outcome data being missing. More detail of the simulation of these MAR mechanisms is provided in section 5.5.3.

The above described simulation scenarios are generated from the base case dataset, which reflects the data as observed in the KAT study. One additional scenario involved the introduction of a five point treatment effect, as the KAT trial did not show a significant difference between the two trial arms[28]. By adding up to three points to the outcome scores in the patellar resurfacing arm, and subtracting up to three points from the OKS

outcome scores in the no patellar resurfacing arm, as appropriate, i.e. such that the upper and lower range of the OKS was not exceeded, an overall treatment effect of approximately five points was generated.

A final set of simulations assesses the effect of including other PROMs data that are correlated with the OKS outcome variable in the MI and IPW models. Here, the SF-12 and EQ-5D-3L data are added at baseline, as well as for all follow-up assessments. In order to keep the sample size of the base case dataset at 983, missing data in the SF-12 and EQ-5D was imputed using Stata's *mi impute* command utilising baseline and PROMs follow-up data to produce a single imputation for each missing value.

Details of the generation of the base case dataset and the missing data mechanism imposed on the full dataset are provided in section 5.5.3.

5.5.3 Generation of the datasets to be used in the simulation

5.5.3.1 Base case dataset

All simulations are based on a complete cases subset of the KAT participants included in the patella resurfacing comparison, consisting of participants with complete data for the relevant baseline variables (treatment allocation, baseline OKS, age, gender, BMI, ASA physical status, centre size, post-operative complications and an indicator for delivery of the allocated intervention) and no missing data for the relevant PROMs follow-up data up to the five year assessment. This dataset contains 983 observations.

5.5.3.2 Simulation of missing data within the simulation datasets

Of 1526 participants randomised to the patella resurfacing vs. no patella resurfacing and not reported to have died at the five-year follow-up, 30% (459) had some missing follow-up data for the OKS. Within these 459 participants, a large number of different missing data patterns were observed, as demonstrated in Figure 5-2.

The longitudinal missing data patterns shown in Table 5-4 are the most commonly observed patterns.

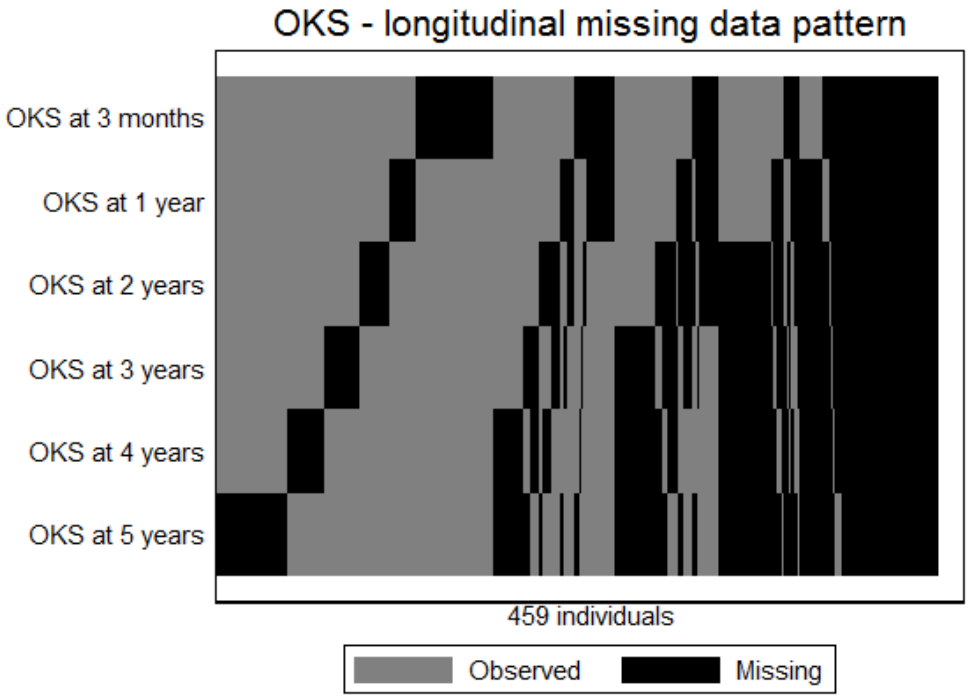


Figure 5-2: Observed longitudinal missing data pattern for the OKS in the KAT trial

Table 5-4: Missing data pattern imposed on the complete cases subset of the KAT trial data – missing the OKS

Missingness pattern	Total	True %	% used in simulation	Cumulative %
No follow-up data available	62	13.51%	22.06%	22.06%
Only three month data missing	49	10.68%	17.44%	39.50%
Only five year data missing	46	10.02%	16.37%	55.87%
Data available to year one	34	7.41%	12.10%	67.97%
Data available to year two	26	5.66%	9.25%	77.22%
Only four year data missing	23	5.01%	8.19%	85.41%
Only three year data missing	22	4.79%	7.83%	93.24%
Data available to year three	19	4.14%	6.76%	100.00%

Replication of the observed missing data mechanism

The eight most commonly observed missing data patterns reported in Table 5-4 are imposed onto the complete cases dataset, according to the relative frequency of each, using an algorithm based on that presented by van Buuren et al[116], and described in detail in Chapter 4. In brief, this algorithm allows the simulation of realistic MAR data, enables the user to vary the overall percentage of missing data, and specify the missing data pattern, as well as the MAR mechanism. Within the algorithm, the variables included in the MAR mechanism are used to predict each participant's probability of having missing data. The coefficients for this equation were obtained in exploratory work whereby eight logistic regression models were fitted (one for each pattern of missing data). The outcome variable used was a binary variable indicating whether or not a participant exhibited the relevant pattern of missing data. The coefficient estimates for each covariate were used in the algorithm to generate missing data. The code for this algorithm is shown in Appendix 11.1.

Simulation of a 'stronger' MAR mechanism

One variation of the simulation considers a 'stronger' MAR mechanism, whereby the influence of variables on the probability of follow-up data being missing is increased for variables outside the analysis model, and decreased for variables contained in the analysis model. This is done by increasing the magnitude of their coefficients. Specifically, the coefficients for the covariates outside the analysis model are increased three-fold, while those contained in the analysis models are halved. The coefficients used in this algorithm for each missing data pattern can be seen in Appendix 11.2.

Simulation of monotone missingness only

Finally, a scenario of solely monotone missingness, i.e. whereby missing data is due to participants dropping out of follow-up or being lost to follow-up, is also generated. The KAT trial data used for this chapter comprises a total of six follow-up points; in this simulation scenario, participants are equally likely to drop out of the follow-up at any one of these time points. The regression model used in previous simulations to estimate the likelihood of participants having no observed follow-up (first missing data pattern in Table 5-4) is used for all drop-out patterns. The Stata code used to generate these drop-out pattern can be seen in Appendix 11.3.

General consideration when simulating missing data

As mentioned in Chapter 4, the *mi impute* code for the MI model requires some missing data in all variables to be imputed, i.e. here all follow-up OKS data. Therefore, if by chance (likely only in small datasets with very low percentages of missing data) no cases of complete loss to follow-up are generated, one such case is added to ensure that the MI code runs without error.

5.5.3.3 Statistical models implemented

Analysis models

The analysis model for the ML, MI and IPW approach is the multilevel mixed-effects linear regression model described in section 5.4.2. The same analysis models are used for all the different simulation scenarios.

MI model

The variables included in the MI model have been described in section 5.4.2. The OKS data for the different follow-up time points are imputed simultaneously using a MICE approach with predictive mean matching using the nearest neighbour approach. All covariates associated with the MAR mechanism are included in the imputation models, which are run separately by treatment arm, where feasible[58, 120].

For the simulation scenario where SF-12 PCS score and EQ-5D-3L data are added to the MI and IPW models, the MI models include all baseline and follow-up data for these PROMs. As all of the additional PROMs data are complete, the data at each follow-up time points are added to the explanatory variables of the MI model to inform the imputation of missing OKS data. The additional variables increase the complexity of the MI models, and thus the

risk of non-convergence. To prevent non-convergence as far as possible, MI models for the simulation scenarios with a total sample size of 100 are simplified, i.e. the MI model uses randomised treatment as a covariate, instead of running separate models by treatment. The Stata code for the MI models can be seen in Appendix 12.

IPW model to calculate weights

The implementation of the IPW model approach has been described in section 5.4.2. All variables associated with the MAR mechanism are included in the logistic regression model to calculate the weights. For the simulation scenario where SF-12 PCS score and EQ-5D-3L data are added to the MI and IPW models, the calculation of the weights reflects this additional information. The PCS score and the EQ-5D-3L composite score at baseline and five years are added to the logistic regression model used to calculate probabilities of participants having complete outcome data.

Depending on the distribution of data, particularly within the categorical variables and percentages of missing data simulated, the logistic regression models do not always converge. Therefore, for some of the iterations of the simulation study, no valid results could be obtained for the IPW model. The Stata code for the IPW models can be seen in Appendix 12.

5.5.4 Data generated from the simulation to assess the comparative performance of the statistical approaches to handling missing data

5.5.4.1 Data generated from base case dataset and observations with imposed missing data

In this chapter, the emphasis lies on the bias introduced into the treatment coefficient within the multilevel mixed-effect models. Therefore, the treatment coefficients, as well as their standard errors are recorded for the full dataset, as well as the ML, MI and IPW analyses.

In addition, the number of instances where no valid results can be obtained from the IPW model, due to non-convergence of the underlying logistic regression model are collected, together with instances where the MI model does not converge.

5.5.4.2 Data generated to assess each model's performance

As in Chapter 4, the performance of each analysis approach compared to the true results (i.e. the results from the datasets without missing data) is assessed using the root mean square error (RMSE) and mean absolute error (MAE). Calculations are in line with those presented in Chapter 4.

5.6 Results

This section presents the results of the simulation study considering different analysis approaches when confronted with missing data in a longitudinally collected PROMs. Results are presented here for RMSE, and corresponding graphs showing the MAEs are shown in Appendix 13. The x-axis of the graphs presents the different sample size scenarios, while the y-axis shows the RMSE or MAE. For clarity, separate plots are provided for the different proportions of missing data.

As in Chapter 4, a general trend that can be observed is that bias in terms of both RMSE and MAE increases both with increasing proportions of missing data, as well as decreasing sample size for each analysis approach. Uncertainty around the results, as measured by the SE of the treatment coefficient, also increases with larger proportions of missing data and smaller sample sizes. Results for the individual simulation scenarios are presented under the relevant headings below.

5.6.1 Feasibility of the different analysis approaches

The ML approach was able to obtain valid results for the multilevel mixed-effects linear regression model in each simulation iteration. Negligible proportions of the MI models did not converge, and no valid results were obtained for these instances (shown in Appendix 13.1). As discussed above, the MI model was simplified for the simulations considering sample sizes of 100 when additional data for the SF-12 and EQ-5D-3L were included into the MI models, and information on non-convergence is provided in Appendix 13.1. Problems, however, were encountered in the estimation of the weights used in the IPW approach, where the logistic regression models did, on occasion, not converge. Table 5-5 shows the percentage of simulations (out of 1,000) for which no valid results could be

obtained for the IPW approach for the scenarios based on the observed data and MAR mechanism. It can been seen that valid results could not be obtained for approximately 21% of the simulation scenarios with a sample size of 100 and 10% of missing data. However, the percentage of non-valid results are 3.2% for the scenarios with 20% missing data and a sample size of 100, and 1.5% with 10% of missing data and a sample size of 250. The failure rates for all other scenarios are below 1% or 0. All IPW models for sample sizes above 250 yielded valid results (not shown in this table).

Table 5-5: Percentage of simulations for which no valid results for the IPW approach could be obtained – simulation using the observed data and observed MAR mechanism

Percentage of participants with simulated missing data	Sample size	
	100	250
10%	20.9%	1.5%
20%	3.2%	0%
30%	0.8%	0%
40%	0.1%	0%
50%	0.1%	0%
60%	0.4%	0%

Similar pattern of failure rates were observed for the simulations considering different MAR mechanisms and underlying data; they are shown in Appendix 13.2.

5.6.2 Simulations using the observed data and observed MAR mechanism

The first simulation scenario considered the observed patterns of missing data and observed MAR mechanism. Figure 5-3 shows the RMSE for the treatment coefficient. Both the ML and the MI approaches perform very similarly irrespective of the sample size or proportion of missing data. The IPW approach performs consistently worse than both ML and MI in terms of RMSE, with differences more pronounced for smaller sample sizes. Similar results are seen for MAE, as presented in Appendix 14.

Figure 5-4 shows the SEs around the different analysis approaches and different simulation scenarios. For smaller proportions of missing data, the SEs estimated from the ML and MI models are almost identical to the true SEs calculated from the full datasets without missing data. They increase only slightly compared to the true SE even when up to 60% of participants have some simulated missing data. The SE obtained for the IPW approach is similar to the true SE for 10% of missing data, and also for 20% when the sample size is large. In the other scenarios, the SE for the IPW is consistently higher than the true SE and those obtained from ML and MI. As with the RMSE, discrepancies in the SE from the IPW approach are also more pronounced for smaller sample sizes.

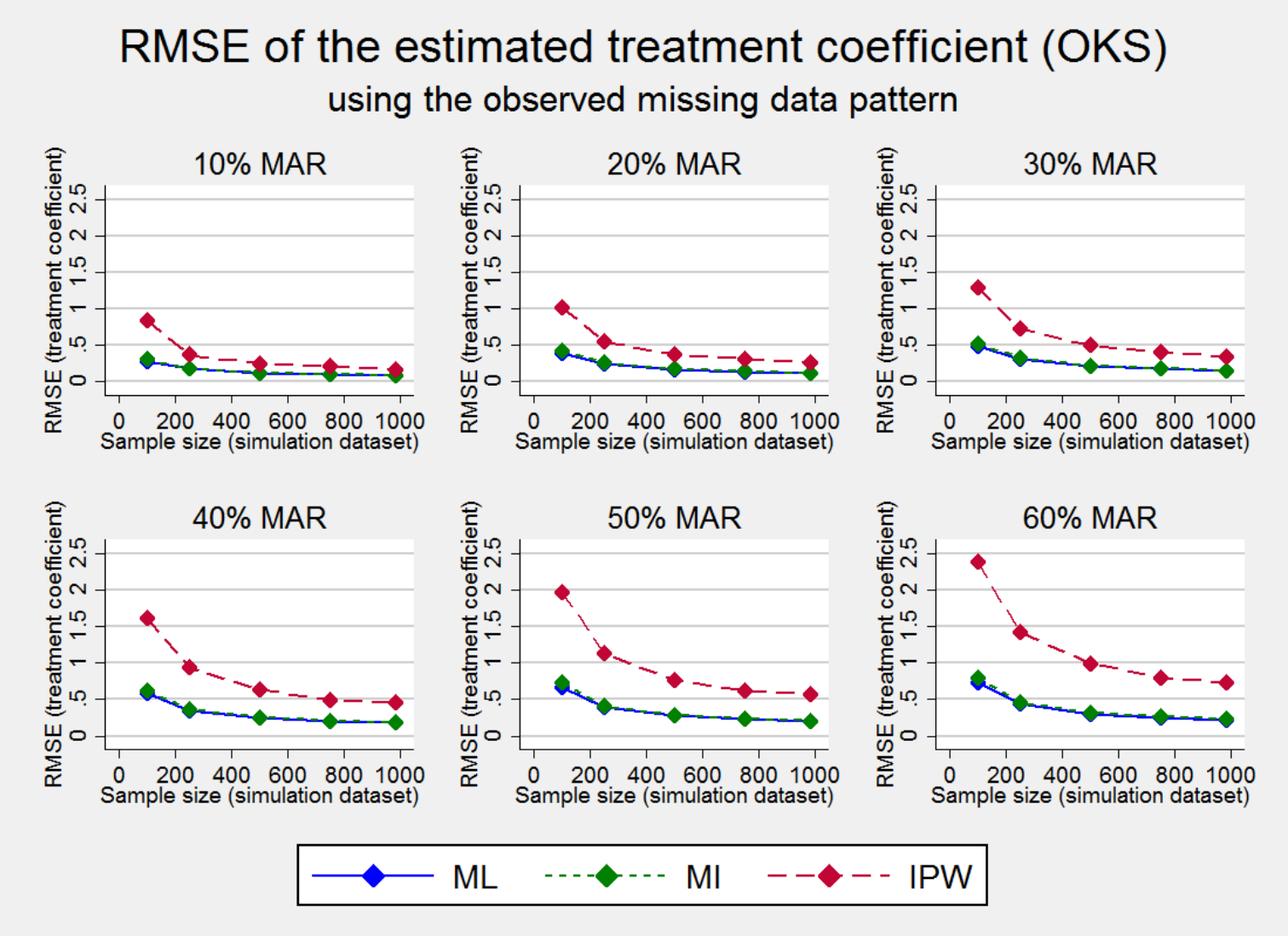


Figure 5-3: RMSE of the estimated treatment coefficient – simulations using the observed missing data pattern

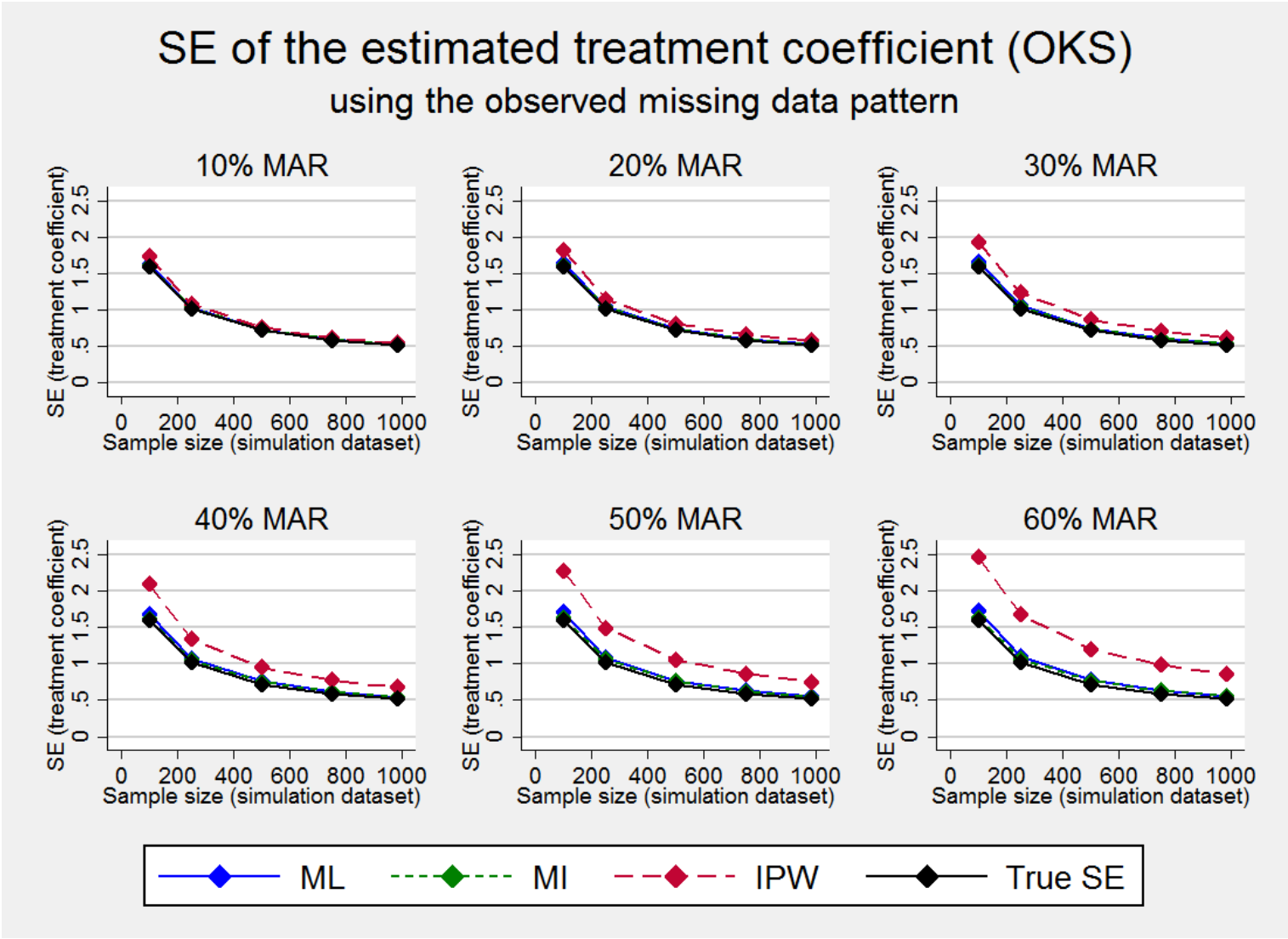


Figure 5-4: SE of the estimated treatment coefficient – simulations using the observed missing data pattern

5.6.3 Simulations using a five point treatment effect and observed MAR mechanism

Figure 5-5 shows the RMSE for the simulations introducing a five point treatment effect into the dataset; the other simulation parameters are the same as those described in section 5.6.2. This change in the underlying data does not appear to affect the performance of either of the analysis approaches, and results are very much in line with those presented above, both for the RMSE, as well as for the SE, as shown in Figure 5-6. Comparison of models by MAE show similar patterns (Appendix 14).

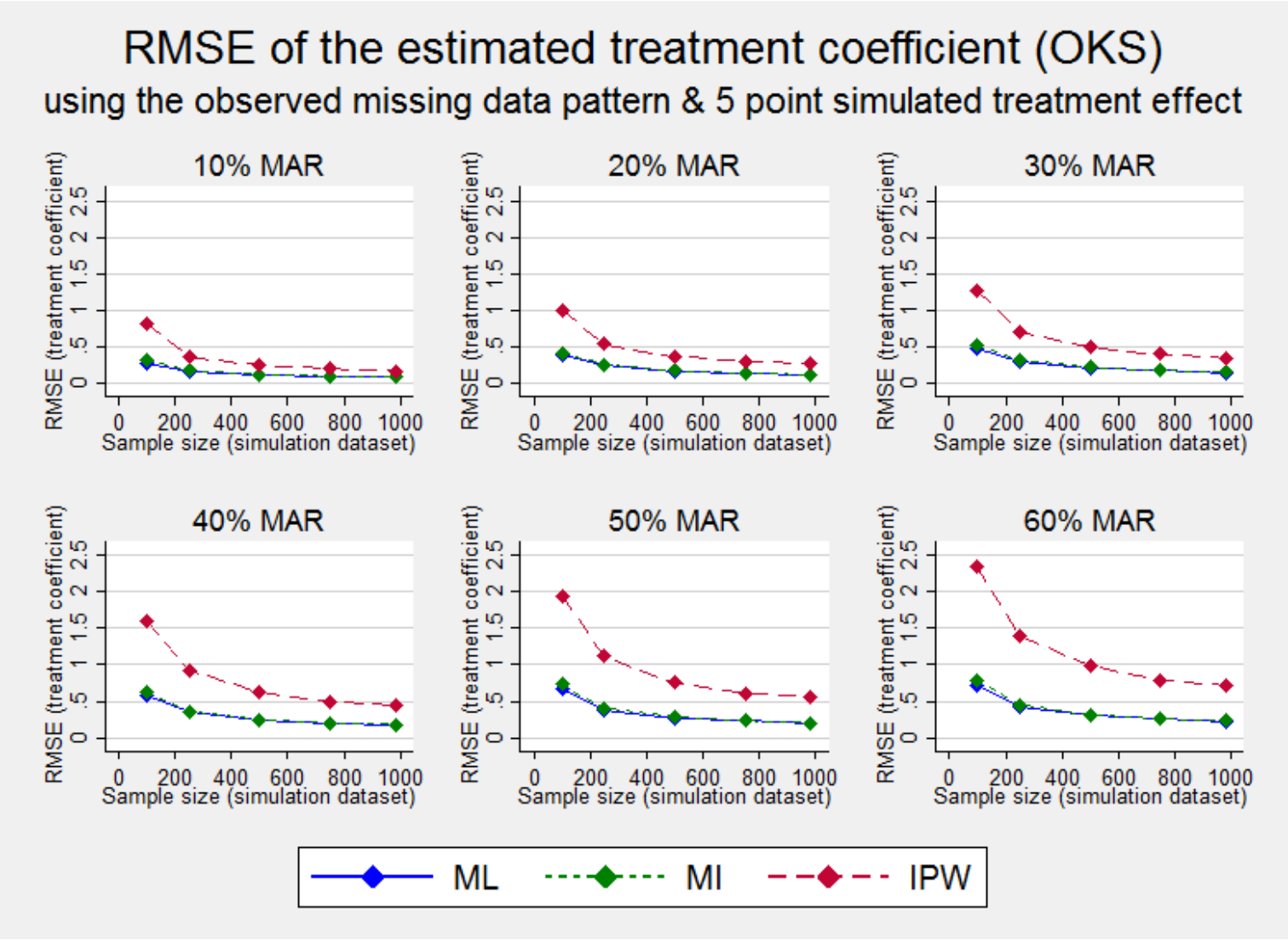


Figure 5-5: RMSE of the estimated treatment coefficient – simulations using the observed missing data pattern and a five point treatment effect

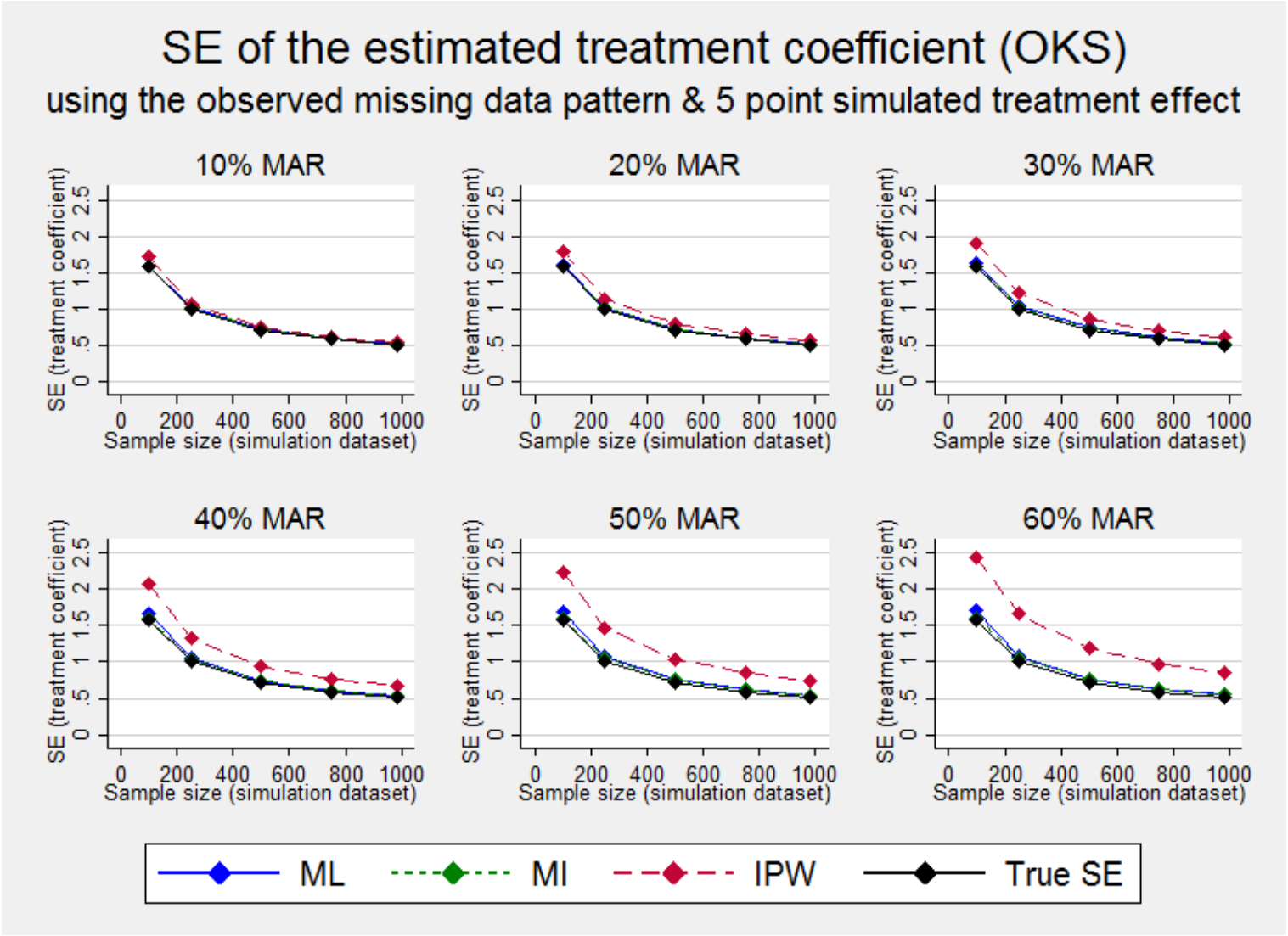


Figure 5-6: SE of the estimated treatment coefficient – simulations using the observed missing data pattern and a five point treatment effect

5.6.4 Simulations using the observed data and a stronger MAR mechanism

Here, the comparative performance of the different analysis approaches is considered when a stronger MAR mechanism is applied to the observed data, as described in section 4.4.3.2. Figure 5-7 shows the RMSEs observed in the treatment coefficients for this simulation scenario. Again, the patterns observed previously remain generally unchanged, although the RMSE for the IPW approach has decreased marginally, particularly for small samples and larger proportions of missing data. The MAE plots in Appendix 14 support these findings.

Figure 5-8 shows the SEs associated with each approach of handling missing outcome data. Again, the patterns are similar to those reported above.

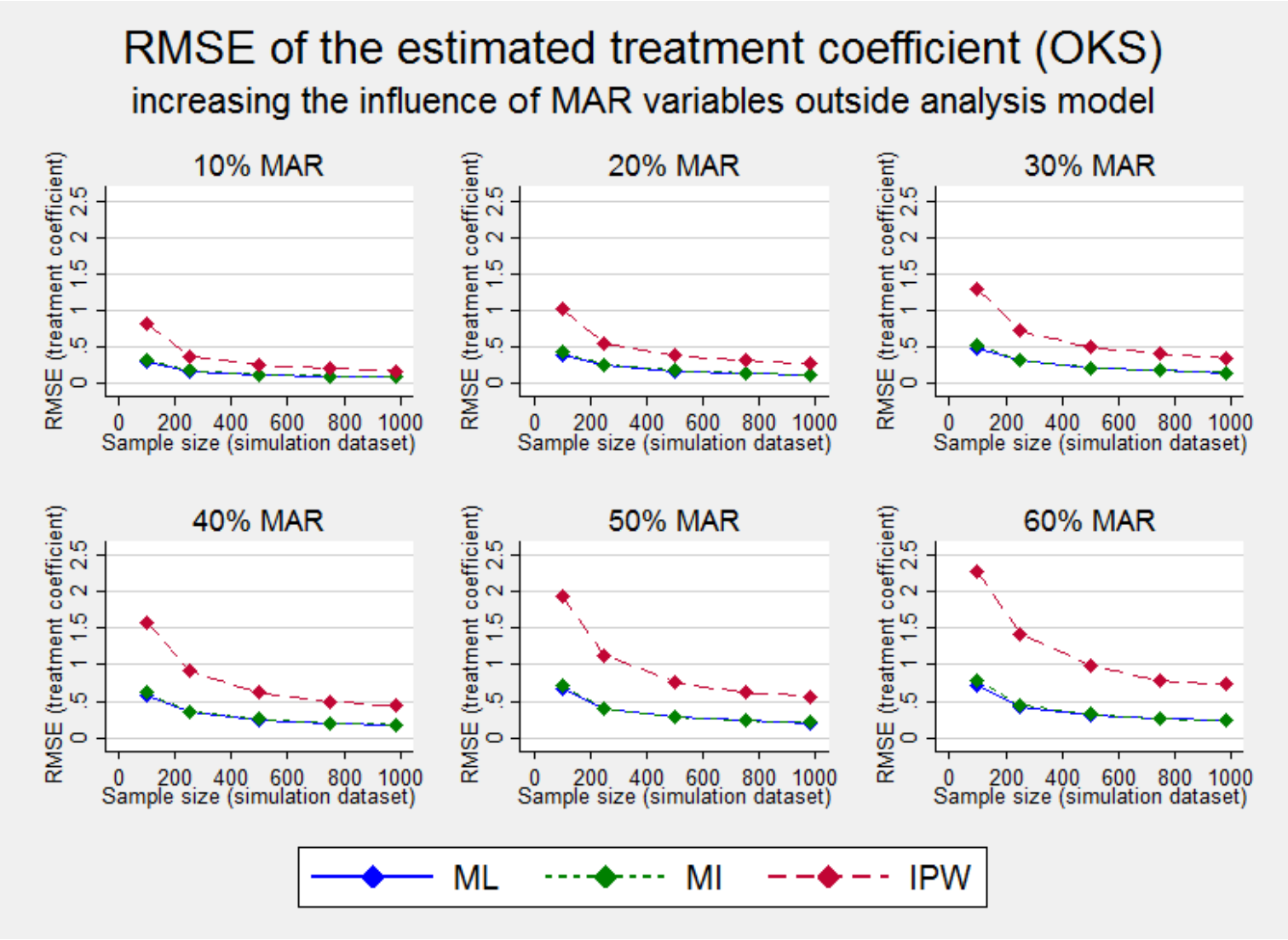


Figure 5-7: RMSE of the estimated treatment coefficient – simulations using the observed missing data pattern and a stronger MAR mechanism

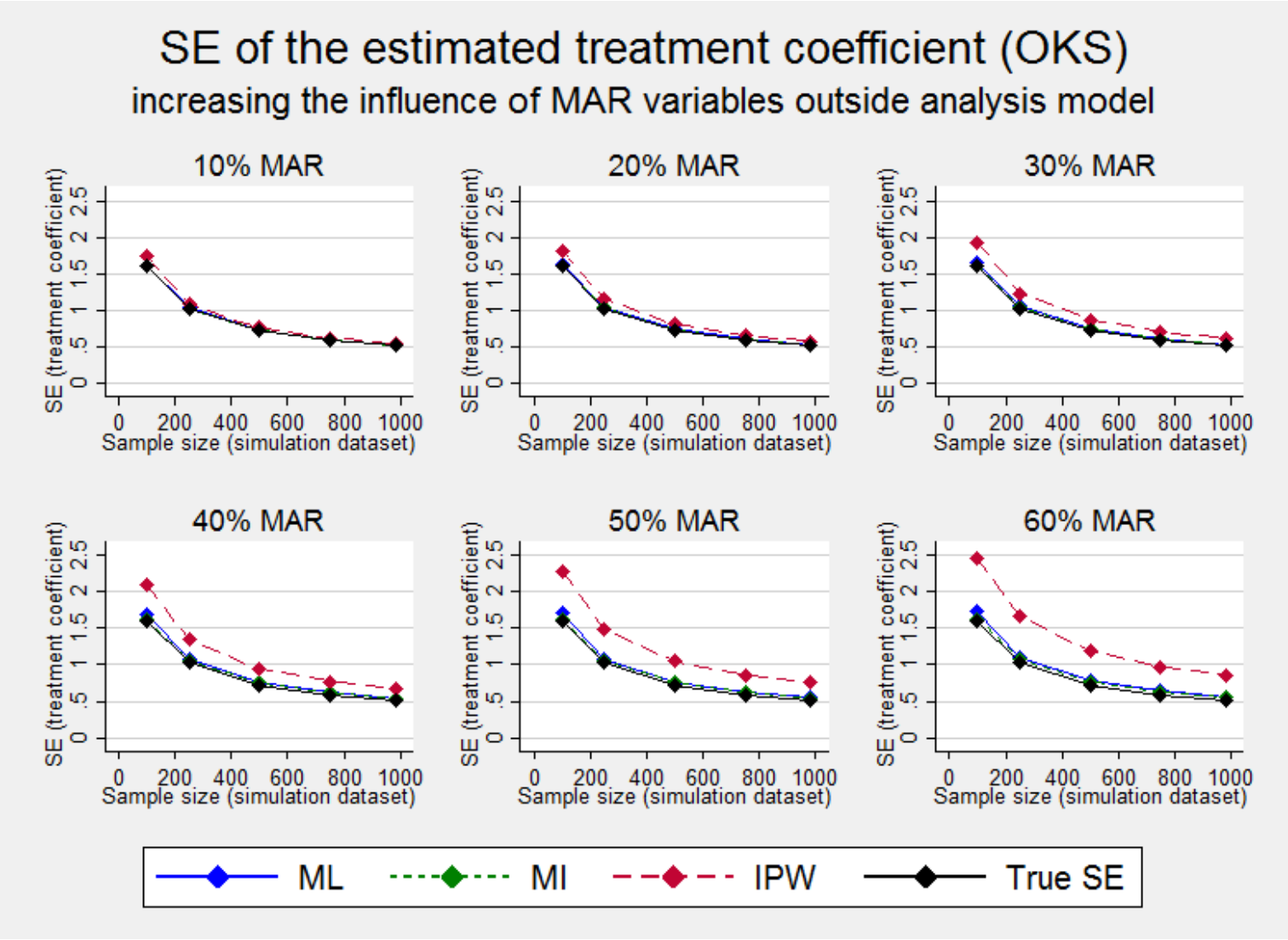


Figure 5-8: SE of the estimated treatment coefficient – simulations using the observed missing data pattern and a stronger MAR mechanism

5.6.5 Simulations using the observed data, observed MAR mechanism and an additional outcome variables in the MI & IPW models

Here, the comparative performance of the different analysis approaches is considered when additional information for other PROMs is added to the MI and IPW models, as described above. The additional auxiliary variables are the SF-12 and EQ-5D-3L.

Figure 5-9 shows the RMSE in the treatment coefficients. The bias in treatment coefficients, as represented by RMSE, obtained from the MI is reduced in this scenarios. The benefits of the MI model in this scenario are more pronounced when larger proportions of participants have missing outcome data. The bias introduced into the IPW model does not appear to have been decreased through the inclusion of this additional information. In fact, compared to the original simulations, the RMSE has increased very slightly across the simulation scenarios. Comparison of models by MAE show similar patterns (Appendix 14). Inclusion of the additional PROMs estimation has no impact on the SEs. Figure 5-10 shows the SEs for this simulation scenario, which are in line with those observed previously.

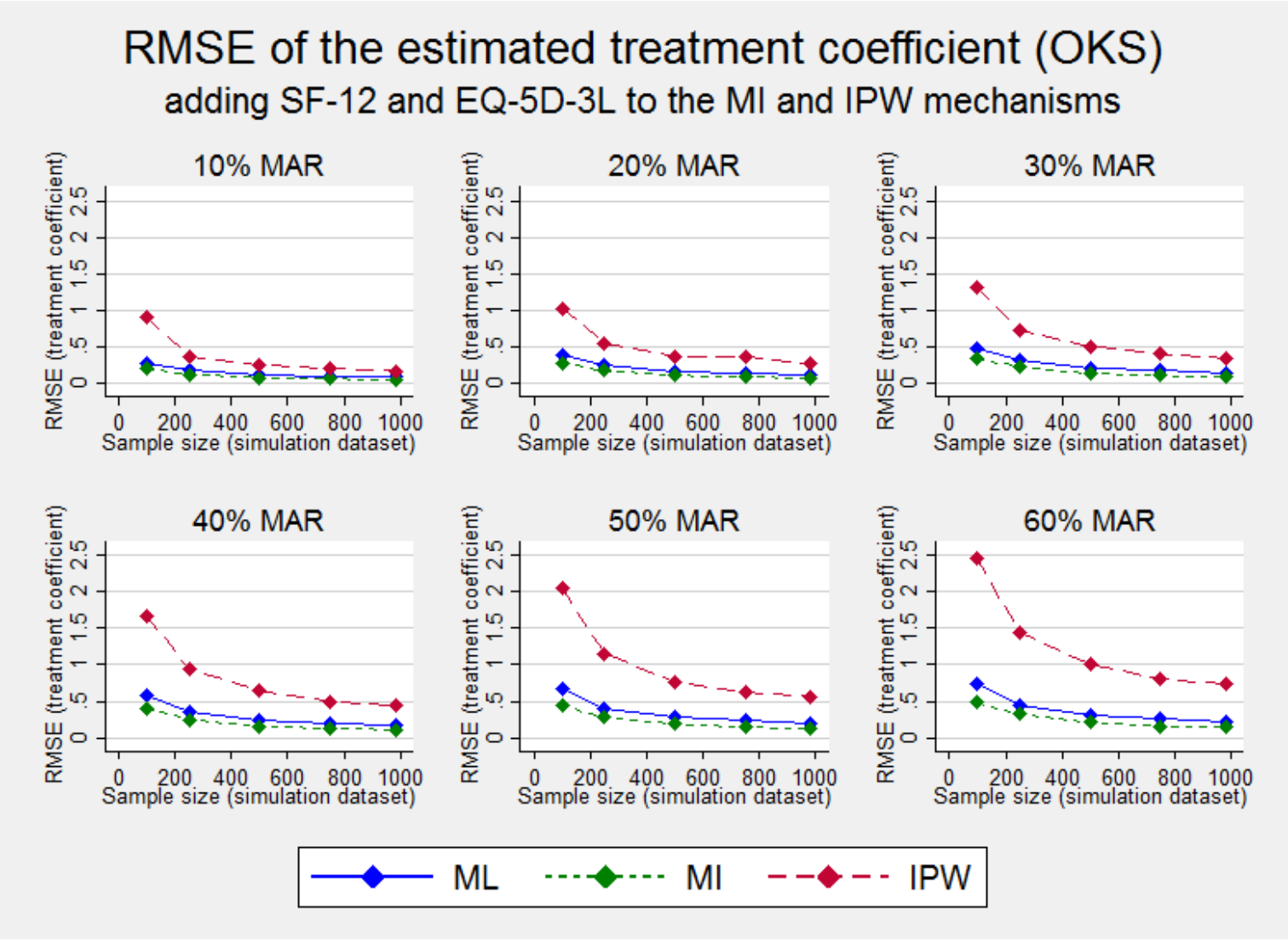


Figure 5-9: RMSE of the estimated treatment coefficient – simulations adding the SF-12 and EQ-5D-3L to the MI and IPW mechanisms

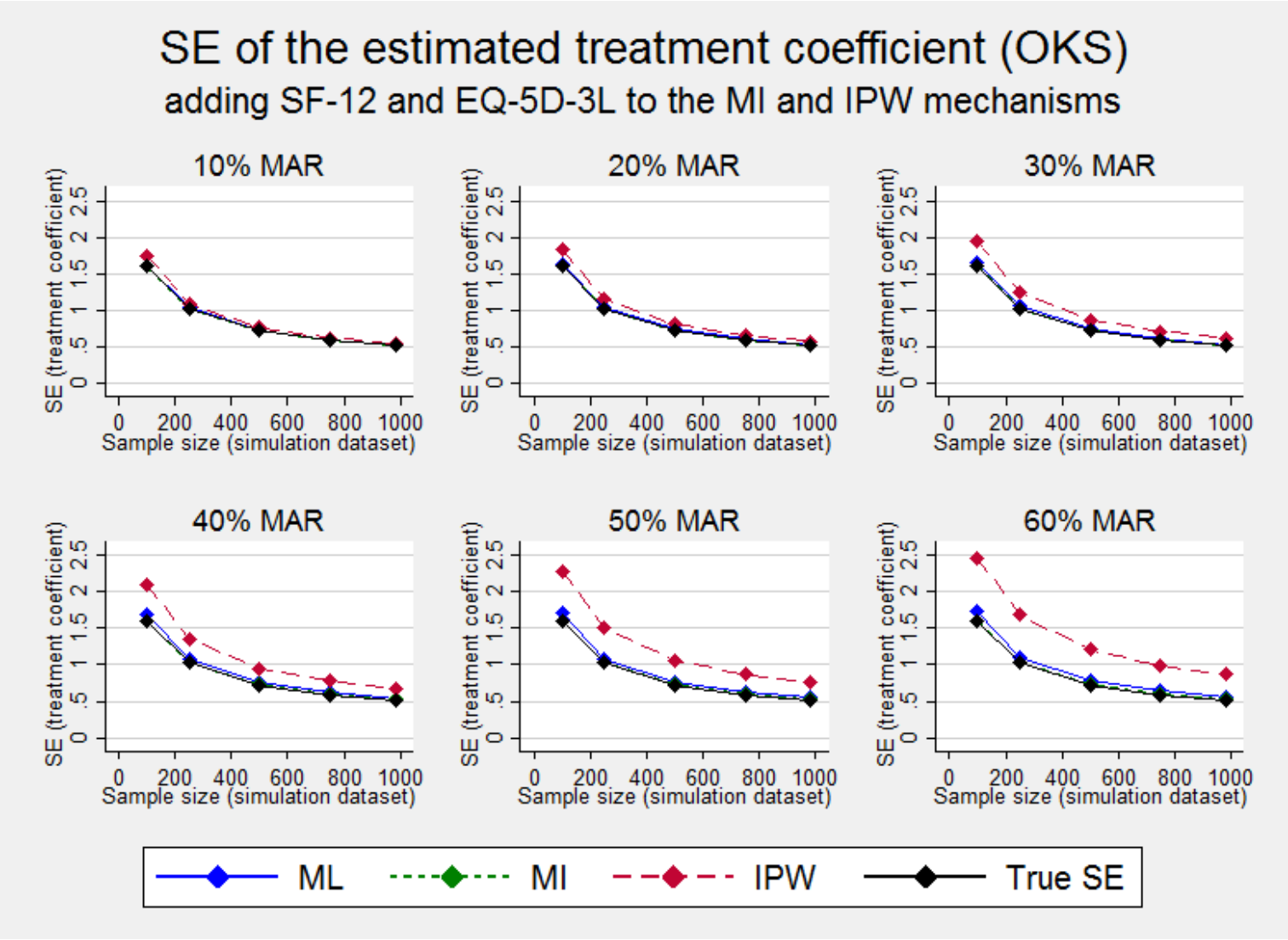


Figure 5-10: SE of the estimated treatment coefficient – simulations adding the SF-12 and EQ-5D-3L to the MI and IPW mechanisms

5.6.6 Simulations considering monotone missingness (i.e. drop-outs) only

As in the previous scenario, this analysis incorporates auxiliary SF-12 and EQ-5D-3L data. However, in this scenario, only monotone missingness, i.e. drop-outs from follow-up are considered, as described in section 4.4.3.2.

Figure 5-11 shows the RMSE in the treatment coefficients for this scenario. As in the previous scenario allowing for auxiliary variables, a small benefit of MI over ML is observed across all the different proportions of missing data; however, these differences are smaller than those observed in the previous simulation scenario. The bias introduced into the IPW model is increased slightly in comparison to the previous scenario. The graphs showing the MAE (Appendix 14) are consistent with these findings.

Figure 5-12 shows the SEs for this simulation scenario, which are in line with those observed previously.

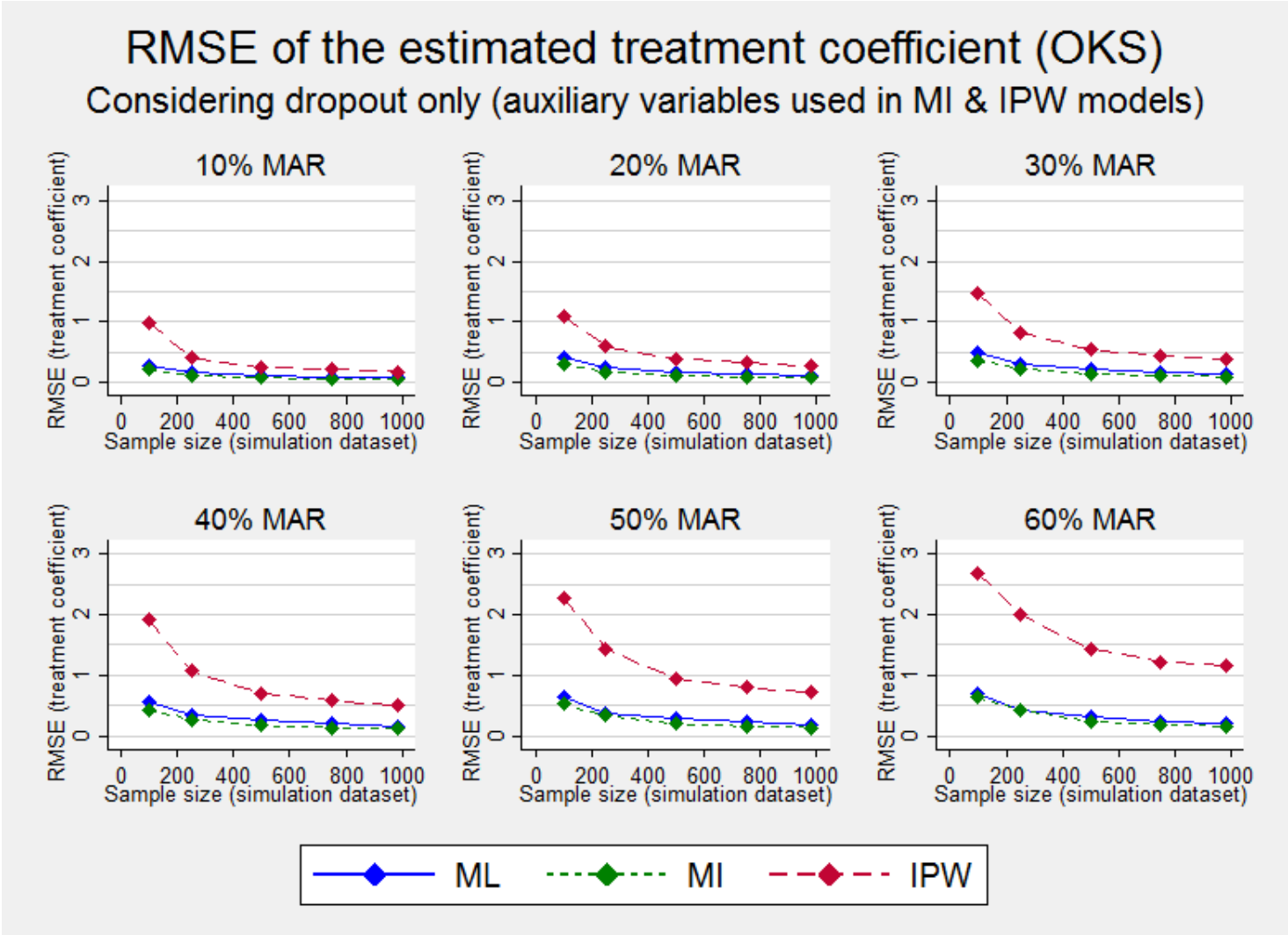


Figure 5-11: RMSE of the estimated treatment coefficient – simulations adding the SF-12 and EQ-5D-3L to the MI and IPW mechanisms while considering dropout only

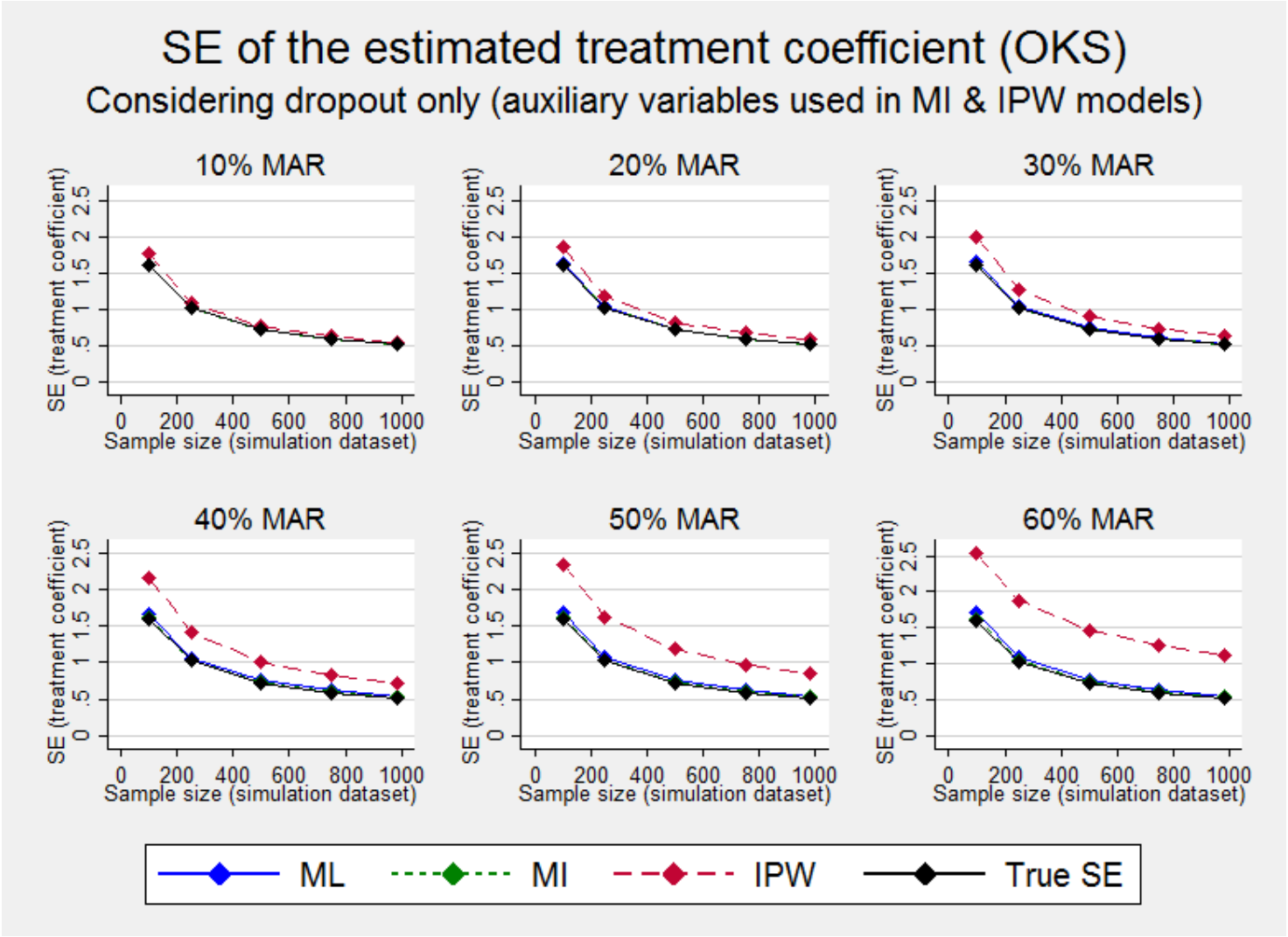


Figure 5-12: SE of the estimated treatment coefficient – simulations adding the SF-12 and EQ-5D-3L to the MI and IPW mechanisms while considering dropout only

5.7 Discussion

The aim of this chapter was to discuss and compare the performance of three different methods for analysing longitudinal PROMs outcomes with some missing data. ML, MI and IPW were investigated in this chapter, as they are widely considered to be appropriate methods to handle missing data in a longitudinal context, and are straightforward to implement using standard statistical software such as Stata.

The methodology underlying each approach was discussed in section 5.3, together with potential benefits and disadvantages of each, and the existing literature on the comparative performance of the methods was presented. The application of each approach was then demonstrated in a case study (section 5.4). This case study was also used as a 'motivating example', to show that the results obtained when applying the different approaches can differ, thus stipulating the need for further direct comparisons of the methods with a focus on RCT and PROMs data. The subsequent simulation study (section 5.5) then offered a direct comparison of the three different approaches under a number of different MAR scenarios, considering a range of sample sizes and proportions of participants with missing data.

The simulation study showed that results obtained for the ML and MI approach were very similar under MAR where the MI model took into account baseline data and data collected very early on in the trial only. This finding was not unexpected, especially as the most important predictors of the outcome were also included in the analysis model, and therefore taken into account in the ML estimation, and is in line with current literature[28, 67]. However, where other follow-up PROMs data was used to inform the MI model, a benefit of MI over ML could be observed. This benefit is likely to be due to the fact they

offer additional post-randomisation data which is correlated with the outcomes, and thus helps in predicting the missing values under a plausible MAR assumption. This is a very important finding which has the potential to contribute to decreasing the bias contained in results from RCT analyses. While it is true that the simulation study considered an ideal scenario, whereby full follow-up information was available for additional PROMs measures, which may not be realistic in many RCT settings, this approach is still likely to produce beneficial results for the MI analyses in practice. This is because most RCTs collect information on a number of different PROMs, which have different completion rates (see Chapter 3). More resources are likely to be devoted to ensuring high completion rates for PROMs used as a primary or key secondary outcome, for example by following patients up using phone calls when the relevant questionnaire was incomplete after the clinic visit or postal follow-up, and these data are therefore more likely to be complete compared to other PROMs collected. Also, it is possible that RCT participants may be more likely to complete questionnaires that are shorter, or considered more relevant to them, thus leaving only certain PROMs incomplete. Alternatively, PROMs may be collected at different time intervals, e.g. so that while no PROMs data may have been obtained at the time point of interest, other data for another PROMs may have been obtained at another follow-up visit at which only a subset of all PROMs was collected. In addition to PROMs, the follow-up data may include information on clinical assessments, re-admissions, further treatments or complications. In short, other follow-up information, both in the form of PROMs or clinical data is likely to be available and may yield valuable information on the missing data of the PROM data of primary interest. Even more importantly, including this post-randomisation information into the MI model also yields the possibility of making the MAR mechanism more plausible[151]. Especially when some of the missing data is related to

a change in health states, the use of additional PROMs outcome data may shift the underlying missing data mechanism from MNAR to MAR if the additional PROMs data is closely correlated to the missing PROMs data[59]. Therefore, where available, it is recommended that other relevant follow-up information should always be added to the MI model as this additional information can be informative of the missing PROMs data and makes the MAR assumption more realistic.

In all the scenarios, IPW performed notably worse than its comparators both in terms of bias introduced, as well as in variability around the treatment coefficient estimates, particularly for larger proportions of missing data. As demonstrated in the motivating example, this can partly be attributed to the fact that the IPW approach only utilises a subset of the data in the analysis, and that the weighting approach may not fully account for the missing data and the mechanism responsible for it. Therefore, one of the conclusions of this chapter is that IPW as utilised here is not recommended for the analysis of RCTs with some missing data in their longitudinal follow-up. This is in line with previous literature, which also does not recommend IPW, particularly when data for several variables are missing, and missing data are non-monotone[152].

Finally, the last two simulation scenarios compared and contrasted bias introduced by the same approaches for monotone vs. intermittent missingness. Bias introduced into the treatment coefficients increased slightly in the scenario considering monotone missingness. This again emphasises the importance of including all collected data into the analysis, and adds to the evidence against using IPW in these settings.

5.7.1 Feasibility of the three different approaches

The findings presented in this chapter have shown that valid results can be obtained for all ML approaches and the vast majority of MI models for the simulation scenarios considered. The IPW approach, however, failed to converge in approximately 20% of simulations for the combination of a sample size of 100 and 10% of missing data. Failures for combinations of other sample sizes and proportions of missing data were much less frequent, or not observed at all, as discussed above. The logistic regression model used to estimate the weights for the IPW approach relies on sufficient numbers of observations being available within all categories of the categorical explanatory variables for participants with and without complete follow-up data. Especially where missing follow-up data is simulated in only 10% of participants, the data distribution is more likely to be such that the logistic regression model does not converge. Where IPW is applied during the analysis of PROMs data, the logistic regression model should be constructed in a way that is appropriate for the underlying data. However, it was not possible to customise the model within this simulation study, as all factors needed to be kept constant to ensure comparability between the different simulations and simulation scenarios.

The complex MI models including additional PROMs follow-up data also failed in a low number of cases.

5.7.2 Novel aspects and limitations of this research

5.7.2.1 Contribution to the literature

This is the first simulation study directly comparing ML, MI and IPW in the context of RCTs focussing on missing PROMs data while utilising complex and realistic MAR mechanisms and missing data patterns. Therefore, the research in this chapter is an important contribution to the literature, providing important guidance to statisticians analysing PROMs data from RCTs. The simulation work covers a wide range of realistic missing data scenarios, and is therefore applicable to a wide range of medical research.

5.7.2.2 Limitations of this research

Although every effort was made to conduct this research as thoroughly and completely as possible, it is not without its limitations.

Firstly, similar to the research presented in Chapter 4, this simulation study considers a limited number of missing data scenarios. Maximum sample sizes were limited by the number of observations in the base case; however, as before, sample sizes ranging from 100 to around 1,000 participants are deemed representative of the vast majority of RCTs. The majority of the simulation studies consider the same missing data pattern, comprising a mixture of intermittent and monotone missingness. The simulation scenarios presented here could arguably have been extended to include other missing data patterns. Indeed, the longitudinal missing data patterns explored in Chapter 3 varied between trials, and to a smaller extend between PROMs regarding the number of participants that represent specific missing data patterns. However, there is consistency in that both monotone and intermittent missing data have been observed for all trials and PROMs observed. Therefore, the longitudinal missing data pattern used in the simulation study are realistic

missing data patterns, and are expected to be generalisable to a large proportion of RCTs. The simulation studies have also been extended to consider only a monotone missing data pattern.

As in Chapter 4, the simulation models are limited to the KAT dataset and also the OKS. While validation in other datasets, and for other PROMs would be beneficial, it is believed that the results from this research are generalisable for the following reasons. The OKS is used as a continuous score, in line with composite scores for many PROMs. Baseline variables have only limited predictive ability of the follow-up OKS data, and there is moderate correlation between the OKS and other PROMs collected at the follow-up assessments. This is in line with other PROMs, and it is therefore not expected that the use of other PROMs would have yielded different conclusions on the comparative performance of ML, MI and IPW.

In line with Chapter 4, MNAR mechanisms were not considered in this chapter. Again, this is because MNAR could be present in a number of ways, either limited to a specific treatment arm or relevant to all treatment groups, and resulting in either increased or decreased treatment effects. As such, findings from simulation studies considering MNAR mechanisms are unlikely to be generalisable because an approach that performs well under a specific MNAR mechanism may be an inappropriate choice in another MNAR scenario. This could lead to bias introduced through missing data being underestimated, as it is impossible to determine the MNAR mechanism based on the available data. For this reason, MNAR scenarios were omitted from this chapter. This is in line with other authors, who acknowledge that any MNAR analysis makes many additional assumptions, and should therefore be considered in the framework of sensitivity analysis[122-124, 153], which is further

discussed in Chapter 6. Including auxiliary follow-up information, in the form of PROMs as well as other data collected during the follow-up, should render MAR assumption more realistic and minimises bias[59, 153-156]. Still, the results from any analysis including missing data should be subject to appropriate sensitivity analysis, investigating potential departures from MAR, and this is further discussed in Chapter 6.

The analysis model used in this simulation study was a multilevel mixed-effects linear regression model. Alternative choices of longitudinal analysis models, such as marginal models with coefficients estimated by GEE were not considered within this work. The choice of model was guided by the pre-specified analysis model in the KAT study, the data of which formed the basis for this methodological work[93]. The interpretation of results for multilevel mixed-effects linear models and marginal models, also referred to as population average models differs. Marginal models produce the expected outcomes across all clusters (given the covariates), whereas multilevel mixed-effects linear regression models can also produce the effect of the covariates on individuals. In the context of this simulation work, the treatment effect in the RCT produced by the analysis model is of primary interest. This estimate can be derived from both approaches, and the literature demonstrates that both GEE and multilevel mixed-effects linear regression models give very similar results for continuous outcome variables[157]. Furthermore, research by Kang et al[146] (discussed in section 5.2.2) demonstrated that multilevel mixed-effects linear regression models produced better results in terms of accuracy and SEs than GEE in the presence of missing data, and multilevel mixed-effects linear regression models were therefore used in this simulation study.

The multilevel mixed-effects linear regression models used in this simulation work were fitted via maximum likelihood, which is the default in Stata's '*mixed*' command. An alternative estimation approach is the restricted maximum likelihood estimation (REML). Both are well-established approaches, with REML thought to provide unbiased estimates of the likelihood although differences between the approaches are described as negligible for large samples, while the ML approach is thought to be more appropriate for small samples[158-160]. Therefore, both approaches were considered appropriate for the purposes of the simulation study performed in this chapter. As such, the choice of the ML estimation approach over REML was influenced primarily by practical reasons: In Stata's '*mixed*' command, the ML estimation has more functionality than the REML estimation (see Stata 14 documentation: Multilevel mixed-effects reference manual). This additional functionality includes the use of robust variance-covariance matrices and weighted estimations, both of which were utilised in the IPW approach. It is for those reasons that ML estimators were utilised in this simulation work.

5.8 Conclusion

The research performed in this chapter presents a direct comparison of ML, MI and IPW when applied to RCT datasets with some missing PROMs follow-up data in a longitudinal context. The results of the simulation study show that the IPW model introduces more bias than the alternative approaches, and should therefore not be used for the analysis of similarly small RCT datasets, especially when some missing outcome data is observed for 30% of participants or more.

ML and MI perform similarly under MAR when no additional follow-up data is available. However, where auxiliary PROMs or other data have been more completely observed during the follow-up, or other post-randomisation data is available, then MI offers benefits in terms of bias reduction and should be favoured over non-imputation based ML approaches. As both approaches assume data being MAR, additional sensitivity analyses considering MNAR scenarios remain imperative to supplement the primary analysis.

Chapter 6 : On the importance of sensitivity analysis to investigate the robustness of randomised controlled trials results with regards to missing outcome data

6.1 Introduction

The work in this thesis has compared different analysis approaches for RCTs with missing PROMs outcome data where the planned analysis takes into account either a single follow-up time point (Chapter 4) or longitudinal follow-up (Chapter 5). The comparative performance of the different approaches was reported in terms of bias and precision for a number of scenarios considering different sample sizes and proportions of missing data, and relevant guidance and advice were derived. However, the results of any of these models rely on a number of assumptions, including firstly that the missing data pattern truly is missing at random (MAR) for many analyses, and secondly that the analysis model or MI model can capture all the relevant variables in the MAR mechanism.

Within a simulation study, these factors can be set by the researcher. The missing data in Chapters 4 and 5 were simulated to be MAR, and therefore the MAR assumption was appropriate. In reality, however, the true underlying missing data mechanism cannot be known, as the assumptions made about the underlying missing data mechanism are untestable given the available data. Researchers can make an educated guess about the underlying missing data mechanism based on the patterns of missingness observed, other available data, as well as reasons clarifying why data are unavailable. However, they can never be certain that their assumptions are correct[17, 18, 59, 122, 161] and that all the relevant variables are taken into account when adjusting for missing data.

Of course, the use of inappropriate assumption can lead to biased results. Therefore, results based on data containing some missing outcomes should always be investigated for robustness with regards to varying the assumptions made about the underlying missing data mechanism, including the possibility that data is missing not at random (MNAR). This process is referred to as sensitivity analysis[15, 18, 162-164]. The addition of sensitivity analysis to the primary analysis of any RCT is crucial to reassure clinicians, researchers and regulatory bodies that the trial results are robust and conclusions are unlikely to vary due to a change in the way missing data were handled.

An overview of the literature on guidance provided on appropriate sensitivity analysis with regards to missing data, defined as analyses considering MNAR mechanisms, is presented in section 6.3. Two approaches to applying MNAR sensitivity analysis that are straightforward to implement using standard statistical software are presented in section 6.4. A case study, described in section 6.5, is used to demonstrate the application of two intuitive and easily implementable sensitivity analyses. The results from these two sensitivity analyses, and their conclusions about the robustness of the underlying trial results will also be discussed. Findings from this chapter are discussed in section 6.6, and conclusions are provided in section 6.6.

6.2 Objectives for this research

The aims of this chapter are to:

- Provide an overview of the recommendations for sensitivity analysis in the current literature with regards to investigating the possibility of data being MNAR
- Present two intuitive and easily implementable approaches to performing MNAR sensitivity analysis that already exist in the literature, and discuss their implementation and implications on the trial results using a case study
- Emphasise the importance of sensitivity analysis when carrying out any statistical analysis in the presence of missing data

6.3 Advice on sensitivity analysis in the current literature

The importance of sensitivity analysis has been recognised by methodologists and regulators, with Li et al stating that "Examining sensitivity to the assumptions about the missing data mechanism (i.e., the sensitivity analysis) should be a mandatory component of the study protocol, analysis, and reporting. This standard applies to all study designs for any type of research question."[18], with others, including the authors of the CONSORT (Consolidated standards of reporting trials) statement, echoing this sentiment[15, 17, 80, 162, 163, 165].

Regulatory bodies, such as the European Medicines Agency (EMA) and the U. S. Food and Drug Administration (FDA), also incorporate the use of sensitivity analysis into their guidance. The EMA states that "sensitivity analyses should show how different assumptions influence results obtained" and "because the performance of any analysis presented (in terms of bias and precision) cannot be fully elucidated, presentation of trial results without adequate investigation of the assumptions made for handling missing data cannot be considered comprehensive"[66]. The FDA adds that "sensitivity analysis should be part of the primary reporting of findings form clinical trials, and that examining sensitivity to the assumptions about the missing data mechanism should be a mandatory component of reporting"[40]. The FDA clarifies that "sensitivity analysis [should] assess the impact of missing data on estimates of treatment differences", but also recommends that more research is still needed on methods of sensitivity analysis and decision making arising from sensitivity analysis[40].

The FDA and EMA provide guidance on the scope and documentation of sensitivity analysis, proposing that "each sensitivity analysis should be designed to assess the effect on the results of the particular assumption made to account for the missing data. The sensitivity

analysis should be planned and described in the protocol and / or in the statistical analysis plan and any changes must be documented and justified in the study report"[66]. Specifically, the EMA proposes that MNAR sensitivity analysis should, where possible, take into account reasons why data are missing, and adds that a clear explanation should be provided for the values imputed. The EMA also states that "MI methods including a pattern mixture approach may be appropriate"[66]. Further, the EMA suggests considering also a worst case scenario: "assigning the best possible outcome to missing values in the control group and the worst possible outcome to those of the experimental group. If this extreme analysis is still favourable then it can be confidently concluded that the results are robust to the handling of missing data".

The advice in these guidelines is very broad and lacks practical advice on implementation[131], which may contribute to the low proportion of existing studies implementing appropriate sensitivity analysis that considers MNAR scenarios, as found in Chapter 2[39]. The implementation of sensitivity analysis is not discussed in other guidance documents, such as the Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement[166, 167].

Methodologists and researchers in the field of statistics and health economics have addressed some of the uncertainty around the implementation of sensitivity analysis. Apart from Bayesian approaches[168-171], which are not considered within this thesis, two frameworks for sensitivity analysis, namely pattern-mixture models and selection models[59, 66, 67, 131, 164, 169, 172-174] have received much attention in the recent literature, and were referred to in the EMA and FDA guidance documents discussed above.

Generally, pattern-mixture models facilitate sensitivity analysis by allowing for differences in the distribution of data between the distinct groups, or pattern, within the dataset. As such, while outcomes for participants with and without complete data that are assumed to be similar under MAR (subject to other variables), these outcomes are allowed to differ within the pattern-mixture model. Different MNAR assumptions for participants with missing data by treatment arm, or different patterns of missing data, can also be implemented in the pattern-mixture model. The difference, or δ, between observed and unobserved data is specified. The choice of δ should be specific to the study being analysed, and could consist of a "jump to reference" whereby those with missing data in the active intervention are assumed to follow the distribution of those with observed data in the control arm[26, 161]. In other examples, δ may be chosen as a clinically plausible value that is added to or deducted from imputed values estimated under MAR. For example, this value might differ based on time since randomisation in longitudinal studies[67, 107], where appropriate. The literature shows a variety of implementations of pattern-mixture models, including cases where δ is applied to specific imputations after MI, or during the MI process for longitudinal data, so that imputations for later time points can take into account corrections at earlier time points based on MNAR assumptions[164].

Selection models also acknowledge that participants with missing data differ from those with observed data. However, here the MNAR assumptions are applied by using a joint model for both the complete data and the missingness mechanism. Again, a parameter δ is defined, which in the context of selection models represents the departure from the MAR assumption, thus representing a MNAR setting. Selection models can be implemented by applying a weighting approach to the imputed values after MI under MAR,

giving imputations that are more likely to stem from the assumed MNAR mechanism a higher relative weight in the chosen analysis model. Based on δ, the weight is calculated for each of the imputed datasets. Subsequently, these weights are used in the calculation of the estimates and variances of the MI approach[67, 174, 175].

The value of δ in both the pattern-mixture and selection model should be based on clinical or expert opinion[169, 174]. Alternatively, researchers may choose to consider a range of possible δ values in order to assess how plausible the assumptions that can cause a change in the study conclusions are i.e. at what point results that are statistically significant under MAR move to being statistically insignificant, and vice versa. This is also referred to as tipping point analysis[131, 164, 176]. This can be used to interpret how robust the results are to varying assumptions about the missing data. If only a very large departure from MAR results in a change in conclusions, this may assure researchers that the results are robust to plausible departures from MAR[177]. Carpenter at al[26] stress that the assumptions need to be accessible and relevant.

Both pattern-mixture and selection models can be used to generate single imputations, or be applied after MI[178]. There are disadvantages of generating single rather than multiple imputations, particularly their inability to adequately account for the uncertainty around the missing values[164], which can affect sensitivity analysis investigating the robustness of trial conclusions[179]. However, the implementation of both approaches, especially in an MI setting, can be computationally demanding and there is a distinct lack of software to apply appropriate MNAR sensitivity analysis, especially after MI. It is thought that this may deter some researchers from implementing such sensitivity analysis[175, 180]. Since appropriate considerations of sensitivity form a crucial component of comprehensive RCT reporting,

this chapter focusses on two approaches to performing MNAR sensitivity analysis that already exist in the literature. These approaches can be easily implemented after any primary analysis that makes a MAR assumption by either including available cases and adjusting for relevant covariates, or performing MI and are described in the following section.

The search strategy implemented to identify the relevant publications discussed above is presented in Appendix 15.

6.4 Exploration of approaches to conduct MNAR sensitivity analysis

In this section, two approaches to conducting sensitivity analysis with a focus on MNAR assumptions are explored.

6.4.1 Stata's *rctmiss* command

The *rctmiss* command has been written by Ian White at the MRC Biostatistics Unit in Cambridge, and can be obtained by typing 'net from http://www.mrc-bsu.cam.ac.uk/IW_Stata/' into Stata. It is designed to allow researchers to investigate the effect of MNAR assumptions during the analysis of RCTs. Specifically, researchers can assess the effect on the treatment coefficient within their analysis model under a range of MNAR scenarios specifying either a selection model or pattern-mixture model approach; here, the focus is on the latter. The researcher can specify δ, which represents the MNAR assumption to be investigated, i.e. it represents the difference in missing outcomes under MAR and MNAR. For example, under the pattern-mixture approach a δ of -5 would investigate the MNAR scenario whereby missing outcomes are assumed to be five points worse on a PROMs scale than they are assumed to be under MAR. The value of δ depends on the outcome measure used, the range of the score, as well as what is considered a clinically relevant or plausible difference. This MNAR assumption can be applied to either one or both of the trial arms, and researchers can investigate either a single departure from MAR, or a range of different MNAR scenarios.

The *rctmiss* command works for both continuous and binary outcomes and can also handle missing baseline data; however, in this chapter, the focus is on continuous outcomes without missing baseline data.

The command generates the treatment coefficients under the specified MNAR scenario, i.e. the estimates obtained under MAR are modified in line with the MNAR assumptions expressed in δ. Results are provided either in table form, listing the trial arm(s) to which the MNAR assumption is applied, the treatment coefficient estimated under MNAR and the corresponding standard error. When a range of MNAR assumptions are investigated, a graphical representation of the MNAR sensitivity analysis results shows the values of delta on the x-axis and the MNAR treatment coefficient with a 95% CI on the y-axis.

By being able to show the results from a range of MNAR analyses, the graph as well as the listing of results can also be used for a tipping point analysis, as discussed above.

Within its documentation, the fact that *rctmiss* only works for two-arm RCTs is listed as a limitation, meaning that it cannot be used for sensitivity analysis of multi-arm comparisons in multi-arm RCTs. Also, the command documentation gives examples for implementation where the analysis model is based on a linear or logistic regression model. Currently, for example, the *rctmiss* command cannot be used for the multilevel mixed-effects models implemented by the 'mixed' command in Stata. Therefore, the *rctmiss* command may not be applicable for sensitivity analysis in all RCT contexts. This highlights the need for alternative approaches to sensitivity analysis that offer this additional flexibility to allow for additional trial arms or analysis models to be included in the sensitivity analysis, and is further discussed in section 6.4.2.

6.4.2 Manually changing the MI imputations according to MNAR

MI is considered a good basis for sensitivity analysis. Particularly where the main analysis is based on MI, MNAR scenarios can be investigated by manually changing the imputed values for participants with missing data[107, 181]. This approach can be considered an application of the pattern-mixture model, as the expected outcomes for certain groups of participants are modified.

This sensitivity analysis could be applied to a trial analysis with any number of trial arms or follow-up time points. The imputed values can be changed in line with the required sensitivity analysis scenario, i.e. imputed outcomes for participants in one trial arm could be decreased by a specified δ (similar to the approach using the *rctmiss* command), while imputed values in the other trial arm are kept unchanged, assuming that these participants follow the MAR mechanism. Alternatively, changes could be applied to both arms equally, or researchers may take into account the reasons for missing data, as proposed by White et al[107]. For example, participants who withdrew from the trials due to deteriorating health conditions or adverse reactions to the intervention could be assigned a larger change from MAR, while those who are lost to follow-up due to relocation and a lack of time to attend follow-up assessments may be assigned values that are closer to MAR. Generally, changes can be applied by either adding or subtracting a constant value to / from the imputed values, or applying changes based on a statistical distribution, i.e. a normal distribution with a specified mean and variance. Whether δ is positive or negative depends on the direction of the score (i.e. do higher scores indicate positive changes such as better functioning, or negative changes, such as higher levels of pain) and on the MNAR made (i.e. are participants with missing data assumed to have better or worse outcomes than those with available data). For longitudinal data, changes to the imputed values can also depend

on time since randomisation. A different δ can be applied to subgroups depending on randomisation allocations and/or reasons for loss to follow-up. Compared to *rctmiss*, this approach also gives researchers the option to implement relevant limits to the changes, such as ensuring that the imputed data, e.g. PROMs scores, do not fall outside their relevant range after applying the MNAR changes; i.e. the OKS can only range from 0 to 48. This option is not possible within *rctmiss*.

6.5 Application and interpretation of both sensitivity analysis approaches

In this section, the above described MNAR sensitivity analysis approaches are applied to a case study. Results for a complete case analysis, a MI model with imputation at the composite score level, as well as different MNAR scenarios are provided.

6.5.1 Datasets used in this case study

Within this case study, two RCT datasets of sample sizes 200 and 1000 are considered, with 20% missing outcome data at a single follow-up time point. The datasets are generated in line with the method described for the simulation work in Chapter 4, whereby the required number of observations are sampled from the KAT participants[92, 93] with complete data for relevant baseline variables and complete OKS follow-up data at five years. Missing data are imposed onto these datasets using the algorithm described in Chapter 4, generating approximately 20% of MAR data based on the missing data patterns observed in the trial.

6.5.2 Results of the CCA and MI assuming data are MAR

Table 6-1 provides information on the datasets used in this case study. They are random subsamples of the KAT participants randomised to patellar resurfacing vs. no patellar resurfacing; this is the reason why the treatment coefficients show different trends. Roughly equal numbers are allocated to each arm; approximately 20% of the participants in both samples and either trial arms have missing outcome data.

Data are analysed using a regression model with the OKS at five years as the outcome variable adjusted for treatment allocation, baseline OKS, age and gender. In addition to the covariates, the MI model also accounts for ASA physical status, height and size of randomising centre.

Table 6-1: Information about the CCA datasets

	Sample size: 200			Sample size: 1000		
	No patellar resurfacing	Patellar resurfacing	Total	No patellar resurfacing	Patellar resurfacing	Total
Participants randomised	94 (47.00%)	106 (53.00%)	200 (100%)	482 (48.20%)	518 (51.80%)	1000 (100%)
Participants with missing OKS data	19 (20.21%)	23 (21.70%)	42 (21.00%)	99 (20.54%)	97 (18.73%)	196 (19.60%)
	Regression coefficient of patellar resurfacing vs. no patellar resurfacing (95% CI):			Regression coefficient of patellar resurfacing vs. no patellar resurfacing (95% CI):		
CCA	-1.164 (-4.495, 2.166)			0.878 (-0.485, 2.241)		
MI (MAR)	-0.924 (-4.247, 2.399)			0.858 (-0.489, 2.205)		

The CCA and MI treatment estimates vary, more so in the smaller sample, which reflects that the MI approach may be able to account for additional variables relevant to the MAR mechanism.

In an RCT context, as opposed to a simulation study, the MAR assumption cannot be verified using the available data. Therefore, it is important to perform MNAR sensitivity analyses investigating how changes to the assumed missing data mechanism would affect the results, and the two approaches discussed in section 6.4 are applied to the two datasets.

6.5.3 Considering MNAR scenarios using the *rctmiss* command

For the sensitivity analysis using the *rctmiss* command, the pattern-mixture approach is used. Initially, MNAR scenarios for deviations from the MAR assumption of up to 10 OKS

points are considered for the sample with 200 observations. The results of the CCA and MI analysis for this dataset are presented in Figure 6-1. The y-axis shows the magnitude of the treatment coefficient, while the x-axis indicates the departures from MAR. The treatment coefficient, which is the difference in the OKS outcome scores between the two treatment groups, and corresponding 95% confidence interval from the CCA can be seen at the centre of the graph, where zero is indicated on the x-axis, i.e. no MNAR assumptions have been applied to the participants with missing data. The differently coloured lines indicate the effect on the treatment coefficient when the MNAR assumption is applied to either only the patellar resurfacing group (blue line – abbreviated "PR only"), the no patellar resurfacing group only (green line – abbreviated "No PR only") or both arms (red line).

The red line shows the effect on the treatment coefficient when the same departure from MAR is applied to both arms. Here, the change in the treatment coefficient is very small, indicated by the almost horizontal red line that joins the treatment coefficients for the different MNAR scenarios. The treatment coefficient adjusted for covariates in the analysis model does not change much because both trial arms have approximately the same proportion of missing data, meaning they are equally affected by departures from MAR in this MNAR scenario. For trials with different proportions of missing by treatment arm, the effect on the treatment coefficient is inflated[151].

The blue line shows that the treatment coefficient increases if only participants with missing data in the patellar resurfacing arm are assumed to have higher scores than under MAR, and also that the treatment coefficient decreases when these participants are assumed to have worse outcomes than under MAR. Vice versa, the green lines show that the treatment coefficient decreases if only participants with missing data in the no patellar resurfacing group are assumed to have better outcomes than under the MAR assumption,

and conversely the treatment coefficient increases if these participants are assumed to have lower OKS scores than under MAR.

The results from the CCA in Table 6-1 indicate that there is no evidence of a treatment difference between the trial arms, with the 95% CI crossing zero. Similarly, for the MNAR scenarios presented in Figure 6-1, the CIs still cross zero, perhaps with the exception of the scenario considering the imputation of missing outcome scores of 10 points lower than under MAR in the patellar resurfacing arm only, which may be borderline significant. These results should be discussed with the clinicians involved in the trial. If the clinical opinion suggested suggest that a 10 point departure from MAR in only one trial arm is very large, and therefore an unlikely scenario, it could be concluded that the analysis for this trial is robust to plausible departures from MAR. As such, the trial results are robust to the variation in the assumptions about the underlying missing data mechanism tested here. Therefore, the trial results are unlikely to be affected by the way missing data were handled.

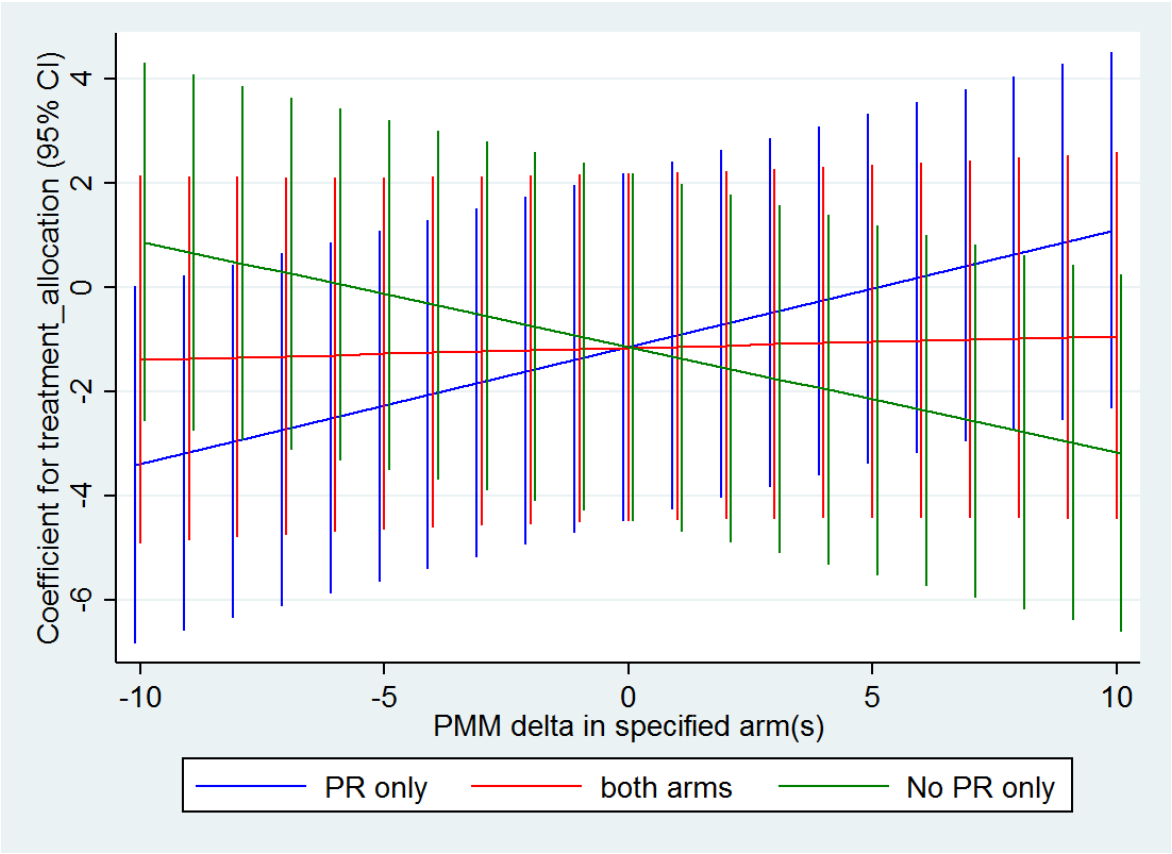


Figure 6-1: Results of the MNAR sensitivity analysis using *rctmiss*, sample size = 200

The results of the sensitivity analysis for the data set with 1000 observations are shown in Figure 6-2. Again, the proportions of missing data in both treatment arms are very similar (Table 6-1), and therefore the effect of applying the same MNAR assumptions to both arms would not have much of an effect on the treatment coefficient. This MNAR scenario is therefore excluded from the graph.

Again, there was no evidence to suggest that the treatment coefficients from the CCA and under MI are significantly different from zero under the MAR assumption. Figure 6-2, however, shows that this treatment coefficient changes to being greater than zero (here shown by the 95% CI no longer crossing zero) under certain assumptions about the missing data. Specifically, a change in conclusions is observed when participants with missing outcome data in the no patellar resurfacing group have OKS outcome scores at least three points worse than assumed under MAR (green line). Vice versa, conclusions also change

when those with missing data in the patellar resurfacing arm are assumed to have OKS outcome scores at least three points better than assumed under MAR (blue line). Both scenarios assume that missing data in the other trial arm assumed to be MAR.

A three point difference in the OKS falls below what is defined as the "smallest possible detectable change for [this] instrument, thus indicating it would fall within measurement error"[12]. For this reason, this departure from MAR may be much more plausible than the difference of 10 points or more discussed above, and means that the results may not be robust to varying the assumptions about the underlying missing data mechanism. From this it can be concluded that the analysis results from this dataset are a lot less robust to missing data, and researchers, clinicians and regulators should be more cautious in accepting the findings without additional research.

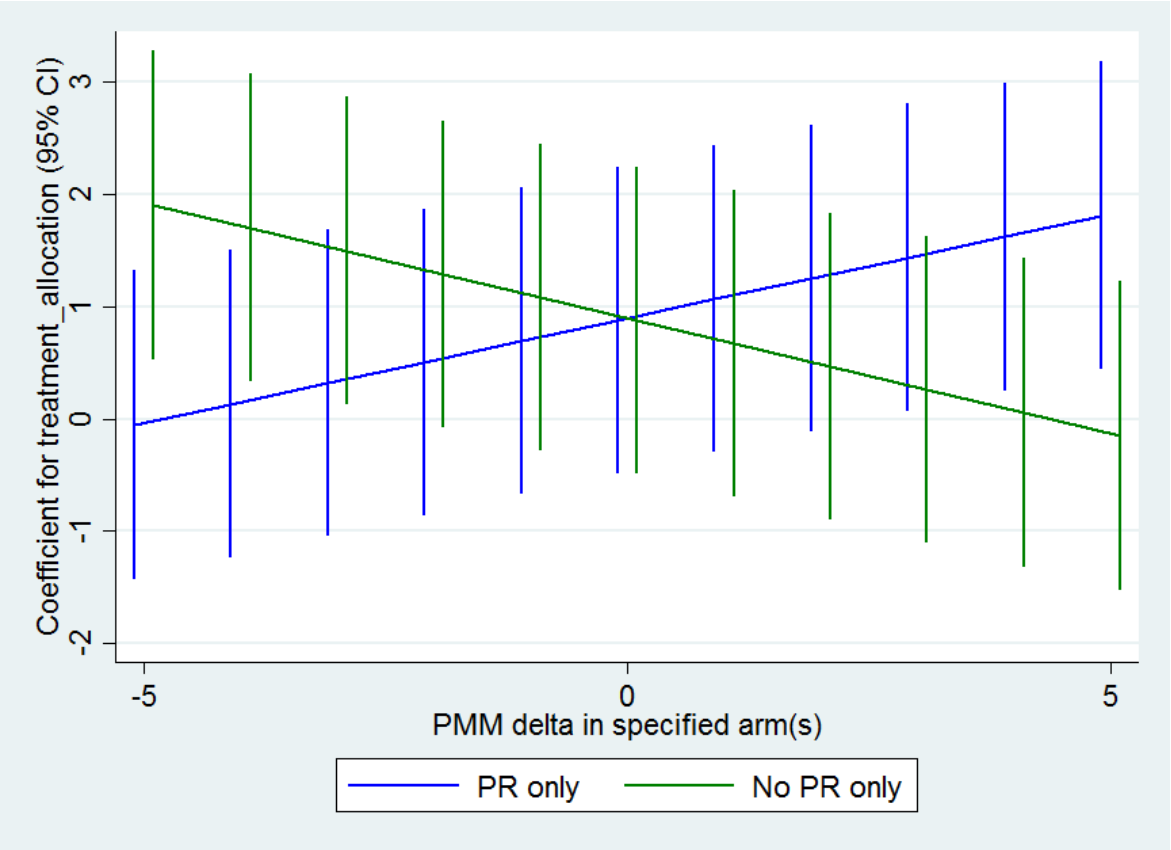


Figure 6-2: Results of the MNAR sensitivity analysis using *rctmiss*, sample size = 1000

6.5.4 Considering MNAR scenarios by manipulating the MI imputations

Table 6-2 contains the results for the treatment coefficients from a CCA, an MI approach, as well as of two MNAR sensitivity analyses based on MI. MNAR is implemented by changing the imputed values for the relevant treatment group in line with the MNAR assumptions made, as discussed above.

The results from the CCA and MI are consistent and lead to the same conclusions as in section 6.5.2, i.e. there is no evidence that the treatment coefficient is statistically significantly different from zero.

In line with Figure 6-1 produced by Stata's *rctmiss* command, the treatment coefficient estimates change by approximately one point when the outcomes of participants in the patellar resurfacing group are assumed to be five points better or worse than under MAR for the data set with 200 observations. However, here neither of these MNAR assumptions results in changing the treatment coefficients to be statistically significantly different from zero, as shown in Table 6-2. The trial conclusions would therefore remain the same, and be robust to the tested MNAR assumptions.

For the case study using the data set of sample size 1000, the treatment coefficients change to being statistically significantly greater than zero under the MNAR scenario assuming that either participants with missing data in the patellar resurfacing arm have outcomes that are five points better than under MAR, or that participants with missing data in the no patellar resurfacing group have outcomes that are five points worse than under MAR. This, again, indicates that this dataset is much less robust to varying the assumptions made about missing data, as discussed in section 6.5.3.

Table 6-2: Results from the MNAR sensitivity analysis whereby MNAR assumptions are applied to the multiply imputed values for the relevant treatment group

	Sample size: 200	Sample size: 1000
	Regression coefficient of patellar resurfacing vs. no patellar resurfacing (95% CI):	Regression coefficient of patellar resurfacing vs. no patellar resurfacing (95% CI):
CCA	-1.164 (-4.495, 2.166)	0.878 (-0.485, 2.241)
MI – assuming MAR	-0.924 (-4.247, 2.399)	0.858 (-0.489, 2.205)
MI after adding 5 points to imputed OKS scores for patella resurfacing arm only	0.090 (-3.179, 3.359)	1.603 (0.209, 2.997)
MI after subtracting 5 points from imputed OKS scores for patella resurfacing arm only	-2.199 (-5.446, 1.048)	-0.181 (-1.602, 1.240)
MI after adding 5 points to imputed OKS scores for no patella resurfacing arm only	0.514 (-2.779, 3.808)	1.716 (0.332, 3.100)

This analysis could be repeated to identify the smallest departure from MAR that would lead to a change in the trial conclusion, similar to the way the graph obtained from Stata's *rctmiss* command can be used.

The code for the manipulation of the imputed values can be found in Appendix 16. The advantage of using this method over using the *rctmiss* command is that the researcher can decide on the assumed MAR mechanism, which can include variables in addition to those in the analysis model. It can be easily amended to also apply the MNAR assumption to longitudinally imputed data, and is therefore more widely applicable. The very useful

graphical representation of the effects on the treatment coefficient for a range of MNAR assumption could also be replicated.

6.6 Discussion

In this chapter, the importance of performing sensitivity analysis investigating MNAR scenarios was reiterated based on the current literature, and advice on the content of such sensitivity analyses by both EMA and the FDA, as well as recommendations by a range of researchers and methodologists, was presented. Following on from this overview, two approaches of applying MNAR sensitivity analysis in Stata were described. These methods were then applied to two datasets of different sample sizes with 20% missing data, and findings were discussed. In one example, even modest departures from MAR led to a change in the trial conclusions. This could have implications for the implementation of new interventions, which may be inappropriately implemented, continued or stopped from being used more widely, depending on the trial results, if appropriate sensitivity analysis is not performed. The strength of the examples presented is that they focus on easily implementable sensitivity analysis based on accessible and plausible assumptions about departures from the MAR assumption.

In line with the advice on sensitivity analysis provided by the EMA[66], there is no concern about missing data if the results from the primary analysis and a wide range of appropriate sensitivity analyses are consistent in either providing or not providing evidence of a treatment effect. However, where the results from the sensitivity analyses are inconsistent, their effect on the conclusions must be discussed. In fact, the EMA states that "in certain circumstances, the influence of missing data is so great that it might not be possible to reliably interpret the results from the trial"[66]. Therefore, limiting the amount of missing data within the follow-up of clinical trials remains the most reliable way of producing unambiguous results that are not affected by missing data. This belief is reiterated in the literature[177, 182, 183], amongst others by Li et al[18] and Little et al[17], who

suggest that “No matter what approach is taken, there is no way to adequately test the robustness of the assumptions about the missing data required by the analysis. This need to rely on untestable assumptions regarding missing data reinforces the importance of preventing missing data in the first place. The key is to design and carry out the trial in a way that limits the problem of missing data.”

6.7 Conclusions

Carrying out of a range of sensitivity analyses, including investigations of MNAR, is crucial to assess the robustness of any primary analysis in the light of the untestable assumptions made about the underlying missing data mechanism. Vital MNAR sensitivity analysis, using plausible departures from MAR, can easily be implemented after any type of primary analysis, including those that utilise MI approaches. The use of a wide range of sensitivity analyses can either reassure researchers of the robustness of the trial results with regards to the way missing data are handled, or may indicate the need for further studies to support the trial conclusions.

Chapter 7 : Conclusions

Introduction

This final thesis chapter presents a summary of the research findings. Limitations of the research are discussed, and the implications of this work, in terms of contributions to the literature, as well as the generalisability of the findings are clarified. Finally, areas for future research are outlined.

7.1 Overview of thesis findings

The thesis started by reviewing current recommendations for the handling, analysis and reporting of missing PROMs outcome data in RCTs, and assessed current RCT reports for their adherence to these standards (Chapter 2). This review provided evidence of considerable discrepancies between methodological guidance and current practice, and called for the need to generate greater awareness of the potential bias introduced through the occurrence and inappropriate handling of missing data. Areas for improvement were identified across the entire RCT lifecycle, including missed opportunities to reduce missing data prospectively, methods to take missing data into account during the analysis, lack of clarity in reporting the quantity of missing data, and assumptions made in the analysis, as well as insufficient assessment of the robustness of these assumptions and hence the validity of the RCT results. Particular points identified in the review were the low uptake of MI, generally considered to be one of the most appropriate methods of handling missing data, the preference of focussing on a single follow-up time point over utilising the longitudinal structure of the data, where applicable, as well as a lack of appropriate sensitivity analysis. The lack of adherence to current methodological guidance with regards

to these aspects in particular was considered an important area for further research and was therefore chosen to be the focus of the following chapters.

Rates and patterns of missing data within a number of PROMs and RCTs, as well as possible predictors of missing data were investigated in Chapter 3. This work was important in its own right to generate a greater understanding of missing PROMs data in RCTs, but was also utilised to inform the subsequent simulation work.

Different approaches to the MI of PROMs outcome data where the relevant analysis considers a single follow-up time point were considered in Chapter 4. Here, the question whether MI should be applied to the composite score, subscale (where available) or item level was addressed using a comprehensive set of simulation studies considering a range of sample sizes, proportions of missing data and different missing data patterns. For missing data following primarily a unit-nonresponse pattern, and large sample sizes, as well as smaller sample sizes where the proportion of missing data is low, imputation at the composite score, subscale and item level tended to perform similarly. However, in a MAR context, imputation at the subscale and/ or score level was preferable for smaller samples with large proportions of missing data. Imputation at the item level became more beneficial as the proportion of item missingness increased. In addition, the work in Chapter 4 also addressed other factors, such as the feasibility of the relevant imputation approach, and the planned analyses, that should be taken into account when deciding on an MI approach to address missing PROMs outcomes data in RCTs.

Chapter 5 assessed the comparative performance of established and validated statistical analysis approaches for handling incomeplete outcomes in longitudinal data in a PROMs and RCT context, namely maximum likelihood (ML), multiple imputation (MI) and inverse

probability weighting (IPW). Simulation studies were used to investigate a number of scenarios which differed with respect to sample sizes, proportions of participants with missing outcome data, as well as different MAR mechanisms. This work showed that IPW is not recommended for handling missing PROMs data in RCT analysis for sample sizes of up to approximately 1,000 participants. ML and MI were found to perform similarly overall in terms of bias introduced into the treatment coefficients under MAR when the MI model adjusted only for data collected at the beginning of the trial. However, MI offered benefits over ML in terms of reducing bias in the treatment coefficient if additional outcome data could be added to the MI model. Such additional data could be other, more completely collected PROMs, or PROMs collected at additional time points, as well as clinical follow-up data. The addition of such follow-up information can be very effective in reducing bias due to missing data in the treatment coefficients. It can also contribute to making the MAR assumption more plausible, and help in creating more robust RCT results.

Throughout the thesis, culminating in Chapter 6, the importance of reducing missing data prospectively, was reiterated, as no statistical analysis can ever make up for the information lost due to missing data. The untestable assumptions underlying all analyses with missing data, including those discussed in Chapters 4 and 5 were emphasised, and the use of appropriate sensitivity analysis was defined as an essential component of any RCT reporting. In Chapter 6, the recommendations for appropriate sensitivity analysis to assess the robustness of RCT analyses in the light of missing data, including MNAR, were therefore reviewed and summarised. In a case study, two approaches to sensitivity analysis considering MNAR, which are easily implementable with standard statistical software,

were then demonstrated. The interpretation and implications of these sensitivity analyses on the study report were subsequently discussed.

7.2 Limitations of this research

The limitations of this research have been discussed in each chapter, and centre mainly around potential concerns regarding generalisability.

The review of the handling and reporting of missing PROMs outcome data in RCTs focussed on publications reporting on RCTs which used at least one of eight pre-specified PROMs. Therefore, there may be a concern as to whether these trials are representative of RCTs utilising PROMs more generally. However, this selection approach was the result of a compromise between limiting the number of PROMs but including a wide range of journals, or including all PROMs, but restricting the review to a low number of academic journals. Both approaches have their benefits and drawbacks in terms of generalisability of results. Changes in reporting standards over time, as well as the reporting of sensitivity analyses that yielded potentially different results from the main analyses, could not be assessed in this review, but form the basis of future work.

A limited number of PROMs and RCTs were investigated in Chapter 3, which identified common missing data patterns and potential predictors of missing data. The exploration of further datasets and PROMs would have yielded additional information, but without such further research it is impossible to ascertain whether this would lead to any different conclusions.

The research into benefits and disadvantages of multiply imputing missing PROMs data at either the item, subscale or composite score level, as presented in Chapter 4, solely utilised data from the KAT study, and therefore was limited to three different PROMs. Therefore, no guidance could be derived for the most appropriate imputation approach for PROMs with more than 12 items, although item imputation in those cases may yield convergence problems. It is also possible that using a different underlying dataset might have produced

different results, and a validation of the findings in a different setting would be valuable. However, the fact that results from earlier research could be reproduced is reassuring and makes it likely that findings from this work are further generalisable.

The simulation work in Chapter 5 was also based exclusively on the KAT dataset, and considered the different approaches for handling missing longitudinal PROMs data for the OKS only, which may raise concerns about the generalisability of the findings. However, it was argued that the use of the OKS as a continuous score makes generalisability more likely. Also, only a limited number of methods to address missing longitudinal PROMs data were considered. Notably, alternative, and possibly more sophisticated, approaches have been suggested in the literature, and were discussed in the chapter. These were not included in the simulation study due to the fact that they are not readily available in standard software, which prevents them from being used more widely at this point in time. However, it is acknowledged that some of these methods may be more beneficial in minimising bias arising from the analysis of missing longitudinal PROMs data. Their omission, although justifiable on practical grounds, is a limitation of this research.

Similarly, the presentation of sensitivity analyses focussed on appropriate, yet easily implementable sensitivity analyses that can be readily appreciated by the wider RCT team and others for whom the trial results are of relevance. Arguably, other, possibly more advanced, methods may allow for the assessment of more realistic and varied sensitivity scenarios.

7.3 Implications of this research

7.3.1 Contribution to the literature

This thesis has contributed to the literature as follows:

1. By establishing current practice in the handling, analysis and reporting of missing PROMs outcome data in RCTs;
2. By generating a better understanding of missing data rates, patterns and mechanisms in PROMs collected within RCTs.
3. By validating and expanding the evidence base regarding whether MI should be applied at the composite score, subscale (where available) or item level when addressing missing PROMs outcome data in RCTs.
4. By evaluating the comparative performance of established statistical methods for analysing longitudinal outcome data, namely ML, MI and IPW, when handling repeatedly measured PROMs outcome data with missing observations in RCTs.
5. By reviewing recommendations for sensitivity analysis to assess the robustness of trial results in the light of missing data, and presenting two intuitive and easily implementable approaches of performing sensitivity analysis considering MNAR data.

To address the first aim, current RCT publications were reviewed, considering adherence to existing methodological guidance. Existing reviews on this topic were added to by reviewing recently published papers and considering steps taken to minimise missing data occurrence, as well as observing the use of appropriate sensitivity. The review focused RCTs utilising eight pre-specified PROMs. Since larger numbers of specific PROMs were identified than in previous reviews, the findings for the standards of handling and reporting missing

data within these PROMs are more robust and generalisable. This research has been presented at a number of conferences, both national and internationally, and has been published in a peer reviewed journal[39].

A better understanding of missing data in RCTs, including rates and patterns of missing data, as well as possible predictors of missing data has been generated within Chapter 3. Through the detailed investigation of missing data patterns across a number of trials, typical missing data patterns were identified. These varied between trials, possibly related to the patient population, to the particular PROMs, or due to differences in their complexity and perceived relevance. However, similarities in terms of higher levels of unit-nonresponse compared to item missingness were identified on a cross-sectional level. Longitudinally, mixtures of intermittent and monotone missingness were identified. This work thus contributes to a better understanding of missing data in PROMs.

The third aim, extending existing research regarding the level at which MI for PROMs data should be applied is also an important contribution to the existing methodological literature and provides researchers with important guidance in the analysis of RCTs, which increasingly utilise PROMs. This is an important contribution to the literature, as the literature review performed in Chapter 2 identified low rates of MI being performed in the RCT analyses of PROMs outcome data. This research reiterated that MI is a very appropriate method of handling missing PROMs outcome data under MAR, and also clarifies whether imputations should be performed at the composite score, subscale or item level, depending on the characteristics of the dataset and missing data within it. Therefore, it is hoped that this work will help to raise reporting standards.

The fourth aim, investigating the comparative performance of established longitudinal analysis methods for PROMs data, was addressed in Chapter 5. This is the first simulation study directly comparing the performance of ML, MI and IPW focussing on PROMs data in an RCT context. Therefore, the findings presented answer important methodological questions, providing guidance to researchers analysing PROMs data.

Chapter 6 focussed on providing guidance on the performance of adequate sensitivity analysis investigating the effect of missing data on trial results. The work in Chapter 2 has identified a distinct lack of appropriate sensitivity analysis in RCT publications. By reviewing the recommended components of sensitivity analyses for missing data and providing examples of easily implementable sensitivity analyses for data MNAR, as well as discussing their interpretation and implications, important guidance is derived for researchers wanting to perform such analyses.

7.3.2 Generalisability of findings

Despite some of the limitations of this research, the findings presented in this thesis are important and largely generalisable.

The review of the current practice of the handling and analysis of missing PROMs outcome data did not restrict the journals included in the review and is therefore, as a minimum, generalisable to analysis and reporting standards of RCTs utilising the eight PROMs used within this review. Also, the findings of the review are in line with results from similar reviews that did not restrict the type of PROMs used, and it can therefore be concluded that the review represents current practice of the handling and reporting of missing PROMs data within RCTs more generally.

The key findings from Chapter 3, mainly that missing data is prevalent in RCTs, but that there is variation between different trials as well as between PROMs, in terms of missing data rates and pattern, and that the identification of possible missing data mechanisms is challenging, are thought to be generalisable beyond the three trials considered.

Although the research on the comparison of different imputation approaches, i.e. applying MI at the item, subscale or composite score level (Chapter 4), was limited to a single RCT data set and three questionnaires, the research is thought to be generalisable to a wider range of similar PROMs. This is because, firstly, findings from previous research were reproduced in the KAT data sets. Also, a range of different missing data patterns, ranging from unit-nonresponse to patterns consisting primarily of item-nonresponse were investigated. These patterns will be relevant for other RCTs and PROMs, and conclusions for the PROMs investigated here were consistent under similar scenarios. This consistency leads to the assumptions that comparable PROMs will also yield similar results under corresponding settings, and that findings are therefore likely to be generalisable more widely.

The same principle applies to the research presented in Chapter 5 on different statistical approaches to handling missing longitudinal PROMs data. Although only one dataset and one PROM, namely the OKS, was used, findings are thought to be generalisable due to a number of factors. Firstly, the OKS was used as a continuous score, which is in line with the analyses of many PROMs. Secondly, a range of missingness patterns was investigated, considering not only a range of monotone and intermittent missingness patterns, but also different MAR mechanisms. Therefore, the findings are expected to be generalisable to any RCT settings where a missing PROMs pattern can be likened to one of the investigated scenarios.

Finally, the proposed sensitivity analyses presented in Chapter 6 are easily implementable using standard statistical software, and can be applied to a wide range of trial analyses, which makes them generalisable to most PROMs analyses conducted within the context of clinical trials.

In conclusion, the research presented in this thesis is expected to be valuable and generalisable to a wide range of clinical trial settings.

7.4 Future research

Future research will expand and validate the work performed within this thesis by using data from the national PROMs study and the Hospital Episode Statistics (HES). This data combines information from PROMs with clinical data and a large amount of demographic information and patient characteristics for a high proportion of patients undergoing elective hip or knee surgery in England. These data are considered valuable in generating a more in-depth understanding of the relationship between missing outcome data and baseline data, as the large numbers of patients included in this dataset are likely to enable the identification of more subtle correlations. This work, which is planned to be undertaken over the next year, will also extend and validate the simulation studies considering imputation at the composite score, subscale and item level, as presented in Chapter 4. Specifically, larger sample sizes will be considered, and findings will be validated using a separate and larger dataset.

The work on improving the handling and reporting of missing PROMs outcome data, and missing data more generally, is also continuing, with additional reviews into this area having been published by other authors since the work on Chapter 2 concluded[61]. The impact of the guidance generated here, together with other guidance in the existing literature, on the importance of minimising missing data, appropriate analyses, accounting for the uncertainty around the imputed values where imputation approaches are used, the clear reporting of assumptions made about missing data in the analysis, as well as the use of appropriate sensitivity analysis, is of great interest. Therefore, additional reviews are planned to assess whether reporting standards change over time with regards to these important factors.

7.5 Concluding remarks

Missing data is widely recognised as a potential source of major bias in the analysis and reporting of RCTs. Health care decisions that are based on RCT results which are biased due to the inappropriate handling of missing data can negatively impact patient care. Therefore, research in this area is of high importance. Through detailed investigations of RCT data and a review of the current literature, this thesis confirms that missing data is prevalent in RCTs, and that there are considerable discrepancies between current and best practice in the handling and reporting of missing data. Additional work presented in this thesis offers guidance on analytical approaches to minimise bias in the analysis of RCT data with missing PROMs outcome data, both for analyses utilising MI for a single follow-up time point, and for analyses considering longitudinal follow-up. The importance of sensitivity analyses to assess the robustness of results based on untestable assumptions made about the missing data is also reiterated. Practical and easily implementable proposals to implement such sensitivity analyses are provided. Throughout the thesis, it is acknowledged that while some missing data is unavoidable in RCTs, and that the analysis approaches proposed are important to address this missing data, emphasis also needs to be put on methods to prospectively minimise missing data occurrence. Both appropriate methodologies for handling missing data, and limiting missing data prospectively are vital factors in the generation of robust RCT results. The potential impact of any remaining missing data should always be assessed in comprehensive sensitivity analyses.

References

1. Altman DG. Better reporting of randomised controlled trials: the CONSORT statement. *Bmj* 1996; **313**(7057): 570-1.
2. Odgaard-Jensen J, Vist GE, Timmer A, et al. Randomisation to protect against selection bias in healthcare trials. *The Cochrane database of systematic reviews* 2011; (4): MR000012.
3. Edwards SJ, Lilford RJ, Hewison J. The ethics of randomised controlled trials from the perspectives of patients, the public, and healthcare professionals. *Bmj* 1998; **317**(7167): 1209-12.
4. Sibbald B, Roland M. Understanding controlled trials. Why are randomised controlled trials important? *Bmj* 1998; **316**(7126): 201.
5. Guyatt G, Feeny D, Patrick D. Measuring health-related quality of life. *Ann Intern Med* 1993; **118**(8): 622-9.
6. Black N. Patient reported outcome measures could help transform healthcare. *Bmj* 2013; **346**: f167.
7. Varagunam M, Hutchings A, Neuburger J, Black N. Impact on hospital performance of introducing routine patient reported outcome measures in surgery. *Journal of health services research & policy* 2014; **19**(2): 77-84.
8. Patrick DL, Burke LB, Powers JH, et al. Patient-reported outcomes to support medical product labeling claims: FDA perspective. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research* 2007; **10 Suppl 2**: S125-37.
9. Jenkinson C, Morley D. Patient reported outcomes. *European journal of cardiovascular nursing : journal of the Working Group on Cardiovascular Nursing of the European Society of Cardiology* 2016; **15**(2): 112-3.
10. Wells GA, Russell AS, Haraoui B, Bissonnette R, Ware CF. Validity of quality of life measurement tools--from generic to disease-specific. *The Journal of rheumatology Supplement* 2011; **88**: 2-6.
11. Kazi AM, Khalid W. Questionnaire designing and validation. *JPMA The Journal of the Pakistan Medical Association* 2012; **62**(5): 514-6.
12. Beard DJ, Harris K, Dawson J, et al. Meaningful changes for the Oxford hip and knee scores after joint replacement surgery. *J Clin Epidemiol* 2015; **68**(1): 73-9.
13. Dawson J, Rogers K, Fitzpatrick R, Carr A. The Oxford shoulder score revisited. *Archives of orthopaedic and trauma surgery* 2009; **129**(1): 119-23.
14. Lohr KN, Zebrack BJ. Using patient-reported outcomes in clinical practice: challenges and opportunities. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation* 2009; **18**(1): 99-107.
15. Altman DG. Missing outcomes in randomized trials: addressing the dilemma. *Open medicine : a peer-reviewed, independent, open-access journal* 2009; **3**(2): e51-3.
16. Hutchings A, Neuburger J, Grosse Frie K, Black N, van der Meulen J. Factors associated with non-response in routine use of patient reported outcome measures after elective surgery in England. *Health and quality of life outcomes* 2012; **10**: 34.
17. Little RJ, D'Agostino R, Cohen ML, et al. The prevention and treatment of missing data in clinical trials. *The New England journal of medicine* 2012; **367**(14): 1355-60.

18. Li T, Hutfless S, Scharfstein DO, et al. Standards should be applied in the prevention and handling of missing data for patient-centered outcomes research: a systematic review and expert consensus. *J Clin Epidemiol* 2014; **67**(1): 15-32.
19. Cook JA, Hislop JM, Altman DG, et al. Use of methods for specifying the target difference in randomised controlled trial sample size calculations: Two surveys of trialists' practice. *Clinical trials* 2014.
20. Bell ML, Fairclough DL. Practical and statistical issues in missing data for longitudinal patient reported outcomes. *Statistical methods in medical research* 2013.
21. Little RJA, Rubin DB. Statistical analysis with missing data. 2nd ed. Hoboken, New Jersey: John Wiley & Sons; 2002.
22. Fielding S, Maclennan G, Cook JA, Ramsay CR. A review of RCTs in four medical journals to assess the use of imputation to overcome missing data in quality of life outcomes. *Trials* 2008; **9**: 51.
23. Eekhout I, de Boer RM, Twisk JW, de Vet HC, Heymans MW. Missing data: a systematic review of how they are reported and handled. *Epidemiology* 2012; **23**(5): 729-32.
24. Rombach I, Rivero-Arias O, Gray AM, Jenkinson C, Burke O. The current practice of handling and reporting missing outcome data in eight widely used PROMs in RCT publications: a review of the current literature. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation* 2016; **25**(7): 1613-23.
25. Bell ML, Fiero M, Horton NJ, Hsu CH. Handling missing data in RCTs; a review of the top medical journals. *BMC medical research methodology* 2014; **14**(1): 118.
26. Carpenter JR, Roger JH, Kenward MG. Analysis of longitudinal trials with protocol deviation: a framework for relevant, accessible assumptions, and inference via multiple imputation. *Journal of biopharmaceutical statistics* 2013; **23**(6): 1352-71.
27. Streiner DL. Missing data and the trouble with LOCF. *Evidence-based mental health* 2008; **11**(1): 3-5.
28. Kenward MG, Carpenter J. Multiple imputation: current perspectives. *Statistical methods in medical research* 2007; **16**(3): 199-218.
29. Fielding S, Fayers P, Ramsay CR. Analysing randomised controlled trials with missing data: choice of approach affects conclusions. *Contemporary clinical trials* 2012; **33**(3): 461-9.
30. Genolini C, Lacombe A, Ecochard R, Subtil F. CopyMean: A new method to predict monotone missing values in longitudinal studies. *Computer methods and programs in biomedicine* 2016; **132**: 29-44.
31. Resche-Rigon M, White IR. Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Statistical methods in medical research* 2016.
32. Liu B, Yu M, Graubard BI, Troiano RP, Schenker N. Multiple imputation of completely missing repeated measures data within persons from a complex sample: application to accelerometer data in the National Health and Nutrition Examination Survey. *Statistics in medicine* 2016.
33. Belger M, Haro JM, Reed C, et al. How to deal with missing longitudinal data in cost of illness analysis in Alzheimer's disease-suggestions from the GERAS observational study. *BMC medical research methodology* 2016; **16**: 83.

34. Bunouf P, Molenberghs G. Implementation of pattern-mixture models in randomized clinical trials. *Pharmaceutical statistics* 2016.
35. Shin T, Davison ML, Long JD. Maximum Likelihood Versus Multiple Imputation for Missing Data in Small Longitudinal Samples With Nonnormality. *Psychological methods* 2016.
36. Erler NS, Rizopoulos D, Rosmalen J, Jaddoe VW, Franco OH, Lesaffre EM. Dealing with missing covariates in epidemiologic studies: a comparison between multiple imputation and a full Bayesian approach. *Statistics in medicine* 2016; **35**(17): 2955-74.
37. Croudace T, Brazier J, Gutacker N, et al. Proceedings of Patient Reported Outcome Measure's (PROMs) Conference Sheffield 2016: advances in patient reported outcomes research. *Health and quality of life outcomes* 2016; **14**(1): 137.
38. Kenward MG, Molenberghs G. Last observation carried forward: a crystal ball? *Journal of biopharmaceutical statistics* 2009; **19**(5): 872-88.
39. Rombach I, Rivero-Arias O, Gray AM, Jenkinson C, Burke O. The current practice of handling and reporting missing outcome data in eight widely used PROMs in RCT publications: a review of the current literature. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation* 2016.
40. O'Neill RT, Temple R. The prevention and treatment of missing data in clinical trials: an FDA perspective on the importance of dealing with it. *Clinical pharmacology and therapeutics* 2012; **91**(3): 550-4.
41. Dawson J, Fitzpatrick R, Murray D, Carr A. Questionnaire on the perceptions of patients about total knee replacement. *Journal of Bone & Joint Surgery - British Volume* 1998; **80**(1): 63-9.
42. Murray DW, Fitzpatrick R, Rogers K, et al. The use of the Oxford hip and knee scores. *The Journal of bone and joint surgery British volume* 2007; **89**(8): 1010-4.
43. Jenkinson C, Layte R. Development and testing of the UK SF-12 (short form health survey). *Journal of health services research & policy* 1997; **2**(1): 14-8.
44. Jenkinson C, Layte R, Jenkinson D, et al. A shorter form health survey: can the SF-12 replicate results from the SF-36 in longitudinal studies? *Journal of public health medicine* 1997; **19**(2): 179-86.
45. EuroQol G. EuroQol--a new facility for the measurement of health-related quality of life. *Health Policy* 1990; **16**(3): 199-208.
46. Oemar M, Oppe M. EQ-5D-3L User Guide - Basic information on how to use the EQ-5D-3L instrument, Version 5.0. 2013. http://www.euroqol.org/fileadmin/user_upload/Documenten/PDF/Folders_Flyers/EQ-5D-3L_UserGuide_2013_v5.0_October_2013.pdf (accessed 01 October 2014).
47. Jenkinson C, Dummett S, Kelly L, et al. The development and validation of a quality of life measure for the carers of people with Parkinson's disease (the PDQ-Carer). *Parkinsonism & related disorders* 2012; **18**(5): 483-7.
48. Jenkinson C, Fitzpatrick R, Peto V, Dummett S, Morley D, Saunders P. The Parkinson's Disease Questionnaires User Manual (PDQ-39, PDQ-8, PDQ Summary Index & PDQ-Carer). Oxford: Oxford : Health Services Research Unit, University of Oxford; 2012.
49. Vickers AJ, Altman DG. Statistics notes: missing outcomes in randomised trials. *Bmj* 2013; **346**: f3438.
50. Carpenter JR, Kenward MG. Multiple Imputation and its Application. 1st ed. Chichester: John Wiley & Sons; 2013.

51. Little RJ, Cohen ML, Dickersin K, et al. The design and conduct of clinical trials to limit missing data. *Statistics in medicine* 2012; **31**(28): 3433-43.
52. Eekhout I, de Vet HC, Twisk JW, Brand JP, de Boer MR, Heymans MW. Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. *J Clin Epidemiol* 2014; **67**(3): 335-42.
53. Fayers PM, Curran D, Machin D. Incomplete quality of life data in randomized trials: missing items. *Statistics in medicine* 1998; **17**(5-7): 679-96.
54. Curran D, Molenberghs G, Fayers PM, Machin D. Incomplete quality of life data in randomized trials: missing forms. *Statistics in medicine* 1998; **17**(5-7): 697-709.
55. Baraldi AN, Enders CK. An introduction to modern missing data analyses. *Journal of school psychology* 2010; **48**(1): 5-37.
56. Sainani KL. Dealing With Missing Data. *Pm&R* 2015; **7**(9): 990-4.
57. Bartlett JW, Frost C, Carpenter JR. Multiple imputation models should incorporate the outcome in the model of interest. *Brain* 2011; **134**(Pt 11): e189; author reply e90.
58. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in medicine* 2011; **30**(4): 377-99.
59. Mallinckrodt CH, Lane PW, Schnell D, Peng Y, Mancuso JP. Recommendations for the primary analysis of continuous endpoints in longitudinal clinical trials. *Drug Inf J* 2008; **42**(4): 303-19.
60. Eekhout I, Enders CK, Twisk JW, de Boer MR, de Vet HC, Heymans MW. Including auxiliary item information in longitudinal data analyses improved handling missing questionnaire outcome data. *J Clin Epidemiol* 2015.
61. Fielding S, Ogbuagu A, Sivasubramaniam S, MacLennan G, Ramsay CR. Reporting and dealing with missing quality of life data in RCTs: has the picture changed in the last decade? *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation* 2016.
62. Powney M, Williamson P, Kirkham J, Kolamunnage-Dona R. A review of the handling of missing longitudinal outcome data in clinical trials. *Trials* 2014; **15**: 237.
63. Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical trials* 2004; **1**(4): 368-76.
64. Noble SM, Hollingworth W, Tilling K. Missing data in trial-based cost-effectiveness analysis: the current state of play. *Health Econ* 2012; **21**(2): 187-200.
65. Deo A, Schmid CH, Earley A, Lau J, Uhlig K. Loss to analysis in randomized controlled trials in CKD. *American journal of kidney diseases : the official journal of the National Kidney Foundation* 2011; **58**(3): 349-55.
66. EMA. Guideline on Missing Data in Confirmatory Clinical Trials2 July 2010, 2010. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/09/WC500096793.pdf (accessed 10/02/2014).
67. Faria R, Gomes M, Epstein D, White IR. A guide to handling missing data in cost-effectiveness analysis conducted within randomised controlled trials. *Pharmacoeconomics* 2014; **32**(12): 1157-70.
68. Fielding S, Fayers PM, Loge JH, Jordhoy MS, Kaasa S. Methods for handling missing data in palliative care research. *Palliative medicine* 2006; **20**(8): 791-8.
69. Fielding S, Fayers PM, McDonald A, McPherson G, Campbell MK, Group RS. Simple imputation methods were inadequate for missing not at random (MNAR) quality of life data. *Health and quality of life outcomes* 2008; **6**: 57.

70. Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *JAMA : the journal of the American Medical Association* 1996; **276**(8): 637-9.
71. Calvert M, Blazeby J, Altman DG, et al. Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension. *JAMA : the journal of the American Medical Association* 2013; **309**(8): 814-22.
72. Singer L, Gould M, Tomlinson G, Theodore J. Determinants of health utility in lung and heart-lung transplant recipients. *Am J Transplant* 2005; **5**(1): 103-9.
73. Horsman J, Furlong W, Feeny D, Torrance G. The Health Utilities Index (HUI): concepts, measurement properties and applications. *Health and quality of life outcomes* 2003; **1**: 54.
74. Dawson J, Fitzpatrick R, Carr A, Murray D. Questionnaire on the perceptions of patients about total hip replacement. *Journal of Bone & Joint Surgery - British Volume* 1996; **78**(2): 185-90.
75. Peto V, Jenkinson C, Fitzpatrick R. PDQ-39: a review of the development, validation and application of a Parkinson's disease quality of life questionnaire and its associated measures. *Journal of neurology* 1998; **245 Suppl 1**: S10-4.
76. Guo Z, Tang HY, Li H, et al. The benefits of psychosocial interventions for cancer patients undergoing radiotherapy. *Health and quality of life outcomes* 2013; **11**: 121.
77. Aaronson N, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials in oncology. *Journal of the National Cancer Institute* 1993; **85**(5): 365-76.
78. Perneger TV, Burnand B. A simple imputation algorithm reduced missing data in SF-12 health surveys. *J Clin Epidemiol* 2005; **58**(2): 142-9.
79. Ware JE, Jr., Gandek B. Overview of the SF-36 Health Survey and the International Quality of Life Assessment (IQOLA) Project. *J Clin Epidemiol* 1998; **51**(11): 903-12.
80. Lepage L, Altman D, Schulz K, et al. The revised CONSORT statement for reporting randomized trials: Explanation and elaboration. *Ann Intern Med* 2001; **134**(8): 663-94.
81. Levin LA, Wallentin L, Bernfort L, et al. Health-related quality of life of ticagrelor versus clopidogrel in patients with acute coronary syndromes-results from the PLATO trial. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research* 2013; **16**(4): 574-80.
82. Hopewell S, Dutton S, Yu LM, Chan AW, Altman DG. The quality of reports of randomised trials in 2000 and 2006: comparative study of articles indexed in PubMed. *Bmj* 2010; **340**: c723.
83. Molnar FJ, Hutton B, Fergusson D. Does analysis using "last observation carried forward" introduce bias in dementia research? *Cmaj* 2008; **179**(8): 751-3.
84. Nilsdotter A, Bremander A. Measures of hip function and symptoms: Harris Hip Score (HHS), Hip Disability and Osteoarthritis Outcome Score (HOOS), Oxford Hip Score (OHS), Lequesne Index of Severity for Osteoarthritis of the Hip (LISOH), and American Academy of Orthopedic Surgeons (AAOS) Hip and Knee Questionnaire. *Arthritis care & research* 2011; **63 Suppl 11**: S200-7.
85. Collins NJ, Misra D, Felson DT, Crossley KM, Roos EM. Measures of knee function: International Knee Documentation Committee (IKDC) Subjective Knee Evaluation Form, Knee Injury and Osteoarthritis Outcome Score (KOOS), Knee Injury and Osteoarthritis Outcome Score Physical Function Short Form (KOOS-PS), Knee Outcome Survey Activities

of Daily Living Scale (KOS-ADL), Lysholm Knee Scoring Scale, Oxford Knee Score (OKS), Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), Activity Rating Scale (ARS), and Tegner Activity Score (TAS). *Arthritis care & research* 2011; **63 Suppl 11**: S208-28.
86. Gill SD, de Morton NA, Mc Burney H. An investigation of the validity of six measures of physical function in people awaiting joint replacement surgery of the hip or knee. *Clin Rehabil* 2012; **26**(10): 945-51.
87. Jenkinson C, Fitzpatrick R, Peto V, Greenhall R, Hyman N. The Parkinson's Disease Questionnaire (PDQ-39): development and validation of a Parkinson's disease summary index score. *Age and ageing* 1997; **26**(5): 353-7.
88. Morley D, Dummett S, Kelly L, et al. The PDQ-Carer: development and validation of a summary index score. *Parkinsonism & related disorders* 2013; **19**(4): 448-9.
89. Brooks R. EuroQol: the current state of play. *Health Policy* 1996; **37**(1): 53-72.
90. Jenkinson C, Chandola T, Coulter A, Bruster S. An assessment of the construct validity of the SF-12 summary scores across ethnic groups. *Journal of public health medicine* 2001; **23**(3): 187-94.
91. Breeman S, Campbell M, Dakin H, et al. Patellar resurfacing in total knee replacement: five-year clinical and economic results of a large randomized controlled trial. *J Bone Joint Surg Am* 2011; **93**(16): 1473-81.
92. MacLennan G. The knee arthroplasty trial (KAT) design features, baseline characteristics, and two-year functional outcomes after alternative approaches to knee replacement. *J Bone Jt Surg Ser A* 2009; **91**(1): 134-41.
93. Murray DW, MacLennan GS, Breeman S, et al. A randomised controlled trial of the clinical effectiveness and cost-effectiveness of different knee prostheses: the Knee Arthroplasty Trial (KAT). *Health technology assessment (Winchester, England)* 2014; **18**(19): 1-235, vii-viii.
94. Group PDMC, Gray R, Ives N, et al. Long-term effectiveness of dopamine agonists and monoamine oxidase B inhibitors compared with levodopa as initial treatment for Parkinson's disease (PD MED): a large, open-label, pragmatic randomised trial. *Lancet* 2014; **384**(9949): 1196-205.
95. PD MED study protocol. 2010. http://www.birmingham.ac.uk/Documents/college-mds/trials/bctu/PDMed/Investigators/PD-MED-Protocol-Version-8.pdf (accessed 29/08 2014).
96. Jenkinson C, Williams A, Ives N, et al. Parkinson's disease questionnaire (PDQ-39) as a primary endpoint in a trial comparing deep brain stimulation with best medical therapy versus best medical therapy alone for advanced parkinson's disease (PD SURG trial): A randomised, open-label trial. *Value in Health* 2012; **15**(4): A148.
97. PD SURG study protocol. 2009. http://www.birmingham.ac.uk/Documents/college-mds/trials/bctu/PDSurg/Investigators/Documentations/PDSURGProtocolV62.pdf (accessed 29/08 2014)).
98. Haukoos JS, Newgard CD. Advanced statistics: missing data in clinical research--part 1: an introduction and conceptual framework. *Acad Emerg Med* 2007; **14**(7): 662-8.
99. Lang KM, Little TD. Principled Missing Data Treatments. *Prevention science : the official journal of the Society for Prevention Research* 2016.

100. Jenkinson C, Heffernan C, Doll H, Fitzpatrick R. The Parkinson's Disease Questionnaire (PDQ-39): evidence for a method of imputing missing data. *Age and ageing* 2006; **35**(5): 497-502.
101. Parenti N, Reggiani ML, Percudani D, Melotti RM. Reliability of American Society of Anesthesiologists physical status classification. *Indian journal of anaesthesia* 2016; **60**(3): 208-14.
102. Sankar A, Johnson SR, Beattie WS, Tait G, Wijeysundera DN. Reliability of the American Society of Anesthesiologists physical status scale in clinical practice. *British journal of anaesthesia* 2014; **113**(3): 424-32.
103. Beale EML. Note on Procedures for Variable Selection in Multiple Regression. *Technometrics* 1970; **12**(4): 909-14.
104. Derksen S, Keselman HJ. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology* 1992; **45**(2): 265-82.
105. Hurvich CM, Tsai C-L. The Impact of Model Selection on Inference in Linear Regression. *The American statistician* 1990; **44**(3): 214-7.
106. Sauerbrei W, Royston P, Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. *Statistics in medicine* 2007; **26**(30): 5512-28.
107. White IR, Horton NJ, Carpenter J, Pocock SJ. Strategy for intention to treat analysis in randomised trials with missing outcome data. *Bmj* 2011; **342**: d40.
108. Curran D, Bacchi M, Schmitz SF, Molenberghs G, Sylvester RJ. Identifying the types of missingness in quality of life data from clinical trials. *Statistics in medicine* 1998; **17**(5-7): 739-56.
109. Peyre H, Leplege A, Coste J. Missing data methods for dealing with missing items in quality of life questionnaires. A comparison by simulation of personal mean score, full information maximum likelihood, multiple imputation, and hot deck techniques applied to the SF-36 in the French 2003 decennial health survey. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation* 2011; **20**(2): 287-300.
110. Blough DK, Ramsey S, Sullivan SD, Yusen R, Nett Research G. The impact of using different imputation methods for missing quality of life scores on the estimation of the cost-effectiveness of lung-volume-reduction surgery. *Health Econ* 2009; **18**(1): 91-101.
111. Harris K, Dawson J, Doll H, et al. Can pain and function be distinguished in the Oxford Knee Score in a meaningful way? An exploratory and confirmatory factor analysis. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation* 2013; **22**(9): 2561-8.
112. Simons CL, Rivero-Arias O, Yu LM, Simon J. Multiple imputation to deal with missing EQ-5D-3L data: Should we impute individual domains or the actual index? *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation* 2015; **24**(4): 805-15.
113. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Statistics in medicine* 2006; **25**(24): 4279-92.
114. Crowther MJ, Lambert PC. Simulating biologically plausible complex survival data. *Statistics in medicine* 2013; **32**(23): 4118-34.
115. StataCorp. Stata Statistical Software: Release 14. *College Station* 2015; **TX: StataCorp LP**.

116. Van Buuren A, Brand JPL, Groothuis-Oudshoorn CGM, Rubin DB. Fully conditional specification in multivariate imputation. *Journal of statistical computation and simulation* 2006; **76**(12): 1049-64.
117. Yu LM, Burton A, Rivero-Arias O. Evaluation of software for multiple imputation of semi-continuous data. *Statistical methods in medical research* 2007; **16**(3): 243-58.
118. Brand JPL, van Buuren S, Groothuis-Oudshoorn K, Gelsema ES. A toolkit in SAS for the evaluation of multiple imputation methods. *Statistica Neerlandica* 2003; **57**(1): 36-45.
119. Khan I, Morris S, Pashayan N, Matata B, Bashir Z, Maguirre J. Comparing the mapping between EQ-5D-5L, EQ-5D-3L and the EORTC-QLQ-C30 in non-small cell lung cancer patients. *Health and quality of life outcomes* 2016; **14**: 60.
120. Bartlett J. FAQs - missing responses - Missing outcomes in a two arm trial. 2012 (accessed August 2010).
121. Chai T, Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geosci Model Dev* 2014; **7**(3): 1247-50.
122. Molenberghs G, Thijs H, Jansen I, et al. Analyzing incomplete longitudinal clinical trial data. *Biostatistics* 2004; **5**(3): 445-64.
123. Horton NJ, White IR, Carpenter J. The performance of multiple imputation for missing covariates relative to complete case analysis. *Statistics in medicine* 2010; **29**(12): 1357; author reply 8.
124. Shen S, Beunckens C, Mallinckrodt C, Molenberghs G. A local influence sensitivity analysis for incomplete longitudinal depression data. *Journal of biopharmaceutical statistics* 2006; **16**(3): 365-84.
125. Beunckens C, Molenberghs G, Kenward MG. Direct likelihood analysis versus simple forms of imputation for missing data in randomized clinical trials. *Clinical trials* 2005; **2**(5): 379-86.
126. Mallinckrodt CH, Clark SW, Carroll RJ, Molenbergh G. Assessing response profiles from incomplete longitudinal clinical trial data under regulatory considerations. *Journal of biopharmaceutical statistics* 2003; **13**(2): 179-90.
127. White IR, Moodie E, Thompson SG, Croudace T. A modelling strategy for the analysis of clinical trials with partly missing longitudinal data. *Int J Methods Psychiatr Res* 2003; **12**(3): 139-50.
128. Enders CK. Analyzing longitudinal data with missing values. *Rehabil Psychol* 2011; **56**(4): 267-88.
129. Mallinckrodt CH, Sanger TM, Dube S, et al. Assessing and interpreting treatment effects in longitudinal clinical trials with missing data. *Biological psychiatry* 2003; **53**(8): 754-60.
130. Lachin JM. Fallacies of last observation carried forward analyses. *Clinical trials* 2016; **13**(2): 161-8.
131. Liublinska V, Rubin DB. Sensitivity analysis for a partially missing binary outcome in a two-arm randomized clinical trial. *Statistics in medicine* 2014; **33**(24): 4170-85.
132. Ibrahim JG, Molenberghs G. Missing data methods in longitudinal studies: a review. *Test* 2009; **18**(1): 1-43.
133. Verbeke G, Fieuws S, Molenberghs G, Davidian M. The analysis of multivariate longitudinal data: a review. *Statistical methods in medical research* 2014; **23**(1): 42-59.
134. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological methods* 2001; **6**(4): 330-51.

135. Doidge JC. Responsiveness-informed multiple imputation and inverse probability-weighting in cohort studies with missing data that are non-monotone or not missing at random. *Statistical methods in medical research* 2016.
136. Fielding S, Fayers P, Ramsay C. Predicting missing quality of life data that were later recovered: an empirical comparison of approaches. *Clinical trials* 2010; **7**(4): 333-42.
137. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in medicine* 2010; **29**(28): 2920-31.
138. Donneau AF, Mauer M, Lambert P, Molenberghs G, Albert A. Simulation-based study comparing multiple imputation methods for non-monotone missing ordinal data in longitudinal settings. *Journal of biopharmaceutical statistics* 2014.
139. Welch CA, Petersen I, Bartlett JW, et al. Evaluation of two-fold fully conditional specification multiple imputation for longitudinal electronic health record data. *Statistics in medicine* 2014; **33**(21): 3725-37.
140. Biering K, Hjollund NH, Frydenberg M. Using multiple imputation to deal with missing data and attrition in longitudinal studies with repeated measures of patient-reported outcomes. *Clinical epidemiology* 2015; **7**: 91-106.
141. Rasbash J, Charlton C, Browne WJ, Healy M, Cameron B. MLwiN Version 2.1. *Centre for Multilevel Modelling* 2009; **University of Bristol**.
142. Mansournia MA, Altman DG. Inverse probability weighting. *Bmj* 2016; **352**: i189.
143. Carpenter JR, Kenward MG, Vansteelandt S. A Comparison of Multiple Imputation and Doubly Robust Estimation for Analyses with Missing Data. *Journal of the Royal Statistical Society Series A (Statistics in Society)* 2006; **169**(3): 571-84.
144. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research* 2013; **22**(3): 278-95.
145. Seaman SR, White IR, Copas AJ, Li L. Combining multiple imputation and inverse-probability weighting. *Biometrics* 2012; **68**(1): 129-37.
146. Kang YG, Lee JT, Kang JY, Kim GH, Kim TK. Analysis of Longitudinal Outcome Data with Missing Values in Total Knee Arthroplasty. *J Arthroplasty* 2016; **31**(1): 81-6.
147. Ferro MA. Missing data in longitudinal studies: cross-sectional multiple imputation provides similar estimates to full-information maximum likelihood. *Annals of epidemiology* 2014; **24**(1): 75-7.
148. Twisk J, de Boer M, de Vente W, Heymans M. Multiple imputation of missing values was not necessary before performing a longitudinal mixed-model analysis. *J Clin Epidemiol* 2013; **66**(9): 1022-8.
149. Kadengye DT, Cools W, Ceulemans E, Van den Noortgate W. Simple imputation methods versus direct likelihood analysis for missing item scores in multilevel educational data. *Behavior research methods* 2012; **44**(2): 516-31.
150. Plumpton CO, Morris T, Hughes DA, White IR. Multiple imputation of multiple multi-item scales when a full imputation model is infeasible. *BMC research notes* 2016; **9**(1): 45.
151. White IR, Kalaitzaki E, Thompson SG. Allowing for missing outcome data and incomplete uptake of randomised interventions, with application to an Internet-based alcohol trial. *Statistics in medicine* 2011; **30**(27): 3192-207.
152. Horton NJ, Kleinman KP. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American statistician* 2007; **61**(1): 79-90.

153. Eekhout I, Enders CK, Twisk JW, de Boer MR, de Vet HC, Heymans MW. Including auxiliary item information in longitudinal data analyses improved handling missing questionnaire outcome data. *J Clin Epidemiol* 2015; **68**(6): 637-45.
154. Cornish RP, Tilling K, Boyd A, Davies A, Macleod J. Using linked educational attainment data to reduce bias due to missing outcome data in estimates of the association between the duration of breastfeeding and IQ at 15 years. *International journal of epidemiology* 2015; **44**(3): 937-45.
155. Graham JW. Missing data analysis: making it work in the real world. *Annual review of psychology* 2009; **60**: 549-76.
156. Hebert PL, Taylor LT, Wang JJ, Bergman MA. Methods for using data abstracted from medical charts to impute longitudinal missing data in a clinical trial. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research* 2011; **14**(8): 1085-91.
157. Twisk JW. Longitudinal data analysis. A comparison between generalized estimating equations and random coefficient analysis. *European journal of epidemiology* 2004; **19**(8): 769-76.
158. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* 1997; **53**(3): 983-97.
159. Cook NR. Restricted Maximum Likelihood: Introduction. Wiley StatsRef: Statistics Reference Online: John Wiley & Sons, Ltd; 2014.
160. Harville DA. Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association* 1977; **72**(358): 320-38.
161. Mallinckrodt CH, Lin Q, Molenberghs M. A structured framework for assessing sensitivity to missing data assumptions in longitudinal clinical trials. *Pharmaceutical statistics* 2013; **12**(1): 1-6.
162. Permutt T. Sensitivity analysis for missing data in regulatory submissions. *Statistics in medicine* 2015.
163. White IR, Carpenter J, Horton NJ. Including all individuals is not enough: lessons for intention-to-treat analysis. *Clinical trials* 2012; **9**(4): 396-407.
164. Ratitch B, O'Kelly M, Tosiello R. Missing data in clinical trials: from clinical assumptions to statistical analysis using pattern mixture models. *Pharmaceutical statistics* 2013; **12**(6): 337-47.
165. Kistin CJ. Transparent reporting of missing outcome data in clinical trials: applying the general principles of CONSORT 2010. *Evidence-based medicine* 2014; **19**(5): 161-2.
166. Husereau D, Drummond M, Petrou S, et al. Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research* 2013; **16**(2): e1-5.
167. Husereau D, Drummond M, Petrou S, et al. Consolidated Health Economic Evaluation Reporting Standards (CHEERS)--explanation and elaboration: a report of the ISPOR Health Economic Evaluation Publication Guidelines Good Reporting Practices Task Force. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research* 2013; **16**(2): 231-50.
168. Hemming K, Hutton JL. Bayesian sensitivity models for missing covariates in the analysis of survival data. *Journal of evaluation in clinical practice* 2012; **18**(2): 238-46.
169. White IR, Carpenter J, Evans S, Schroter S. Eliciting and using expert opinions about dropout bias in randomized controlled trials. *Clinical trials* 2007; **4**(2): 125-39.

170. Roberts EM, English PB. Analysis of multiple-variable missing-not-at-random survey data for child lead surveillance using NHANES. *Statistics in medicine* 2016.
171. Zhu H, Ibrahim JG, Tang N. Bayesian Sensitivity Analysis of Statistical Models with Missing Data. *Statistica Sinica* 2014; **24**(2): 871-96.
172. Janssens M, Molenberghs G, Kerstens R. Handling of missing data in long-term clinical trials: a case study. *Pharmaceutical statistics* 2012; **11**(6): 442-8.
173. Keene ON, Roger JH, Hartley BF, Kenward MG. Missing data sensitivity analysis for recurrent event data using controlled imputation. *Pharmaceutical statistics* 2014; **13**(4): 258-64.
174. Rezvan PH, White IR, Lee KJ, Carlin JB, Simpson JA. Evaluation of a weighting approach for performing sensitivity analysis after multiple imputation. *BMC medical research methodology* 2015; **15**: 83.
175. Carpenter JR, Kenward MG, White IR. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Statistical methods in medical research* 2007; **16**(3): 259-75.
176. Yan X, Lee S, Li N. Missing data handling methods in medical device clinical trials. *Journal of biopharmaceutical statistics* 2009; **19**(6): 1085-98.
177. Scharfstein DO, Hogan J, Herman A. On the prevention and analysis of missing data in randomized clinical trials: the state of the art. *J Bone Joint Surg Am* 2012; **94 Suppl 1**: 80-4.
178. Ayele BT, Lipkovich I, Molenberghs G, Mallinckrodt CH. A multiple-imputation-based approach to sensitivity analyses and effectiveness assessments in longitudinal clinical trials. *Journal of biopharmaceutical statistics* 2014; **24**(2): 211-28.
179. Moreno-Betancur M, Chavance M. Sensitivity analysis of incomplete longitudinal data departing from the missing at random assumption: Methodology and application in a clinical trial with drop-outs. *Statistical methods in medical research* 2016; **25**(4): 1471-89.
180. Heraud-Bousquet V, Larsen C, Carpenter J, Desenclos JC, Le Strat Y. Practical considerations for sensitivity analysis after multiple imputation applied to epidemiological studies with incomplete data. *BMC medical research methodology* 2012; **12**: 73.
181. Groenwold RH, Donders AR, Roes KC, Harrell FE, Jr., Moons KG. Dealing with missing outcome data in randomized trials and observational studies. *American journal of epidemiology* 2012; **175**(3): 210-7.
182. Wittes J. Missing inaction: preventing missing outcome data in randomized clinical trials. *Journal of biopharmaceutical statistics* 2009; **19**(6): 957-68.
183. Wright CC, Sim J. Intention-to-treat approach to data from randomized controlled trials: a sensitivity analysis. *J Clin Epidemiol* 2003; **56**(9): 833-42.

Appendix 1: Database search strategy for the review of the literature in Chapter 2

This appendix shows the database search strategy used for the review looking at the current practice of handling, analysis and reporting of missing PROMs outcome data in RCTs. PubMed, EMBASE, Web of Science and NHS EED (for EQ-5D-3L and HUI only) were searched. Search terms used, and final number of papers identified from each search are shown below.

Of note, the tables below show the number of identified searches overall, while the PRISMA diagram only includes publications for the pre-specified years.

Table 1: PubMed search strategy

Search	Search Terms	Number of results
18	Search **(#1 AND #2 AND #10)**	2591
17	Search **(#1 AND #2 AND #9)**	225
16	Search **(#1 AND #2 AND #8)**	369
15	Search **(#1 AND #2 AND #7)**	3462
14	Search **(#1 AND #2 AND #6)**	1055
13	Search **(#1 AND #2 AND #5)**	354
12	Search **(#1 AND #2 AND #4)**	55
11	Search **(#1 AND #2 AND #3)**	1456
10	Search **(HUI OR "Health Utilities Index" OR "Health Utility Index")**	21074
9	Search **("PDQ-39" OR "PDQ 39" OR "PDQ39" OR "PDQ-8" OR "PDQ 8" OR "PDQ8" OR PDQ OR "Parkinson's Disease Questionnaire" OR "Parkinsons Disease Questionnaire" OR "Parkinson Disease Questionnaire")**	904
8	Search **("QLQ-C30" OR "QLQ C30" OR "QLQC30")**	1126
7	Search **("SF-36" OR "SF 36" OR "SF36")**	10501
6	Search **("SF-12" OR "SF 12" OR "SF12")**	2870
5	Search **(OHS OR "oxford hip score")**	3448
4	Search **(OKS OR "oxford knee score")**	398
3	Search **("EQ-5D" OR "EQ5D" OR "EQ 5D")**	3111
2	Search **(clinical* OR trial OR RCT)**	1480503
1	Search **random***	427289

Table 2: EMBASE search strategy

Search	Search Terms	Number of results
34	18 not 26	146
33	17 not 25	135
32	16 not 24	472
31	15 not 23	3170
30	14 not 22	426
29	13 not 21	82
28	12 not 20	61
27	11 not 19	813
26	18 and "Journal: Conference Abstract" [Publication Type]	29
25	17 and "Journal: Conference Abstract" [Publication Type]	91
24	16 and "Journal: Conference Abstract" [Publication Type]	300
23	15 and "Journal: Conference Abstract" [Publication Type]	880
22	14 and "Journal: Conference Abstract" [Publication Type]	137
21	13 and "Journal: Conference Abstract" [Publication Type]	17
20	12 and "Journal: Conference Abstract" [Publication Type]	4
19	11 and "Journal: Conference Abstract" [Publication Type]	363
18	1 and 2 and 10	175
17	1 and 2 and 9	226
16	1 and 2 and 8	772
15	1 and 2 and 7	4050
14	1 and 2 and 6	563
13	1 and 2 and 5	99
12	1 and 2 and 4	65
11	1 and 2 and 3	1176
10	(HUI or "Health Utilities Index" or "Health Utility Index").mp. [mp=title, abstract, subject headings, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword]	1757
9	("PDQ-39" or "PDQ 39" or "PDQ39" or "PDQ-8" or "PDQ 8" or "PDQ8" or PDQ or "Parkinson's Disease Questionnaire" or "Parkinsons Disease Questionnaire" or "Parkinson Disease Questionnaire").mp. [mp=title, abstract, subject headings, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword]	1272
8	("QLQ-C30" or "QLQ C30" or "QLQC30").mp. [mp=title, abstract, subject headings, heading word, drug trade name,	3257

	original title, device manufacturer, drug manufacturer, device trade name, keyword]	
7	("SF-36" or "SF 36" or "SF36").mp. [mp=title, abstract, subject headings, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword]	19675
6	("SF-12" or "SF 12" or "SF12").mp. [mp=title, abstract, subject headings, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword]	3445
5	(OHS or "oxford hip score").mp. [mp=title, abstract, subject headings, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword]	1321
4	(OKS or "oxford knee score").mp. [mp=title, abstract, subject headings, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword]	460
3	("EQ-5D" or "EQ5D" or "EQ 5D").mp. [mp=title, abstract, subject headings, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword]	5146
2	(clinical* or trial or RCT).mp. [mp=title, abstract, subject headings, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword]	4844039
1	random*.mp.	862922

Table 3: Web of Science search strategy

Search	Search Terms	Number of results
26	#10 AND #2 AND #1 Refined by: [excluding] DOCUMENT TYPES=(PROCEEDINGS PAPER) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, CCR-EXPANDED, IC Timespan=All years	110
25	#9 AND #2 AND #1 Refined by: [excluding] DOCUMENT TYPES=(MEETING ABSTRACT) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, CCR-EXPANDED, IC Timespan=All years	103
24	#8 AND #2 AND #1 Refined by: [excluding] DOCUMENT TYPES=(MEETING ABSTRACT OR PROCEEDINGS PAPER) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, CCR-EXPANDED, IC Timespan=All years	560
23	#7 AND #2 AND #1 Refined by: [excluding] DOCUMENT TYPES=(MEETING ABSTRACT OR PROCEEDINGS PAPER) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, CCR-EXPANDED, IC Timespan=All years	2,779
22	#6 AND #2 AND #1 Refined by: [excluding] DOCUMENT TYPES=(MEETING ABSTRACT OR PROCEEDINGS PAPER) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, CCR-EXPANDED, IC Timespan=All years	318
21	#5 AND #2 AND #1 Refined by: [excluding] DOCUMENT TYPES=(MEETING ABSTRACT) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, CCR-EXPANDED, IC Timespan=All years	65
20	#4 AND #2 AND #1 Refined by: [excluding] DOCUMENT TYPES=(PROCEEDINGS PAPER) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, CCR-EXPANDED, IC Timespan=All years	40
19	#3 AND #2 AND #1 Refined by: [excluding] DOCUMENT TYPES=(PROCEEDINGS PAPER OR MEETING ABSTRACT)	699

	Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, CCR-EXPANDED, IC Timespan=All years	
18	#10 AND #2 AND #1 Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, CCR-EXPANDED, IC Timespan=All years	118
17	#9 AND #2 AND #1 Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, CCR-EXPANDED, IC Timespan=All years	105
16	#8 AND #2 AND #1 Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, CCR-EXPANDED, IC Timespan=All years	606
15	#7 AND #2 AND #1 Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, CCR-EXPANDED, IC Timespan=All years	2,933
14	#6 AND #2 AND #1 Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, CCR-EXPANDED, IC Timespan=All years	330
13	#5 AND #2 AND #1 Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, CCR-EXPANDED, IC Timespan=All years	67
12	#4 AND #2 AND #1 Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, CCR-EXPANDED, IC Timespan=All years	41
11	#3 AND #2 AND #1 Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, CCR-EXPANDED, IC Timespan=All years	720
10	**TOPIC:** *(HUI OR "Health Utilities Index" OR "Health Utility Index")* Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, CCR-EXPANDED, IC Timespan=All years	1,952
9	**TOPIC:** *("PDQ-39" OR "PDQ 39" OR "PDQ39" OR "PDQ-8" OR "PDQ 8" OR "PDQ8" OR PDQ OR "Parkinson's Disease Questionnaire" OR "Parkinsons Disease Questionnaire" OR "Parkinson Disease Questionnaire")* Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, CCR-EXPANDED, IC Timespan=All years	824
8	**TOPIC:** *("QLQ-C30" OR "QLQ C30" OR "QLQC30")* Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, CCR-EXPANDED, IC Timespan=All years	2,274
7	**TOPIC:** *("SF-36" OR "SF 36" OR "SF36")*	15,325

	Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, CCR-EXPANDED, IC Timespan=All years	
6	**TOPIC:** *("SF-12" OR "SF 12" OR "SF12")* Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, CCR-EXPANDED, IC Timespan=All years	2,089
5	**TOPIC:** *(OHS OR "oxford hip score")* Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, CCR-EXPANDED, IC Timespan=All years	1,229
4	**TOPIC:** *(OKS OR "oxford knee score")* Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, CCR-EXPANDED, IC Timespan=All years	452
3	**TOPIC:** *("EQ-5D" OR "EQ5D" OR "EQ 5D")* Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, CCR-EXPANDED, IC Timespan=All years	3,176
2	**TOPIC:** *(clinical* OR trial OR RCT)* Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, CCR-EXPANDED, IC Timespan=All years	2,608,443
1	**TOPIC:** *(random*)* Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, CCR-EXPANDED, IC Timespan=All years	1,114,195

Table 4: NHS EED search strategy

2	(HUI) OR ("Health Utilities Index") OR ("Health Utility Index") IN NHSEED	139
1	("EQ-5D") OR ("EQ5D") OR ("EQ 5D") IN NHSEED	606

Appendix 2: Data extraction for the review of the literature in Chapter 2

This appendix lists the information that was extracted from each eligible paper, together with options for categorical data in (Table 5).

Table 5: List of items extracted from the identified articles

Variable name	Description	Possible answers
Pro	The paper was identified in the search for which PRO?	EQ-5D/ HUI/ OHS/ OKS/ PDQ-8/ PDQ-39/ QLQ-C30/ SF-12/ SF-36
Author	Lead author of the publication	
Year	Year of publication	2009-2013
Disease area	What is the main disease area/ characteristic by which the study population can be described?	
Type of analysis	The focus of the analysis is on clinical outcomes or health economics	Clinical/ HE
Primary/ secondary	Is the PRO a primary or secondary outcome in the context of this paper	Primary/ secondary
Primary outcome time point	What is the primary outcome time point for the relevant PRO? (all converted into month)	Time in months
Maximum follow-up	Until what time from randomisation is the relevant PRO measured to? (all converted into month)	Time in months
FUP assessment time points	Are follow-up data measured repeatedly, or only once?	Single/ repeated
Strategies mentioned to avoid missing data in PRO	Are any steps mentioned in the article to reduce the occurrence in missing data in the relevant PRO measure?	Yes/ no
Details of strategy	Details of the strategy to reduce the occurrence in missing data in the relevant PRO measure (if appropriate)	Free text
Total N randomised	Total number of participants randomised into the trial (to the two arms used in this review for trials with more than two arms)	N
Randomised into active arm	Number of participants randomised into the active comparison arm	
Randomised into control arm	Number of participants randomised into the control arm of the study	

Variable name	Description	Possible answers
Is it clear how much data is available at follow-up	Is it clear how much data for the relevant PRO is available at the primary follow-up times point?	Yes/ no
Total N with data collected at follow-up	Total number of participants with available PRO data at the primary follow-up time point	
Active participants with data collected at follow-up	Number of participants in the active treatment arm with available PRO data at the primary follow-up time point	
Control participants with data collected at follow-up	Number of participants in the control treatment arm with available PRO data at the primary follow-up time point	
Reasons for missing data given	Have reasons been provided for why data is missing, beyond withdrawals and death?	
Distinction between complete and partial missingness	Does the article distinguish between complete missingness, and missing items within a partly completed questionnaire?	Yes/ no
Differential missingness assessed	Has differential missingness been assessed, i.e. has the baseline (or other) data been compared between those with complete and missing outcome data?	Yes/ no
Analysis population	Which is the primary analysis population for the PRO?	Modified ITT/ ITT/ PP/ unclear
Missing data mentioned in methods section	Has missing data been mentioned in the method or analysis section	Yes/ no
Assumed missing data mechanism	What details are provided on the assumed missing data mechanism?	MCAR/ MAR/ MNAR/ not described
Primary imputation method	What is the primary imputation methods used in the analysis?	CC/ LOCF/ mean imputation/ repeated measures/ MI/ unclear
Justification for imputation method	Is any justification provided in the article regarding the chosen primary imputation method?	Yes/ no
Sensitivity analysis done	Was any sensitivity analysis performed in the context of the PRO analysis?	Yes/ no
Sensitivity analysis assumption	Assumption about the missing data mechanism in the sensitivity analysis	MCAR/ MAR/ MNAR/ not described/ N/A

Variable name	Description	Possible answers
Sensitivity analysis detail	What did sensitivity analysis include	
Potential influence of missing data mentioned	Has the potential impact of the missing data on the trial results been mentioned?	Yes/ no

Appendix 3: Details of keywords used in literature review (Chapter 2)

This appendix details the terms used in the keyword search of articles. As a minimum, title, abstract and methods sections were read in full; CONSORT diagrams and tables (where available) were considered to identify available data at the relevant analysis time points. Key word searchers were used to identify other relevant information, and the terms used for this are shown in Figure 1.

Identify primary endpoint as assessment schedule

- Primary / outcome/ endpoint
- Month/year/ week

Check for info on reducing the amount of missing data

- Phone
- Letter/ (e)mail
- Post
- Sent/ send
- home
- Remind(er)
- Minimise/ minimize – maximise/ze
- Limit
- Avoid
- Reduc(e) – increase(e)
- Missing
- Responders
- Complete/ Incomplete
- Available
- Loss/ lost
- Follow
- Compliance/comply
- Withdraw(al) / attrition

Information on missingness: (also search tables and CONSORT)

- Baseline
- Differ
- Similar
- Assume/ assumption
- Item
- question

Assess sensitivity analysis in paper:

- Intention (to treat)
- Imput(ation), replace
- dropout
- Sensitivity
- secondary

Figure 1: T**erms used for the key word search of identified articles**

Appendix 4: Relevant CRFs from the KAT study

This appendix includes extracts from the KAT case report forms (CRFs) relevant to the analyses presented in Chapter 3, and also used in the simulation studies presented in Chapters 4 and 5.

A sample copy of the EQ-5D-3L, including the copyright statement as per EurpoQoL guidelines, is also reproduced. Information on the numbering of items for the OKS and SF-12, as referred to in the chapter, is added in red.

kat

Knee Arthroplasty Trial

Study Centre No

Patient Study No

CONFIDENTIAL

KAT STUDY

PARTICIPANT DETAILS FORM

For completion by local KAT researcher

This study is funded by the NHS Research and Development Health Technology Assessment Programme.

PERSONAL INFORMATION
Please check faxed patient information and write amendments below, if necessary.

Title *(Mr, Mrs etc)* **Surname**

First Names

Date of Birth Day Month Year

House Name

House Number

Street Name

District

Town/City

County

Postcode

Telephone No *(including code)*

Place of Birth *(including county)*

Marital Status Single Married Divorced Widowed

Sex Male Female

Maiden name if female and ever married

NHS Number

Hospital Number *(if known)*

CHI Number *(Scotland only)*

DESCRIPTIVE INFORMATION ABOUT THE PARTICIPANT

1. Weight ☐☐☐ **kgs**

2. Height ☐☐☐ **cms**

3. Type of knee arthritis? Osteoarthritis ☐ (Cross X one box)

Rheumatoid ☐

4. Is the arthritis in? Single knee ☐

Both knees ☐

General ☐

5. Are there any other conditions which are affecting the participant's mobility? No ☐

Yes ☐

If **Yes**, please specify _____

6. Has the participant had any previous knee surgery? No ☐

Yes ☐

If **Yes**, was it

Ipsilateral osteotomy ☐

Ipsilateral patellectomy ☐

Contralateral previous knee replacement ☐

Other ☐

If Other, please specify _____

Please send this completed form, together with the participant entry questionnaire, the surgeon's form, and the hospital care form, to the KAT co-ordinating office, when the participant leaves hospital. Prepaid envelopes are provided.

The EQ-5D-3L

By placing a tick in one box in each group below, please indicate which statements best describe your own health state today.

Mobility

I have no problems in walking about	❑
I have some problems in walking about	❑
I am confined to bed	❑

Self-Care

I have no problems with self-care	❑
I have some problems washing or dressing myself	❑
I am unable to wash or dress myself	❑

Usual Activities *(e.g. work, study, housework, family or leisure activities)*

I have no problems with performing my usual activities	❑
I have some problems with performing my usual activities	❑
I am unable to perform my usual activities	❑

Pain / Discomfort

I have no pain or discomfort	❑
I have moderate pain or discomfort	❑
I have extreme pain or discomfort	❑

Anxiety / Depression

I am not anxious or depressed	❑
I am moderately anxious or depressed	❑
I am extremely anxious or depressed	❑

The SF-12

YOUR GENERAL HEALTH

The following questions ask for your views about your health, how you feel and how well you are able to do your usual activities.

If you are unsure about how to answer any questions please give the best answer you can and make any of your own comments if you like. Do not spend too much time in answering as your immediate response is likely to be the most accurate.

2. In general, would you say your health is: *(Please X one box)*

SF-12 item 1

Excellent	Very Good	Good	Fair	Poor
☐	☐	☐	☐	☐

3. The following questions are about activities you might do during a typical day. Does <u>your health now limit you</u> in these activities? If so, how much? *(Please X one box on each line)*

		Yes, limited a lot	Yes, limited a little	No, not limited at all
SF-12 item 2a	**Moderate activities**, such as moving a table, pushing a vacuum cleaner, bowling or playing golf	☐	☐	☐
SF-12 item 2b	Climbing **several** flights of stairs	☐	☐	☐

4. During the <u>past 4 weeks</u>, have you had any of the following problems with your work or other regular daily activities <u>as a result of your physical health?</u> *(Please X one box on each line)*

		All of the time	Most of the time	Some of the time	A little of the time	None of the time
SF-12 item 3a	**Accomplished less** than you would like	☐	☐	☐	☐	☐
SF-12 item 3b	Were you limited in the **kind** of work or other activities	☐	☐	☐	☐	☐

Study Centre No ☐☐

Patient Study No ☐☐☐☐☐

5. **During the <u>past 4 weeks</u>, have you had any of the following problems with your work or other regular daily activities <u>as a result of any emotional problems</u> (such as feeling depressed or anxious)?** *(Please X one box on each line)*

		All of the time	Most of the time	Some of the time	A little of the time	None of the time
SF-12 item 4a	**Accomplished less** than you would like	☐	☐	☐	☐	☐
SF-12 item 4b	Didn't do work or other activities as **carefully** as usual	☐	☐	☐	☐	☐

6. **During the <u>past 4 weeks</u> how much did <u>pain</u> interfere with your normal work (including work both outside the home and housework)?** *(Please X one box)*

	Not at all	A little bit	Moderately	Quite a bit	Extremely
SF-12 item 5	☐	☐	☐	☐	☐

7. **These questions are about how you feel and how things have been with you during the <u>past 4 weeks</u>. For each question, please give the one answer that comes closest to the way you have been feeling.** *(Please X one box on each line)*

How much time during the <u>past 4 weeks</u>:

		All of the time	Most of the time	Some of the time	A little of the time	None of the time
SF-12 item 6a	Have you felt calm and peaceful?	☐	☐	☐	☐	☐
SF-12 item 6a	Did you have a lot of energy?	☐	☐	☐	☐	☐
SF-12 item 6a	Have you felt downhearted and low?	☐	☐	☐	☐	☐

8. **During the <u>past 4 weeks</u>, how much of the time has your physical health or emotional problems interfered with your social activities (like visiting with friends, relatives, etc)?** *(Please X one box)*

	All of the time	Most of the time	Some of the time	A little of the time	None of the time
SF-12 item 7	☐	☐	☐	☐	☐

The OKS

HEALTH PROBLEMS CAUSED BY YOUR KNEE

The following questions ask about problems which may have been caused by your right knee during the **past 4 weeks**. (Please X one box for each question.)

9. During the past 4 weeks how would you describe the pain you have from your right knee?

OKS Q1

None	Very mild	Mild	Moderate	Severe
☐	☐	☐	☐	☐

10. During the past 4 weeks have you had any trouble with washing and drying yourself (all over) because of your right knee?

OKS Q2

No trouble at all	Very little trouble	Moderate trouble	Extreme difficulty	Impossible to do
☐	☐	☐	☐	☐

11. During the past 4 weeks have you had any trouble getting in and out of a car or using public transport because of your right knee? (whichever you tend to use).

OKS Q3

No trouble at all	Very little trouble	Moderate trouble	Extreme difficulty	Impossible to do
☐	☐	☐	☐	☐

12. During the past 4 weeks for how long have you been able to walk before the pain from your right knee becomes severe? (with or without a stick).

OKS Q4

No pain at all, or no pain for more than 30 mins	16 to 30 mins	5 to 15 mins	Around the house only	Not at all - pain severe on walking
☐	☐	☐	☐	☐

13. During the past 4 weeks after a meal (sat at a table), how painful has it been for you to stand up from a chair because of your right knee?

OKS Q5

Not at all painful	Slightly painful	Moderately painful	Very painful	Unbearable
☐	☐	☐	☐	☐

5

14. During the <u>past 4 weeks</u> have you been limping when walking, <u>because of your right knee</u>?

OKS Q6

Rarely/ never	Sometimes or just at first	Often, not just at first	Most of the time	All of the time
☐	☐	☐	☐	☐

15. During the <u>past 4 weeks</u> could you kneel down and get up again afterwards? (thinking of your right knee)

OKS Q7

Yes, easily	With little difficulty	With moderate difficulty	With extreme difficulty	No, impossible
☐	☐	☐	☐	☐

16. During the <u>past 4 weeks</u> have you been troubled by <u>pain from your right knee</u> in bed at night?

OKS Q8

No nights	Only 1 or 2 nights	Some nights	Most nights	Every night
☐	☐	☐	☐	☐

17. During the <u>past 4 weeks</u> how much has <u>pain from your right knee</u> interfered with your usual work (including housework)?

OKS Q9

Not at all	A little bit	Moderately	Greatly	Totally
☐	☐	☐	☐	☐

18. During the <u>past 4 weeks</u> have you felt that your right knee might suddenly 'give way' or let you down?

OKS Q10

Rarely/ never	Sometimes, or just at first	Often, not just at first	Most of the time	All of the time
☐	☐	☐	☐	☐

19. During the <u>past 4 weeks</u> could you do the household shopping <u>on your own</u>? (thinking of your knee)

OKS Q11

Yes, easily	With little difficulty	With moderate difficulty	With extreme difficulty	No, impossible
☐	☐	☐	☐	☐

20. During the <u>past 4 weeks</u> could you walk down one flight of stairs? (thinking of your knee)

OKS Q12

Yes, easily	With little difficulty	With moderate difficulty	With extreme difficulty	No, impossible
☐	☐	☐	☐	☐

Thank you again!

Information collected during the surgery

Study Centre No

kat

Knee Arthroplasty Trial

Patient Study No

Patient Name

Date of birth Day Month Year

Date of operation Day Month Year

Confirmation of treatment allocated by randomisation (mark each line with a cross X)

A: Metal backed	☐	**Non-metal backed**	☐	**Not randomised**	☐
B: Patellar Resurfacing	☐	**No patellar resurfacing**	☐	**Not randomised**	☐
C: Mobile bearing	☐	**Fixed bearing**	☐	**Not randomised**	☐
D: Uni-compartmental arthroplasty	☐	**Total knee replacement**	☐	**Not randomised**	☐

Please attach stickers for all components used for the operation or fill in the appropriate information:

CAT. NO. _____
Component _____

CAT. NO. _____
Component _____

CAT. NO. _____
Component _____

CAT. NO. _____
Component _____

RIGHT KNEE

PTO FOR SURGEON'S FORM

kat
Knee Arthroplasty Trial

SURGEON'S FORM

Please complete straight after the operation by marking boxes with a cross (X) or giving details as requested

Pre-operative

The study knee is the right knee

Fixed flexion deformity No ☐ Yes ☐ → How many degrees? ☐☐°

Valgus/Varus deformity No ☐ Yes ☐ → Varus ☐ Valgus ☐

Was the deformity? Mild ☐ Moderate ☐ Severe ☐

Was it correctable? No ☐ Yes ☐

Intra-operative

Patella

Normal ☐ Partial loss of cartilage ☐ Full thickness cartilage loss ☐ Less than 5mm bone loss ☐ More than 5mm bone loss ☐

Anterior cruciate ligament Intact ☐ Damaged ☐ Absent ☐

Posterior cruciate ligament Intact ☐ Recessed or damaged ☐ Divided ☐

Lateral retinacular release No ☐ Yes ☐

Was cement used for:

Tibia No ☐ Yes ☐ **Femur** No ☐ Yes ☐ **Patella** No ☐ Yes ☐

Intra-operative complications? No ☐ Yes ☐

If 'yes', Patella fracture ☐ Other ☐

If other, please specify

Was the consultant's usual surgical technique followed? No ☐ Yes ☐

If 'no', please give details

Did the patient receive the allocated procedure? No ☐ Yes ☐

If 'no', please give details

At end of operation

Fixed flexion deformity No ☐ Yes ☐ → How many degrees? ☐☐°

Name of surgeon performing operation

Grade of operator

Consultant ☐ Associate specialist/staff grade ☐ SPR ☐ SHO ☐

Grade of senior surgeon present

Consultant ☐ Associate specialist/staff grade ☐ SPR ☐

RIGHT KNEE **PTO FOR THEATRE FORM**

The hospital readmission form

Kat
Knee Arthroplasty Trial

THE KAT STUDY
HOSPITAL RE-ADMISSION FORM

Patient Study No

The re-admission is related to:

Knee replacement ☐ **Surgery in general** ☐ **Not related** ☐ **Other** ☐

First name ☐☐☐☐☐☐☐☐☐☐ **Surname** ☐☐☐☐☐☐☐☐☐☐

Hospital ☐

Date of Admission **Day** ☐☐ / **Month** ☐☐ / **Year** ☐☐☐☐

Date of Discharge **Day** ☐☐ / **Month** ☐☐ / **Year** ☐☐☐☐

Discharged home ☐
Transferred to a rehabilitation unit ☐ Name of unit ☐
Died ☐
Other ☐

If Other, please specify

☐

Reason for Readmission *(Please cross as appropriate)*

Surgical ☐ **Medical** ☐ **Other** ☐

Please give details below

☐

Please continue over the page.

This study is funded by the NHS Research and Development Health Technology Assessment Programme.

PTO

Did the patient require any further knee surgery? **NO** ☐ **YES** ☐

If Yes, what?

Manipulation under anaesthesia	☐	Washout	☐
Debridement	☐	Aspiration	☐
1st stage revision	☐	2nd stage revision	☐
Revision	☐	1st patella resurfacing	☐
Above knee amputation	☐	Other	☐

If Other, please specify (Eg excision arthroplasty, arthrodesis or internal fixation of fractures)

If revision procedure or 2nd stage revision procedure please specify which component

Patella revision ☐ **Tibial revision** ☐ **Femur revision** ☐

What was the reason for revision?

Infection ☐
Loosening ☐
Pain ☐
Mechanical failure/fracture ☐
Dislocation ☐
Other ☐

If Other, please specify

Please attach stickers for all components used for the operation or fill in the appropriate information and attach a photocopy of the participant's operation notes!

Cat No .....
Component .....

Cat No .....
Component .....

Cat No .....
Component .....

Cat No .....
Component .....

Appendix 5: Relevant CRFs from the PD MED study

This appendix includes copies of the CRFs from the PD MED trial that are relevant to the analyses presented in Chapter 3.

Appendix H

RANDOMISATION NOTEPAD

Prepare for the randomisation questions by filling in sections A, B, C, D, E and F on this pad before telephoning the toll free randomisation service on 0800 953 0274 for immediate randomisation, or fax form to 0121 415 9135 for allocation by next working day.

PART A: IDENTIFYING DETAILS

Randomising Consultant…….. Hospital Name……..

Patient's Full name:…… Gender: Male ☐ Female ☐ Title: Mr/Mrs/Ms/Miss/Other……..

PART B: ELIGIBILITY

Is the patient demented? No ☐ Yes ☐

Is the patient able to complete the questionnaire? No ☐ Yes ☐ *(with help, if necessary)*

Has the patient given written informed consent? No ☐ Yes ☐

PART C: PATIENT'S MEDICAL DETAILS

Date of initial diagnosis of PD (month/year) …../….. Yoehn & Yahr Stage ……. (see protocol, appendix A)

Stage of PD Early ☐ Later ☐

If Early: Any previous PD therapy? No ☐ <1 month ☐ 1 – 3 months ☐ 3 – 6 months ☐ > 6 months ☐

If previous therapy, please specify……..

If later: Patient previously entered in PD MED trial? No ☐ Yes ☐ if yes, PD MED trial number…….

Current therapy: DA: Yes ☐ No ☐ MAOBI: Yes ☐ No ☐ COMTI: Yes ☐ No ☐

PART D: TREATMENT DETAILS

Willing to randomise to MAOBI: No ☐ Yes ☐ Willing to randomise to LD alone (early PD only): No ☐ Yes ☐

If allocated DA, which DA will be prescribed?.....

If allocated MAOBI, which MAOBI will be prescribed?.....

If allocated COMTI, which COMTI will be prescribed?..... (later PD only)

PART E: QUESTIONNAIRES

Has the patient completed? PDQ39: No ☐ Yes ☐ Euroqol EQ-5D: No ☐ Yes ☐

Has the MMSE been administrated? No ☐ Yes ☐

PART F: CARER DETAILS

Does the patient have a regular carer? No ☐ Yes ☐ If yes, name of principal carer……..

Has the carer completed the SF-36? No ☐ Yes ☐ Relationship……

If No, reason (eg no carer, carer declined to take part)……

PART G: TREATMENT ALLOCATION from RANDOMISATION SERVICE 0800 953 0274

Early PD LD only ☐ Dopamine agonist ☐ MAOB inhibitor ☐

Later PD Dopamine agonist ☐ MAOB inhibitor ☐ COMT inhibitor ☐

PD MED trial number:……..

Contact person…….. Fax No: …… Telephone No:……..

(for queries or fax allocations)

Please file the top copy of this form in the patient notes, and return the bottom copy along with the questionnaires listed in Part E (and F if applicable) and consent form within one week of trial entry to the PD MED Trial Office. A Freepost envelope is supplied for return to The University of Birmingham, Birmingham Clinical Trials Unit, Division of Medical Sciences, Robert Aitken Institute, **FREEPOST RRKR-JUZR-HZHG**, Birmingham B15 2TT

PD MED Randomisation Notepad Version 8 Aug 2010

21

ANNUAL FOLLOW-UP FORM
EARLY DISEASE

Appendix K

Part A: Identification Details

To be completed by patient's hospital doctors

Patient's initials: ☐☐ PD MED Trial No E☐☐☐☐☐

Date of birth: / / Hospital number: _____

Hospital _____

Is the diagnosis still idiopathic Parkinson's Disease? No ☐ Yes ☐

If not, what is the most likely diagnosis? _____

N.B. The patient will still be followed up within PD MED

Part B: Current Disease Status

Date of assessment: ____/____/____

Patient's current Hoehn & Yahr stage: ☐

Hoehn and Yahr Stages

Stage 1.0 Unilateral involvement only
Stage 1.5 Unilateral and axial involvement
Stage 2.0 Bilateral involvement without impairment of balance
Stage 2.5 Mild bilateral involvement with recovery on retropulsion (pull) test
Stage 3.0 Mild to moderate bilateral involvement, some postural instability but physically independent
Stage 4.0 Severe disability, still able to walk and to stand unassisted
Stage 5.0 Wheelchair bound or bedridden unless aided.

Please ask the patient if they have suffered (a) any involuntary movements, other than tremor, and demonstrate typical athetoid dyskinesia to them or (b) wearing off of one dose of medication before the next is due. If the reply is affirmative, or if you or the carer have witnessed these phenomena, please record the findings below.

Has the patient developed motor complications? No ☐ Yes ☐

What type of motor complications have developed?

Dyskinesia No ☐ Yes ☐ **If Yes**, date started (mo/yr): ____/____

Fluctuations No ☐ Yes ☐ **If Yes**, date started (mo/yr): ____/____

Has the patient developed dementia? No ☐ Yes ☐ **If Yes**, date of diagnosis (mo/yr) ____/____
(as defined by the clinician's usual criteria)

Has the patient been institutionalised? No ☐ Yes ☐ **If Yes**, date admitted (mo/yr) ____/____

Type of home: Nursing ☐ Residential ☐

Has the patient died? No ☐ Yes ☐ **If Yes**, date of death: ____/____/____

Cause of death: _____

If the patient has died, please give details of therapy prior to death in Part C.

Part C: Current Therapy

Please give details of the patient's current drug therapy for PD including treatment related to PD
(e.g. anti-depressants, anti-psychotic):

Example of patients current drug therapy

Drug	Dose	Total daily dose (mg)	Date Started
Sinemet Plus	100mg x 5 daily	500	5/10/00
Bromocriptine	10mg + 5mg + 10mg	25	25/5/99

Drug	**Dose**	**Total daily dose (mg)**	**Date Started**
			/ /
			/ /
			/ /
			/ /

If the medication has changed since the last follow-up, **please record the changes and reasons on the reverse side to this form**

Assessor: _____ Signature: _____ Date: / /

Please return this form to: **PD MED Trial Office, The University of Birmingham, Birmingham Clinical Trials Unit, Robert Aitken Institute, FREEPOST RRKR-JUZR-HZHG, Birmingham, B15 2TT**

Year of follow-up

1 2 3 4 5

DRUG	Date Started approximate	Date Stopped approximate	Reason stopped or changed (please specify side effects)

If appropriate, please send a copy of GP letter

PDQ-39 QUESTIONNAIRE

Please complete the following

Please tick one box for each question

	Due to having Parkinson's disease, how often during the last month have you....	**Never**	**Occasionally**	**Sometimes**	**Often**	**Always** or cannot do at all
1	Had difficulty doing the leisure activities which you would like to do?	☐	☐	☐	☐	☐
2	Had difficulty looking after your home, e.g. DIY, housework, cooking?	☐	☐	☐	☐	☐
3	Had difficulty carrying bags of shopping?	☐	☐	☐	☐	☐
4	Had problems walking half a mile?	☐	☐	☐	☐	☐
5	Had problems walking 100 yards?	☐	☐	☐	☐	☐
6	Had problems getting around the house as easily as you would like?	☐	☐	☐	☐	☐
7	Had difficulty getting around in public?	☐	☐	☐	☐	☐
8	Needed someone else to accompany you when you went out?	☐	☐	☐	☐	☐
9	Felt frightened or worried about falling over in public?	☐	☐	☐	☐	☐
10	Been confined to the house more than you would like?	☐	☐	☐	☐	☐
11	Had difficulty washing yourself?	☐	☐	☐	☐	☐
12	Had difficulty dressing yourself?	☐	☐	☐	☐	☐
13	Had problems doing up buttons or shoe laces?	☐	☐	☐	☐	☐

*Please check that you have ticked **one box for each question** before going on to the next page*

Due to having Parkinson's disease, how often <u>during the last month</u> have you....		***Please tick <u>one</u> box for each question***				
		Never	**Occasionally**	**Sometimes**	**Often**	**Always** or cannot do at all
14	Had problems writing clearly?	☐	☐	☐	☐	☐
15	Had difficulty cutting up your food?	☐	☐	☐	☐	☐
16	Had difficulty holding a drink without spilling it?	☐	☐	☐	☐	☐
17	Felt depressed?	☐	☐	☐	☐	☐
18	Felt isolated and lonely?	☐	☐	☐	☐	☐
19	Felt weepy or tearful?	☐	☐	☐	☐	☐
20	Felt angry or bitter?	☐	☐	☐	☐	☐
21	Felt anxious?	☐	☐	☐	☐	☐
22	Felt worried about your future?	☐	☐	☐	☐	☐
23	Felt you had to conceal your Parkinson's from people?	☐	☐	☐	☐	☐
24	Avoided situations which involve eating or drinking in public?	☐	☐	☐	☐	☐
25	Felt embarrassed in public due to having Parkinson's disease?	☐	☐	☐	☐	☐
26	Felt worried by other people's reaction to you?	☐	☐	☐	☐	☐
27	Had problems with your close personal relationships?	☐	☐	☐	☐	☐
28	Lacked support in the ways you need from your spouse or partner?	☐	☐	☐	☐	☐
	If you do not have a spouse or partner tick here		☐			
29	Lacked support in the ways you need from your family or close friends?	☐	☐	☐	☐	☐

*Please check that you have ticked **one box for each question** before going on to the next page*

Due to having Parkinson's disease, how often <u>during the last month</u> have you....		***Please tick <u>one</u> box for each question***				
		Never	**Occasionally**	**Sometimes**	**Often**	**Always or cannot do at all**
30	Unexpectedly fallen asleep during the day?	☐	☐	☐	☐	☐
31	Had problems with your concentration, e.g. when reading or watching TV?	☐	☐	☐	☐	☐
32	Felt your memory was bad?	☐	☐	☐	☐	☐
33	Had distressing dreams or hallucinations?	☐	☐	☐	☐	☐
34	Had difficulty with your speech?	☐	☐	☐	☐	☐
35	Felt unable to communicate with people properly?	☐	☐	☐	☐	☐
36	Felt ignored by people?	☐	☐	☐	☐	☐
37	Had painful muscle cramps or spasms?	☐	☐	☐	☐	☐
38	Had aches and pains in your joints or body?	☐	☐	☐	☐	☐
39	Felt unpleasantly hot or cold?	☐	☐	☐	☐	☐

*Please check that you have ticked **one box for each question** before going on to the next page*

Thank you for completing the PDQ 39 questionnaire

EuroQoL EQ-5D

Appendix F

Please answer the questions by ticking one box in each group.

Please indicate which statement best describes your own health today.

Do not tick more than one box in each group.

1 **Mobility**

- I have no problems walking about ☐
- I have some problems in walking about ☐
- I am confined to bed ☐

2 **Self care**

- I have no problems with self-care ☐
- I have some problems washing or dressing myself ☐
- I am unable to wash or dress myself ☐

3 **Usual activities** (e.g. work, study, housework, family or leisure activities)

- I have no problems with performing my usual activities ☐
- I have some problems with performing my usual activities ☐
- I am unable to perform my usual activities ☐

4 **Pain / Discomfort**

- I have no pain or discomfort ☐
- I have moderate pain or discomfort ☐
- I have extreme pain or discomfort ☐

5 **Anxiety/ Depression**

- I am not anxious or depressed ☐
- I am moderately anxious or depressed ☐
- I am extremely anxious or depressed ☐

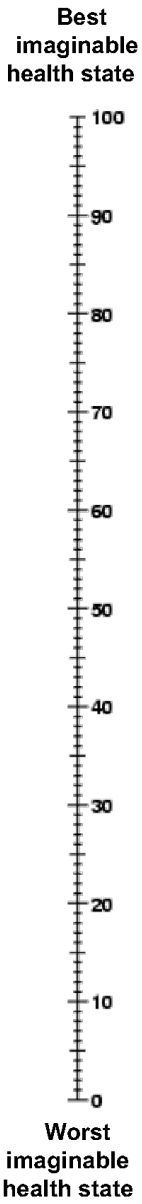
Please continue overleaf

Your own health state today

To help people say how good or bad their health state is, we have drawn a scale (rather like a thermometer) on which the best state you can imagine is marked by 100 and the worst state you could imagine is marked by 0

We would like you to indicate on the scale how good or bad your own health is today, in your opinion. Please do this by drawing a line from the box to whichever point on the scale indicates how good or bad your current health state is.

Your own health state today



Please complete the following

PD MED Trial Number

Patient Initials: .....

Date of Birth: ..... / ..... / .....

Date Completed: ..... / ..... / .....

Trial office use only

Date Sent: / /

Date Received: ..... / ..... / .....

Date Entered: ..... / ..... / .....

Baseline 6 month 1yr 2yr 3yr 4yr 5yr

(please circle as appropriate)

Appendix 6: Relevant CRFs from the PD SURG study

This appendix contains extracts from the PD SURG trial CRFs that are relevant to the analyses presented in Chapter 3. The PDQ-39 and the EQ-5D-3L are shown in Appendix 5 and are not reproduced here.

APPENDIX D: RANDOMISATION NOTEPAD

Prepare for the randomisation questions by filling in sections A, B C and D on this pad before telephoning the toll free randomisation service on 0800 953 0274 for immediate randomisation, or fax form to 0121 415 9135 for allocation by next working day.

PART A: IDENTIFYING DETAILS

Randomising consultant: ..... Hospital name: .....

Patient's full name: ..... Sex: Male ☐ Female ☐

Date of birth: ...../...../..... Hospital number: .....

PART B: PATIENT'S MEDICAL DETAILS

Date of initial diagnosis of PD (mo/yr): ...../..... Hoehn & Yahr stage: .....

Has the patient previously received?

Dopamine agonist No ☐ Yes ☐

Selegiline No ☐ Yes ☐

COMT inhibitor No ☐ Yes ☐

Apomorphine No ☐ Yes ☐

Reason for considering surgery (more than one box may be ticked):

Tremor ☐ Dyskinesia ☐ Severe "off" periods" ☐ Other, specify: .....

PART C: PLANNED TREATMENT

If allocated to surgery, which operation will be performed?

STN Stimulation ☐ STN Lesion ☐

GPi Stimulation ☐ GPi Lesion ☐

If allocated to medical therapy, will apomorphine be prescribed? No ☐ Yes ☐

PART D: QUESTIONNAIRES These must be completed prior to randomisation.

Has the patient completed the following:

PDQ-39 No ☐ Yes ☐ EuroQol EQ-5D No ☐ Yes ☐

UPDRS performed in ON state: No ☐ Yes ☐ In OFF state: No ☐ Yes ☐

Has DRS2 been performed: No ☐ Yes ☐

Has the patient given written informed consent? No ☐ Yes ☐

PART E: TREATMENT ALLOCATION

Treatment: Immediate Surgery ☐ Medical Management ☐

PD SURG trial number: S☐☐☐

PART F: CARER DETAILS

Does the patient have a regular carer? No ☐ Yes ☐

If yes, name of principal carer: .....

Has the carer given written informed consent? No ☐ Yes ☐

Has the carer completed the SF36? No ☐ Yes ☐

APPENDIX I: HOEHN & YAHR STAGING SYSTEM

Stage 0	No signs of Parkinson's disease
Stage 1.0	Unilateral involvement only
Stage 1.5	Unilateral and axial involvement
Stage 2.0	Bilateral involvement without impairment of balance
Stage 2.5	Mild bilateral involvement with recovery on retropulsion (pull) test
Stage 3.0	Mild to moderate bilateral involvement, some postural instability but physically independent
Stage 4.0	Severe disability, still able to walk and to stand unassisted
Stage 5.0	Wheelchair bound or bedridden unless aided

APPENDIX L: POST-OPERATIVE FORM

Surgeon ..... Date of Admission .....

Date of Surgery ..... Date of Discharge .....

PART A: PRE-OPERATIVE ASSESSMENT

Number of clinic visits prior to surgery .....

Please identify the staff who were present at these visits to assess the patient *(please tick all who were present)*

Neurosurgeon ☐ Neurologist ☐ PD Nurse ☐

Neuropsychologist ☐ Neuropsychology assistant ☐ Other ☐ *specify* .....

Did the patient stay in hospital for any pre-operative assessment Yes ☐ No ☐

If *yes*, how many days hospital stay was the pre-operative assessment ..... days

Please state how many tests/ procedures were used/ carried out during the pre-operative assessment

EMG ..... CT scan ..... MRI scan ..... CT-MRI fusion ..... X-ray .....

PART B: THEATRE

Was the theatre dedicated to the PD surgery for the entire day? Yes ☐ No ☐

If *no*, please state the duration of theatre time required for the preparation, procedure and recovery ..... hours ..... minutes

Did you use a planning station for the planning of the operation? Yes ☐ No ☐

If *yes*, which type? Stealth ☐ Radionics ☐ Brain Lab ☐ Zeiss ☐

Please state the number of each of the following staff present in theatre for preparation, procedure and recovery

Surgeon ..... Neurosurgical technician ..... Anaesthetist .....

Registrar ..... Theatre Nurse ..... PD Nurse ..... Anaesthetist Assistant .....

ODA ..... Electrophysiologist ..... Electrophysiology Technician .....

Other, please state grade and number .....

Was a robotic arm used in this procedure? Yes ☐ No ☐

What type of stereotactic frame did you use?

CRW (Radionics) ☐ Lecksell-G (Electra) ☐ Leibenger ☐ Other ☐ specify .....

PART C: PROCEDURE

Note: if staged procedure, a form needs completing for each procedure

Bilateral simultaneous ☐ Bilateral staged ☐ If staged: First ☐ Second ☐

Unilateral ☐

Target STN ☐ GPi ☐ **Technique** Left Stimulation ☐ Lesion ☐

Right Stimulation ☐ Lesion ☐

Number of tracts Left ..... Right .....

Surgery abandoned No ☐ Yes ☐ If yes, state reason: .....

PART D: LOCALISATION

Localisation during surgery; please state number of times each test used

CT scan ..... MRI scan ..... Visual field test ..... Ventriculography .....

ECG ..... X-ray ..... Other radiological (state) .....

Externalisation of electrode for test period prior to implantation of pulse generator

Microelectrode ..... Semi microelectrode ..... Impedance .....

Microstimulation ..... Macrostimulation .....

Other electrophysiological (state) .....

PART E: STIMULATOR MODEL NUMBERS

	Number used	Make and odel number (if known) or use stickers
Implantable pulse generator	.....	.....
DBS Electrode	.....	.....
Extension Lead	.....	.....
Therapy Controller	.....	.....
Accessory Kit	.....	.....

PART F: POST OPERATIVE MANAGEMENT

Please state the number of times each test was used in the days following surgery:

CT scan ..... MRI scan ..... Visual field test .....

ECG ..... X-ray ..... Other radiological (state) .....

Please note MRI images may be reviewed centrally.

Following surgery and pre-discharge, which staff been involved in turning stimulator on, testing electrode, adjusting voltage etc. Approximately how long in total have they been involved?

Neurosurgeon	Yes ☐	No ☐	..... hrs	..... mins
Neurosurgical technician	Yes ☐	No ☐	..... hrs	..... mins
Neurologist	Yes ☐	No ☐	..... hrs	..... mins
Registrar	Yes ☐	No ☐	..... hrs	..... mins
PD Nurse	Yes ☐	No ☐	..... hrs	..... mins
Theatre nurse	Yes ☐	No ☐	..... hrs	..... mins

PART F: INTRA AND POST OPERATIVE ADVERSE EVENTS

Please indicate any adverse events by ticking the appropriate box(es) below.

Death ☐	Intracerebral haematoma ☐	Hemiparesis ☐	Infection ☐
Hemiballism ☐	Dystonia ☐	Confusion ☐	Seizure ☐
Drowsiness ☐	Eyelid apraxia ☐	Diplopia ☐	

Problem with Frame (specify): .....

Anaesthetic Complications (specify): .....

Other (specify): .....

Did the adverse event prolong hospitalisation? No ☐ Yes ☐ If yes, for how long days .....

APPENDIX M: SIX MONTH POST-OP FORM

Surgeon .....

Date of Surgery .....

PART A: POST OPERATIVE ADVERSE EVENTS

Please indicate any adverse events by ticking the appropriate box(es) below.

Remember to ask about admissions to other hospitals.

Device related

Electrode malplacement ☐	Electrode displacement ☐	Electrode fracture ☐
Lead fracture ☐	Infection ☐	Battery malfunction ☐
Skin erosion ☐		

Other ☐ (specify):.....

Side Effects of Stimulation (with optimum anti PD effects)

Hemiballism ☐	Dystonia ☐	Parasthesia ☐
Worsening motor function ☐	Memory impairment ☐	Depression/anxiety ☐
Personality change ☐	Diplopia ☐	Eyelid apraxia ☐
Confusion ☐	Dysarthria ☐	Aphasia ☐

Other ☐ (specify): .....

Did the adverse event require hospitalisation? No ☐ Yes ☐ If yes, for how long.....days

PART C STIMULATION PARAMETERS

Pulse width .....

Rate .....

	LEFT Channel 1	**RIGHT Channel 2**
Electrodes in use	0☐ 1☐ 2☐ 3☐	4☐ 5☐ 6☐ 7☐
Amplitude (volts)		
Range (if applicable)		

PART B CURRENT PD MEDICATION

Drug	**Total daily dosage (mg)**	**Date started (if new)**
.....		
.....		
.....		
.....		

If the patient has ceased taking any PD medication since surgery, please state below:

Drug	**Date Stopped**
.....	
.....	
.....	

Appendix 7: Stata code to generate MAR data

This appendix shows the Stata code used to impose MAR data following onto the complete cases subset of the KAT trial. Here, the missing data pattern observed for the OKS in the KAT trial are reproduced. The overall percentage of participants with some missing outcome data (`miss_loop'), the sample size (`ss_loop') are defined as local macros at the start of the program.

```
clear all

*set all parameters that can be changed
local n = 1001 // number of samples + 1
local ss = `ss_loop' // sample size of simulated dataset
*indicate % of observations with missing data (i.e. 20 for 20%)
local miss_perc = `miss_loop'
local imps = `miss_loop' // specify number of imputations to be performed
*i.e. for non-basecases, have number of imputations == % of data missing,
*as per White recommendations
*dataset prepared in KAT_simulation_data_prep programme
use "$data\kat_sim", clear
*take a random sample of the pre-defined sample size
sample `ss', count

gen p = `miss_perc' / 100 // percentage of missing data
*as the OKS can still be calculated when up to two items are missing, scale up
*the percentage in line with missing data pattern:
replace p = p/0.7305

noisily di  as text `c' " out of " `n'-1
set more off

sort id

** IMPOSING MISSING DATA ON FULL DATASET
// 1.   Generate a random variable, X, from the uniform distribution (i.e.
U[0,1]).
// Then assign the patients to 1 of the observed missing patterns by
// comparing X with the cumulative probabilities
// observed for each MD pattern:
/* simplify use the following pattern:
Missingness pattern  Total True % % used in simulation Cumulative %
Unit non-response    244   67.22%       73.05%              73.05%
Only item 7 missing  52    14.33%       15.57%              88.62%
Only item 4 missing  11    3.03%        3.29%               91.92%
Only item 6 missing  9     2.48%        2.69%               94.61%
Only item 9 missing  7     1.93%        2.10%               96.71%
Only item 10 missing 5     1.38%        1.50%               98.20%
Only item 1 missing  3     0.83%        0.90%               99.10%
Only item 12 missing 3     0.83%        0.90%               100.00%
*/

local seed2 =`c'*(`c'+ 35211)
set seed `seed2'
gen x = uniform()
gen mpattern = 1 if x>=0 & x<=0.7305
replace mpattern = 2 if x>0.7305 & x<=0.8862
replace mpattern = 3 if x>0.8862 & x<=0.9192
```

```
replace mpattern = 4 if x>0.9192 & x<=0.9461
replace mpattern = 5 if x>0.9461 & x<=0.9671
replace mpattern = 6 if x>0.9671 & x<=0.9820
replace mpattern = 7 if x>0.9820 & x<=0.9910
replace mpattern = 8 if x>0.9910 & x<=1
*percentages of MD pattern when taking into account only participants with the
* 8 most common MD pattern

// 2.   Create a linear score for all patients using the beta coefficients in
// Table 2 according to the missing pattern assigned

*focus on pattern that refer to participants with some missing data
gen lscore = .7758 -.0627762*comp_b_alloc + .0245076*age ///
 -.0503112*oks_bl -.0233877*height ///
 +  .024583*asa1 +  .4514279* asa3 ///
 +  .3692998*site_size_m + .2606951*site_size_l if mpattern==1

replace lscore = -2.725454 -.001319*comp_b_alloc -.0097468*age ///
 + .061564*oks_bl -.0008646*height ///
 -.2032481*asa1 -1.007218* asa3 ///
 -1.14821*site_size_m -.8104573*site_size_l if mpattern==2

replace lscore = -10.56717 -.406437*comp_b_alloc + .0638233*age ///
 + .0536308*oks_bl -.0014529*height ///
 -.4217216*asa1 +  1.325208* asa3 ///
 +  .5651921*site_size_m + 0*site_size_l if mpattern==3

replace lscore = -20.19695 -1.184716*comp_b_alloc + .1151257*age ///
 -.0592787*oks_bl +.0562759*height ///
 -.5103956*asa1 -.6014382* asa3 ///
 -.9228536*site_size_m -1.216499*site_size_l if mpattern==4

replace lscore = -11.7455 -.2835892*comp_b_alloc + .1199281*age ///
 + .0257165*oks_bl -.0169899*height ///
 +  .7655309*asa1 +  .0634544* asa3 ///
 +  .0977824*site_size_m + 0*site_size_l if mpattern==5

replace lscore =  -3.283268 -1.396968*comp_b_alloc + .0550662*age ///
 + .0106849*oks_bl -.0355495*height ///
 +  0*asa1 -.003731* asa3 ///
 +  .2043287*site_size_m + 0*site_size_l if mpattern==6

replace lscore = -9.671319 -.753164*comp_b_alloc -.0144333 *age ///
 + .0656632*oks_bl +.0242518*height ///
 +  0*asa1 +  0* asa3 ///
 + .0294031*site_size_m + 0*site_size_l if mpattern==7

replace lscore = -11.56452 -.7392675*comp_b_alloc + .0879289*age ///
 + .0052501*oks_bl -.0011796*height ///
 +  0*asa1 +  0* asa3 ///
 -.0657889*site_size_m + 0*site_size_l if mpattern==8
 *updated based on logistic regression

// 3.
centile lscore if mpattern==1, centile(33 66)
gen oddsmiss = 1 if mpattern==1 & lscore<=r(c_1)
replace oddsmiss = 2 if mpattern==1 & lscore>r(c_1) & lscore<=r(c_2)
replace oddsmiss = 3 if mpattern==1 & lscore>r(c_2)

centile lscore if mpattern==2, centile(33 66)
replace oddsmiss = 1 if mpattern==2 & lscore<=r(c_1)
replace oddsmiss = 2 if mpattern==2 & lscore>r(c_1) & lscore<=r(c_2)
replace oddsmiss = 3 if mpattern==2 & lscore>r(c_2)

centile lscore if mpattern==3, centile(33 66)
replace oddsmiss = 1 if mpattern==3 & lscore<=r(c_1)
replace oddsmiss = 2 if mpattern==3 & lscore>r(c_1) & lscore<=r(c_2)
replace oddsmiss = 3 if mpattern==3 & lscore>r(c_2)
```

```
centile lscore if mpattern==4, centile(33 66)
replace oddsmiss = 1 if mpattern==4 & lscore<=r(c_1)
replace oddsmiss = 2 if mpattern==4 & lscore>r(c_1) & lscore<=r(c_2)
replace oddsmiss = 3 if mpattern==4 & lscore>r(c_2)

centile lscore if mpattern==5, centile(33 66)
replace oddsmiss = 1 if mpattern==5 & lscore<=r(c_1)
replace oddsmiss = 2 if mpattern==5 & lscore>r(c_1) & lscore<=r(c_2)
replace oddsmiss = 3 if mpattern==5 & lscore>r(c_2)

centile lscore if mpattern==6, centile(33 66)
replace oddsmiss = 1 if mpattern==6 & lscore<=r(c_1)
replace oddsmiss = 2 if mpattern==6 & lscore>r(c_1) & lscore<=r(c_2)
replace oddsmiss = 3 if mpattern==6 & lscore>r(c_2)

centile lscore if mpattern==7, centile(33 66)
replace oddsmiss = 1 if mpattern==7 & lscore<=r(c_1)
replace oddsmiss = 2 if mpattern==7 & lscore>r(c_1) & lscore<=r(c_2)
replace oddsmiss = 3 if mpattern==7 & lscore>r(c_2)

centile lscore if mpattern==8, centile(33 66)
replace oddsmiss = 1 if mpattern==8 & lscore<=r(c_1)
replace oddsmiss = 2 if mpattern==8 & lscore>r(c_1) & lscore<=r(c_2)
replace oddsmiss = 3 if mpattern==8 & lscore>r(c_2)

*4.Calculate the number in each ODDSMISS subgroup within each pattern, noddmiss
sort mpattern id, stable
by mpattern: egen x1 = count (oddsmiss) if oddsmiss==1
by mpattern: egen x2 = count (oddsmiss) if oddsmiss==2
by mpattern: egen x3 = count (oddsmiss) if oddsmiss==3

egen noddmiss = rsum(x1 x2 x3)
drop x1 x2 x3

// 5.   Calculate the probability of being missing for each subject,
* PMISS, based on a total % of missing P
* oodsmiss - probability of having observed MD pattern
* p - previously defined overall percentage of participants with missing data
* ss - number of participants in dataset - consider to make this a variable
*  that can be changed later on to assess effect of imputation on smaller
*  samples                                                          **!!!**
gen pmiss = (oddsmiss*p*`ss'*0.7305)/(6*noddmiss) if mpattern==1
replace pmiss = (oddsmiss*p*`ss'*0.1557)/(6*noddmiss) if mpattern==2
replace pmiss = (oddsmiss*p*`ss'*0.0329)/(6*noddmiss) if mpattern==3
replace pmiss = (oddsmiss*p*`ss'*0.0269)/(6*noddmiss) if mpattern==4
replace pmiss = (oddsmiss*p*`ss'*0.0210)/(6*noddmiss) if mpattern==5
replace pmiss = (oddsmiss*p*`ss'*0.0150)/(6*noddmiss) if mpattern==6
replace pmiss = (oddsmiss*p*`ss'*0.0090)/(6*noddmiss) if mpattern==7
replace pmiss = (oddsmiss*p*`ss'*0.0090)/(6*noddmiss) if mpattern==8

// 6. Generate another random variable, Y, from a uniform distribution (i.e.
U[0,1])
local seed3 = (`c'*187)+`c'
set seed `seed3'
gen y = uniform()
gen missing = 0
replace missing = 1 if y<pmiss

// 7. Create new data set
gen oks1_miss = oks1_5y
replace oks1_miss=. if (mpattern==1 & missing==1) | (mpattern==7 & missing==1)
gen oks2_miss = oks2_5y
replace oks2_miss= . if mpattern==1 & missing==1
gen oks3_miss = oks3_5y
replace oks3_miss=. if mpattern==1 & missing==1
gen oks4_miss = oks4_5y
replace oks4_miss=. if (mpattern==1 & missing==1) | (mpattern==3 & missing==1)
```

```
gen oks5_miss = oks5_5y
replace oks5_miss=. if mpattern==1 & missing==1
gen oks6_miss = oks6_5y
replace oks6_miss=. if (mpattern==1 & missing==1) | (mpattern==4 & missing==1)
gen oks7_miss = oks7_5y
replace oks7_miss=. if (mpattern==1 & missing==1) | (mpattern==2 & missing==1)
gen oks8_miss = oks8_5y
replace oks8_miss=. if mpattern==1 & missing==1
gen oks9_miss = oks9_5y
replace oks9_miss=. if (mpattern==1 & missing==1) | (mpattern==5 & missing==1)
gen oks10_miss = oks10_5y
replace oks10_miss=. if (mpattern==1 & missing==1) | (mpattern==6 & missing==1)
gen oks11_miss = oks11_5y
replace oks11_miss=. if mpattern==1 & missing==1
gen oks12_miss = oks12_5y
replace oks12_miss=. if (mpattern==1 & missing==1) | (mpattern==8 & missing==1)

*in the dataset with missing values:
*calculate OKS - or create missing values where scoring manual does not
* allow calculations due to missing values
egen oks_num_miss = rowmiss(oks1_miss - oks12_miss)
egen oks_rowtotal_miss = rowtotal(oks1_miss - oks12_miss)
egen oks_rowmean_miss = rowmean(oks1_miss - oks12_miss)
gen oks_miss = oks_rowtotal_miss if oks_num_miss==0
replace oks_miss = oks_rowtotal_miss + round(oks_rowmean_miss) ///
 if oks_num_miss==1
replace oks_miss = oks_rowtotal_miss + 2*round(oks_rowmean_miss) ///
 if oks_num_miss==2
replace oks_miss = . if oks_num_miss > 2

*calculate subscales:
/* in line with email received from David Churchman at ISIS
(email received on 10/11/2015, 12:41):
One missing per subscale is allowed. If one item (per subscale) is missing then
use mean value of other items in that subscale to fill in the missing value.
Advice received from Jill Dawson via ISIS
*/
*pain subscale
egen oks_num_miss_p = rowmiss(oks1_miss oks4_miss oks5_miss oks6_miss ///
 oks8_miss oks9_miss oks10_miss)
egen oks_rowtotal_miss_p = rowtotal(oks1_miss oks4_miss oks5_miss oks6_miss ///
 oks8_miss oks9_miss oks10_miss)
egen oks_rowmean_miss_p = rowmean(oks1_miss oks4_miss oks5_miss oks6_miss ///
 oks8_miss oks9_miss oks10_miss)
gen oks_pain_miss = (oks_rowtotal_miss_p)/28*100 if oks_num_miss_p==0
replace oks_pain_miss = ///
 (oks_rowtotal_miss_p + round(oks_rowmean_miss_p))/28*100 ///
 if oks_num_miss_p==1
replace oks_pain_miss = . if oks_num_miss_p > 1

*function subscale
egen oks_num_miss_f = rowmiss(oks2_miss oks3_miss oks7_miss oks11_miss ///
 oks12_miss)
egen oks_rowtotal_miss_f = rowtotal(oks2_miss oks3_miss oks7_miss oks11_miss ///
 oks12_miss)
egen oks_rowmean_miss_f = rowmean(oks2_miss oks3_miss oks7_miss oks11_miss ///
 oks12_miss)
gen oks_func_miss = (oks_rowtotal_miss_f)/20*100 if oks_num_miss_f==0
replace oks_func_miss = ///
 (oks_rowtotal_miss_f + round(oks_rowmean_miss_f))/20*100 ///
 if oks_num_miss_f==1
replace oks_func_miss = . if oks_num_miss_f > 1


label var oks_pain_miss "OKS pain subscale with simulated missing data"
label var oks_func_miss "OKS functino subscale with simulated missing data"

label variable oks_miss "OKS with simulated missing data"
```

```
label variable oks1_miss "Missing OKS Q1"
label variable oks2_miss "Missing OKS Q2"
label variable oks3_miss "Missing OKS Q3"
label variable oks4_miss "Missing OKS Q4"
label variable oks5_miss "Missing OKS Q5"
label variable oks6_miss "Missing OKS Q6"
label variable oks7_miss "Missing OKS Q7"
label variable oks8_miss "Missing OKS Q8"
label variable oks9_miss "Missing OKS Q9"
label variable oks10_miss "Missing OKS Q10"
label variable oks11_miss "Missing OKS Q11"
label variable oks12_miss "Missing OKS Q12"
```

Appendix 8: MAE plots for the comparison of applying MI at the composite score, subscale or item level

This appendix contains graphs showing the MAE of the different MI approaches, i.e. imputing either at the composite score, subscale or item level. The graphs supplement the assessment of the MI approaches using the RMSE as presented in Chapter 4.

8.1 Comparative performance of the different MI approaches: OKS

8.1.1 Results for the OKS simulations – using the observed missing data pattern

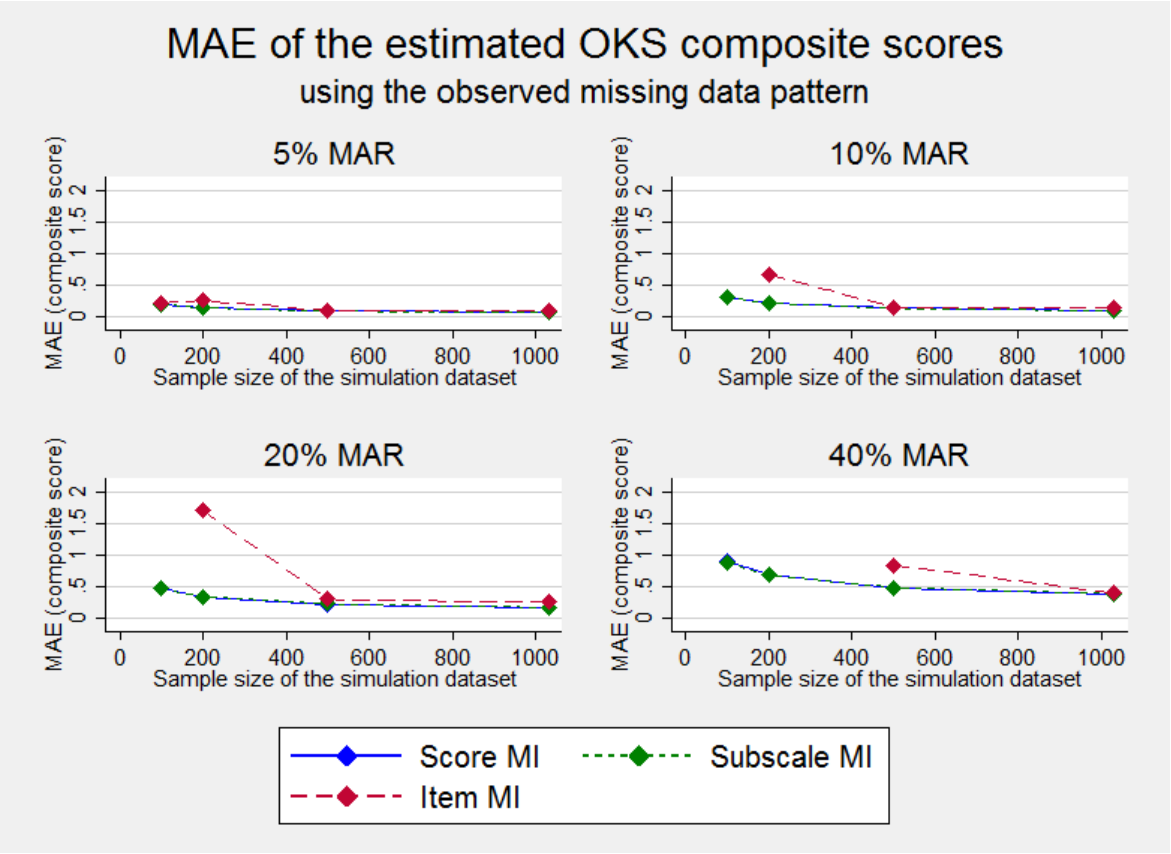


Figure 2: MAE in the OKS composite score estimates

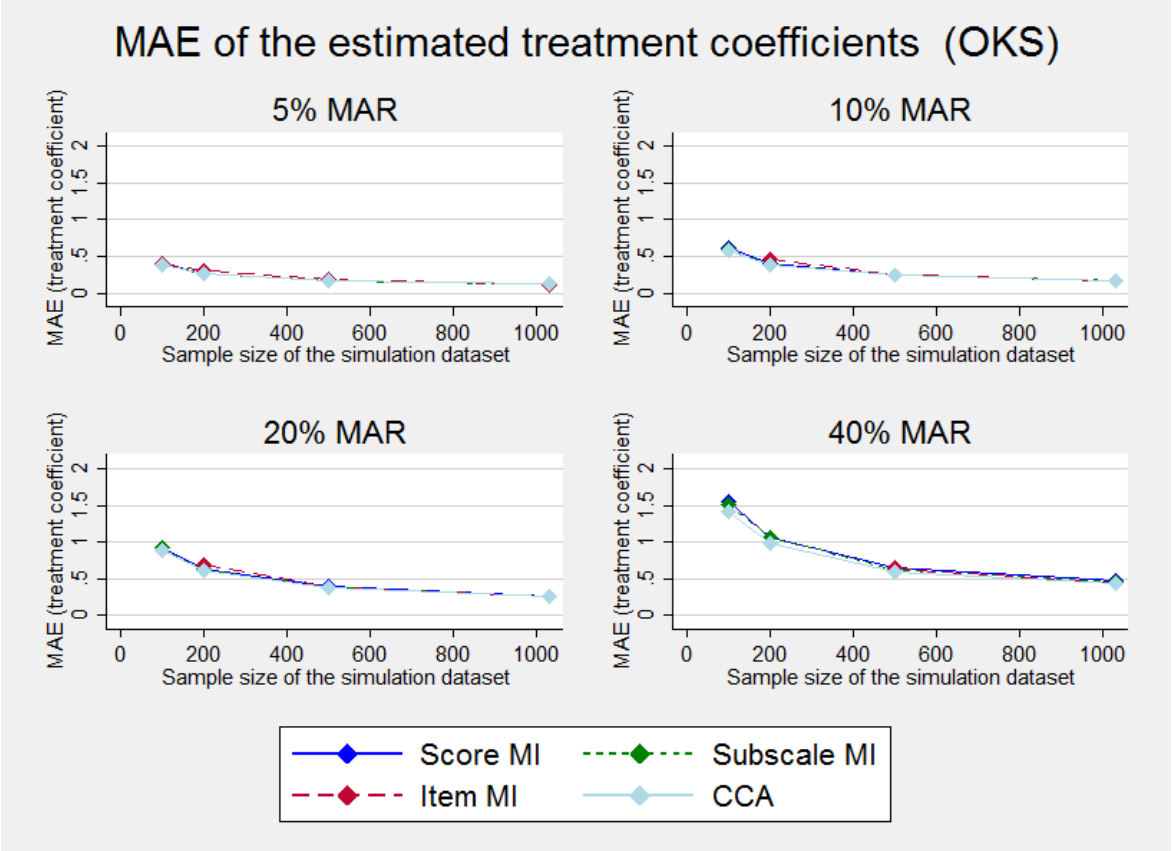


Figure 3: MAE in the treatment coefficient estimates using the imputed OKS as the outcome variable in the regression model

8.1.2 Results for the OKS simulations – simulating a unit nonresponse MAR mechanism

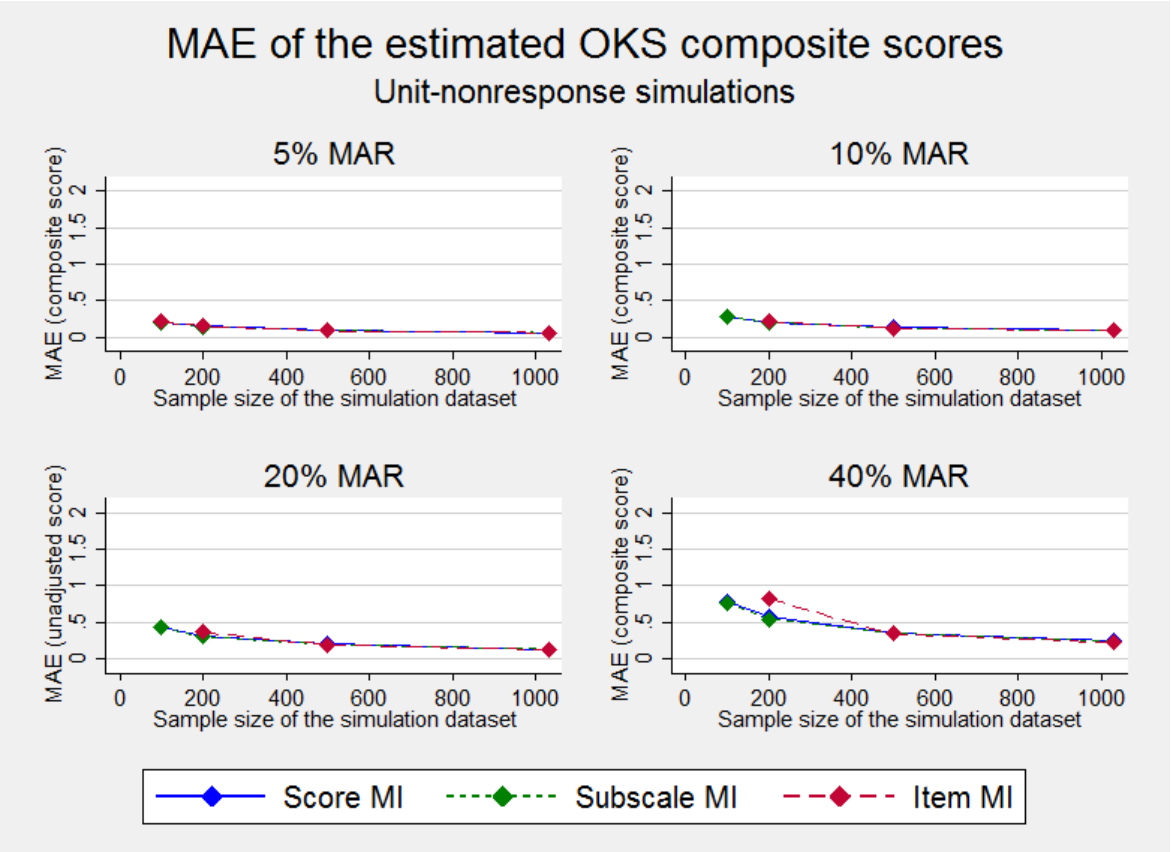


Figure 4: MAE in the OKS composite score estimates (unit-nonresponse simulations)

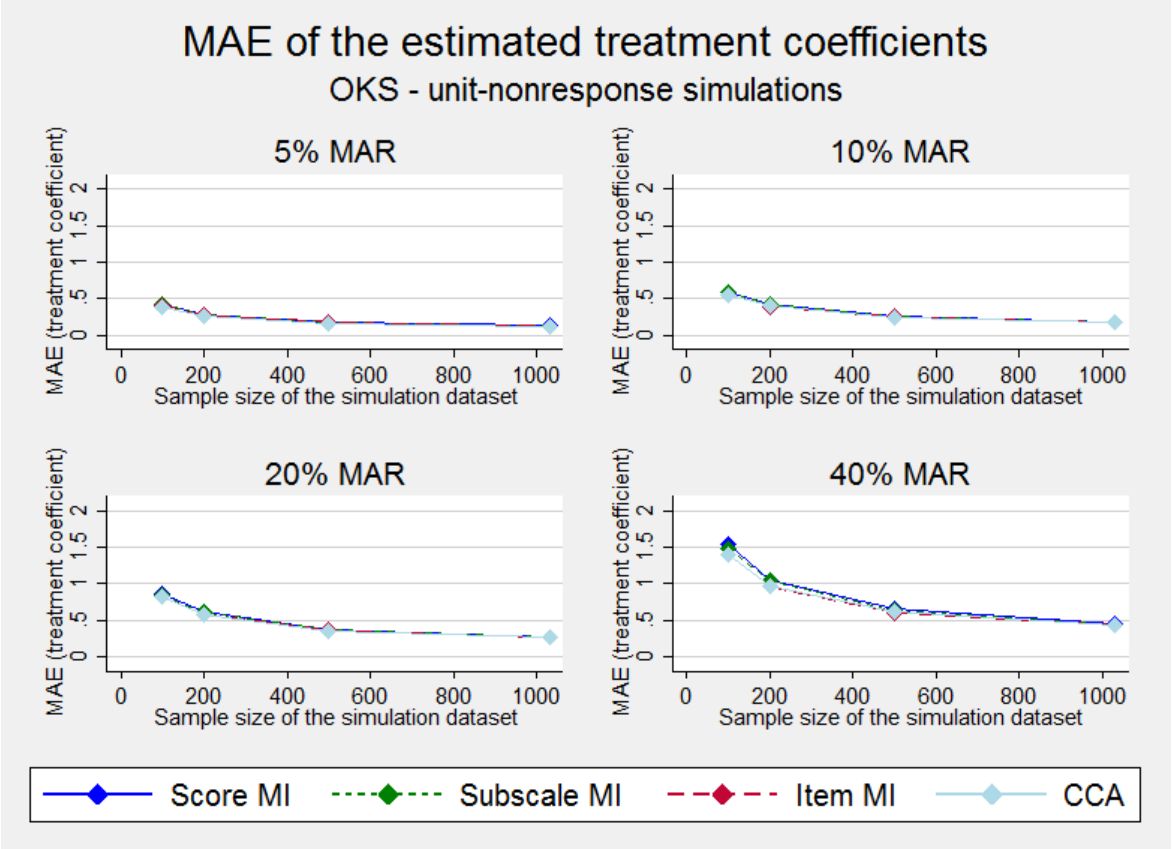


Figure 5: MAE in the treatment coefficient estimates using the imputed OKS as the outcome variable in the regression model

8.1.3 Results for the OKS simulations – simulating 70% item missingness

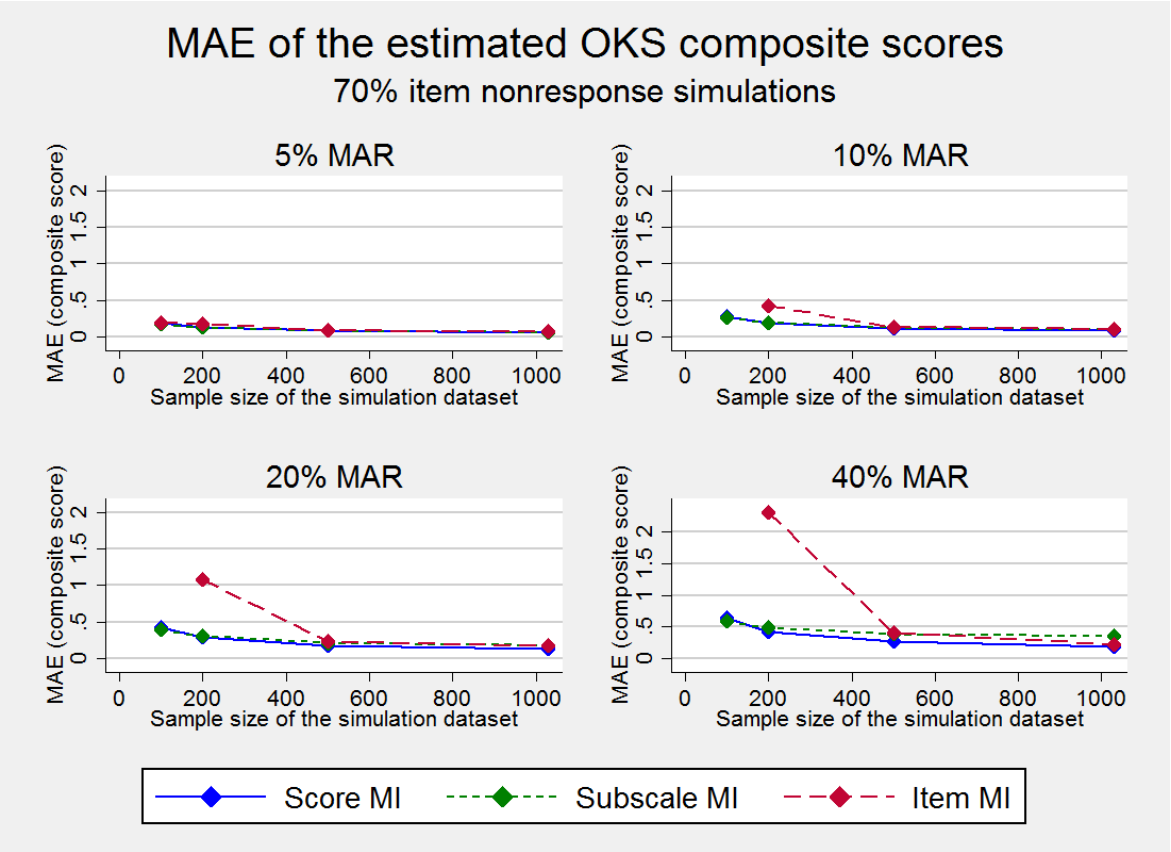


Figure 6: RMSE in the OKS composite score estimates (70% item missingness simulations)

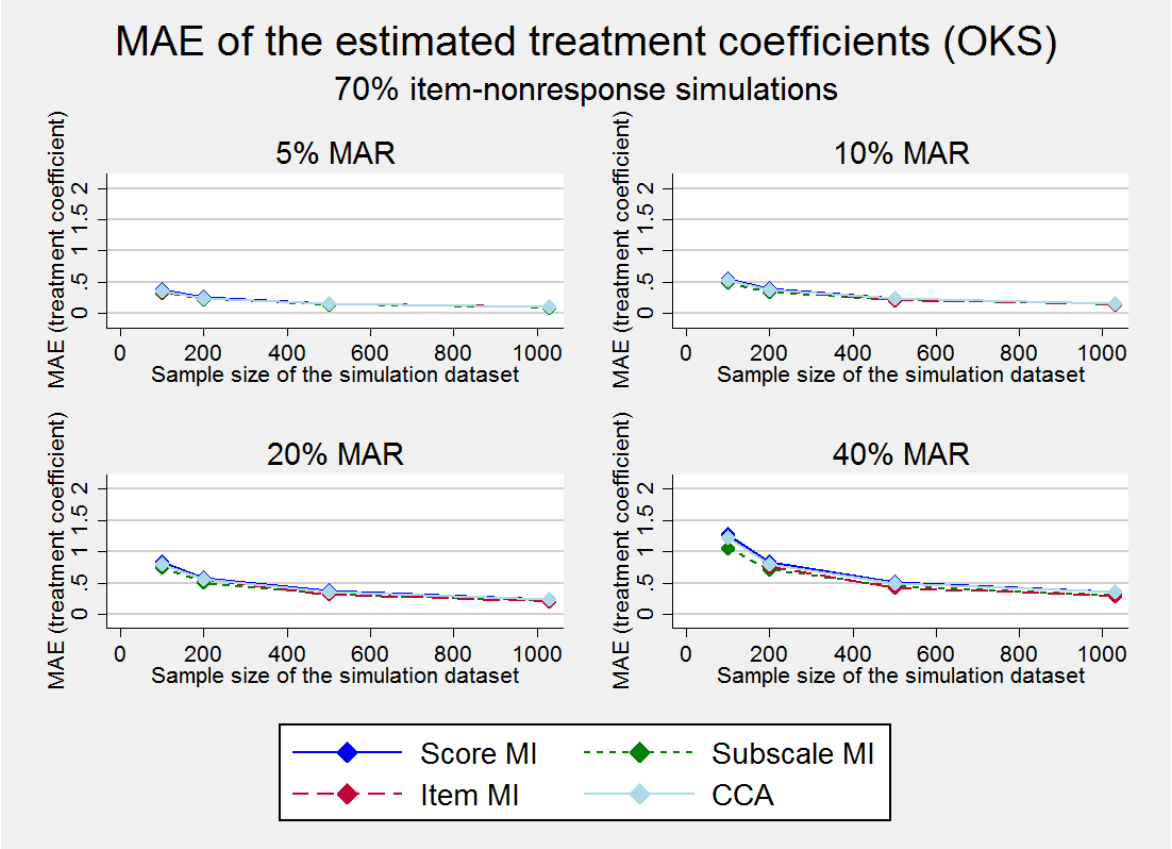


Figure 7: RMSE in the treatment coefficient estimates using the imputed OKS as the outcome variable in the regression model (70% item missingness simulations)

8.1.4 Results for the OKS simulations – introducing a five point treatment effect

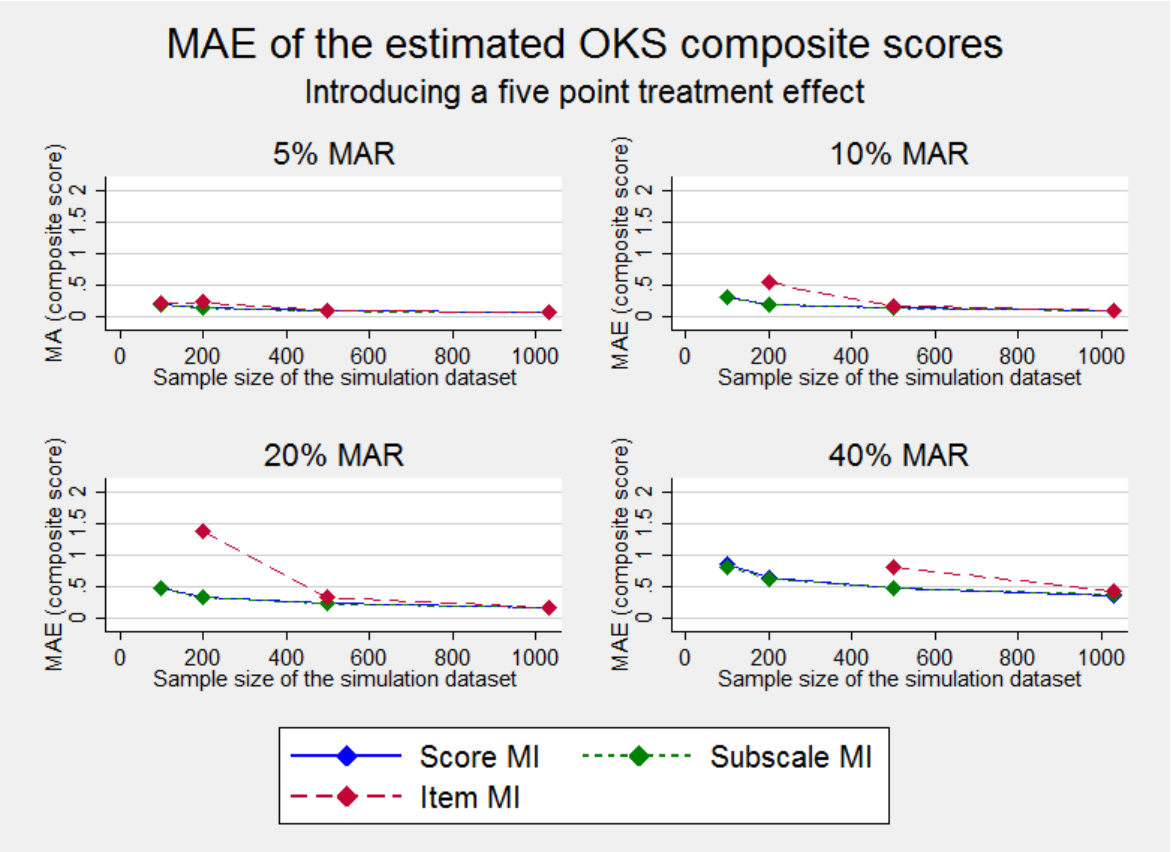


Figure 8: MAE in the OKS composite score estimates (introducing a five point treatment effect)

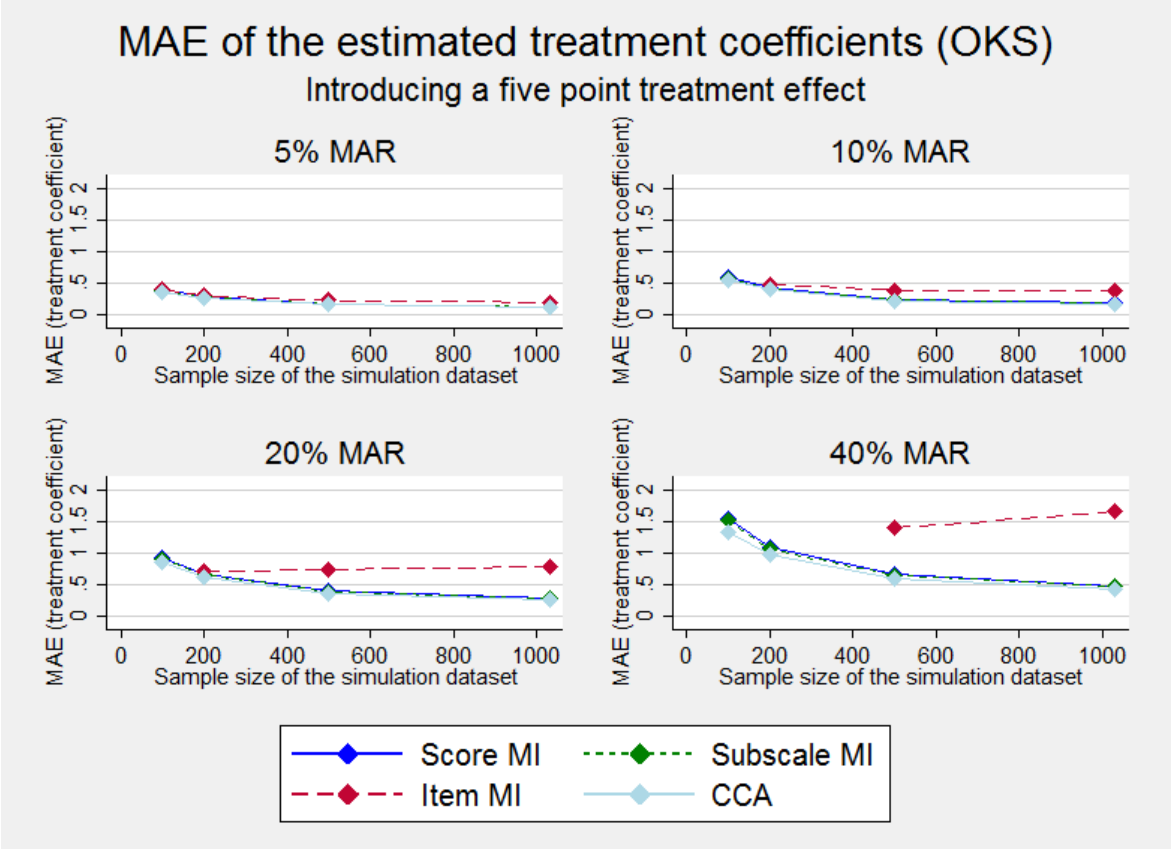


Figure 9: MAE in the treatment coefficient estimates using the imputed OKS as the outcome variable in the regression model (introducing a five point treatment effect)

8.1.5 Results for the OKS simulations – Comparing scoring in line with the scoring manual versus no mean imputations

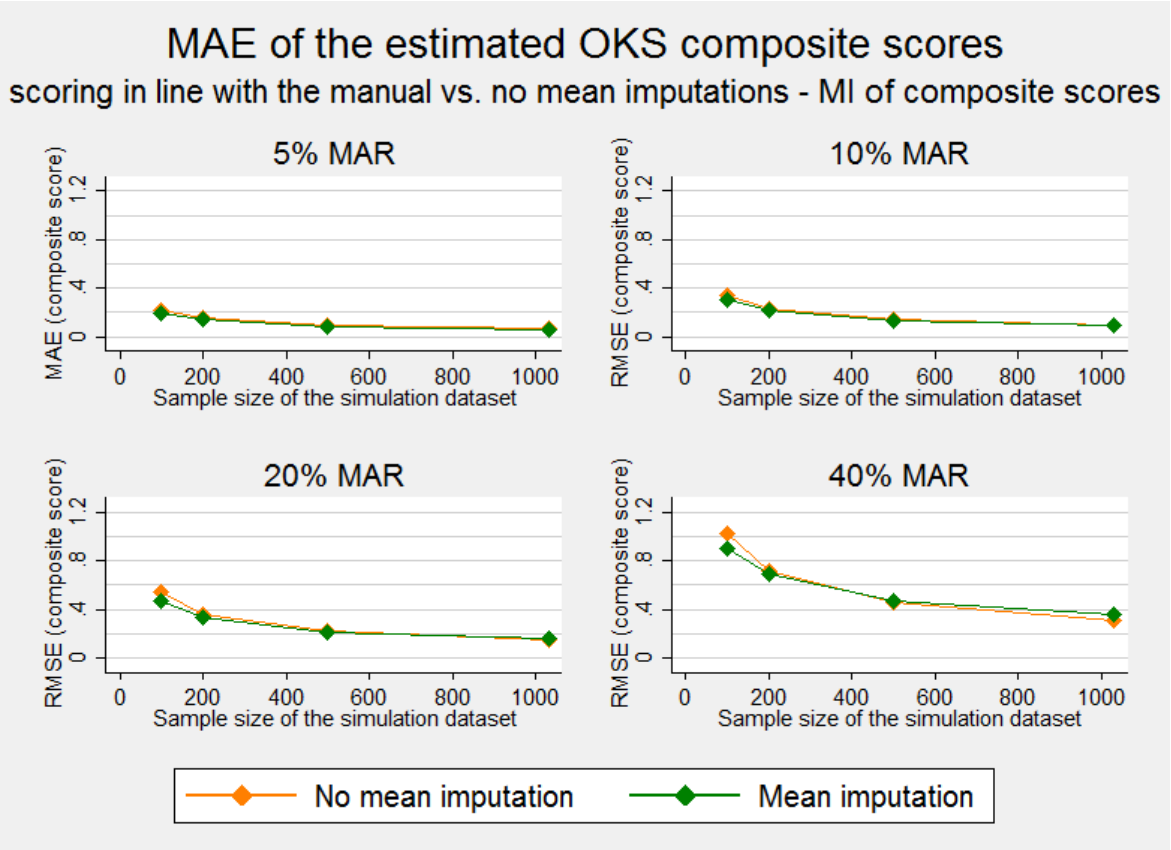


Figure 10: MAE in the OKS composite score estimates comparing the effect of following the scoring manual vs. not using mean imputation for MI at the composite score level (using the observed missing data patterns)

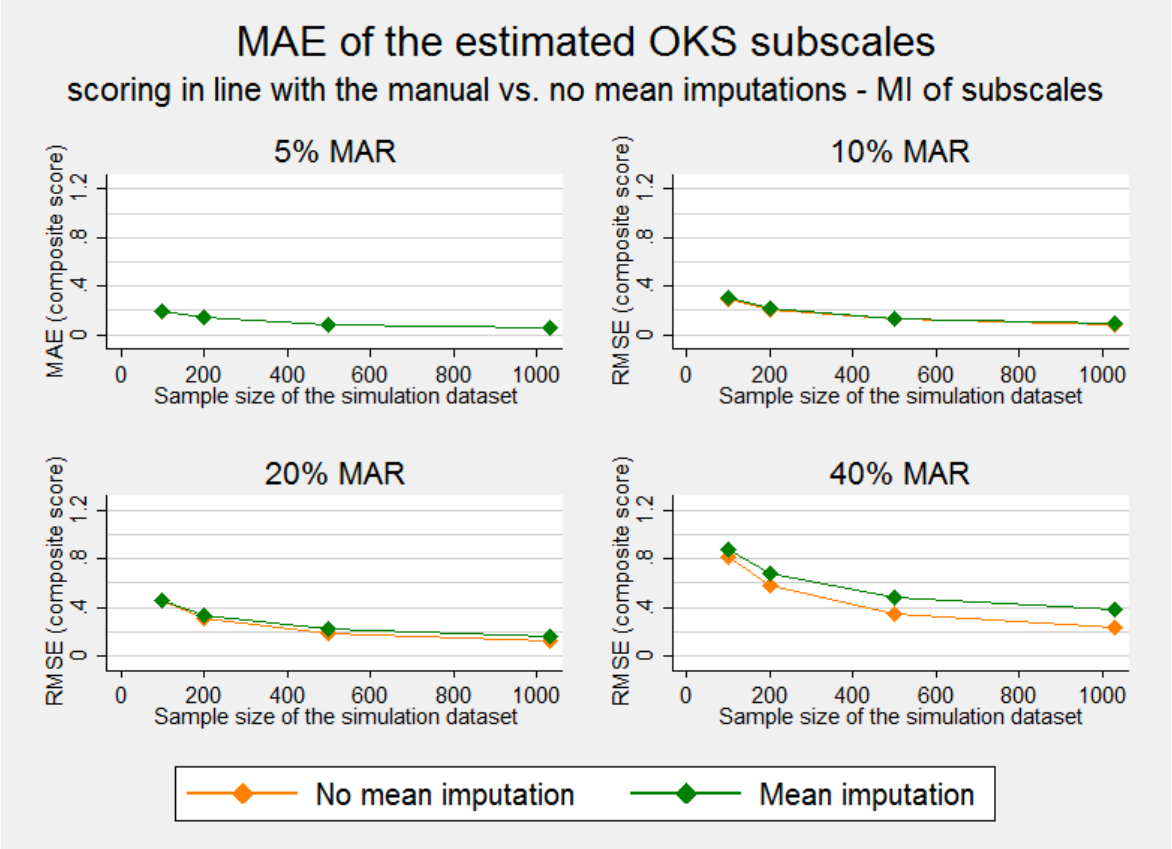


Figure 11: MAE in the OKS composite score estimates comparing the effect of following the scoring manual vs. not using mean imputation for MI at the subscale level (using the observed missing data patterns)

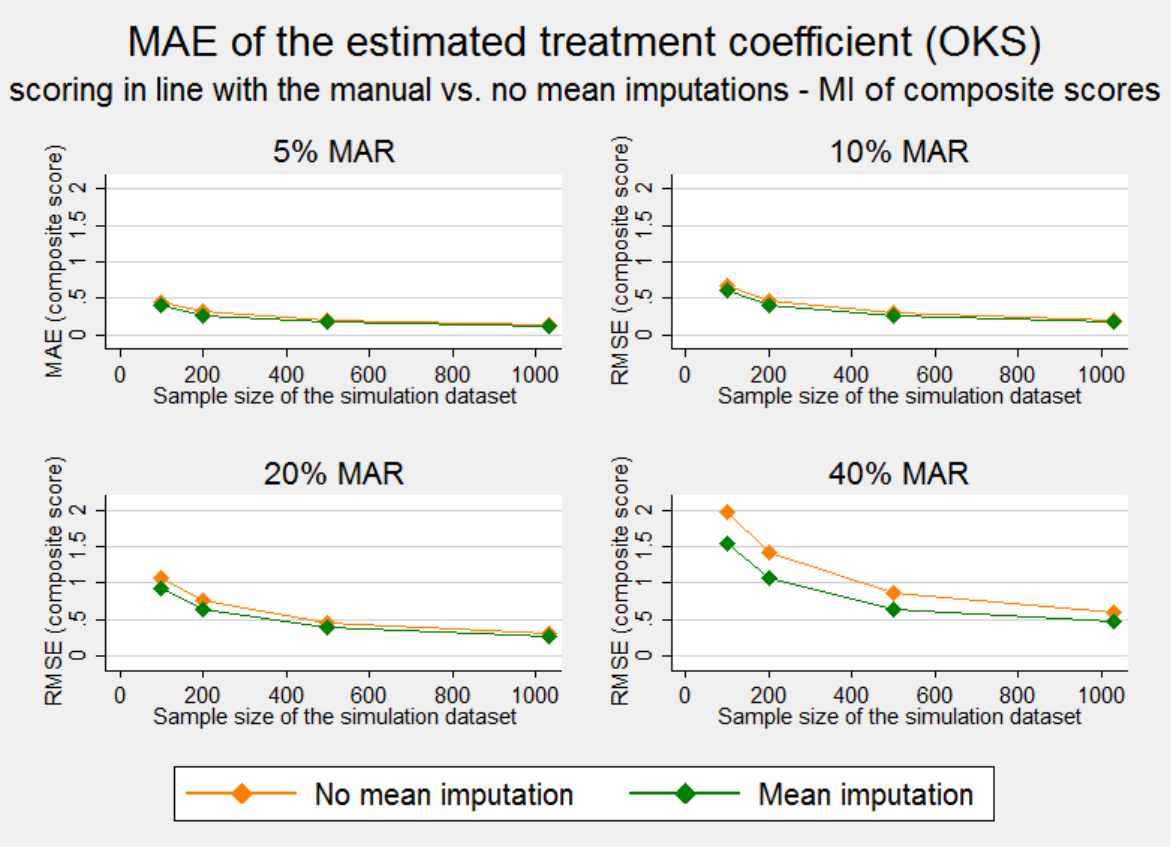


Figure 12: MAE in the treatment coefficient estimates comparing the effect of following the scoring manual vs. not using mean imputation for MI at the composite score level (using the observed missing data patterns)

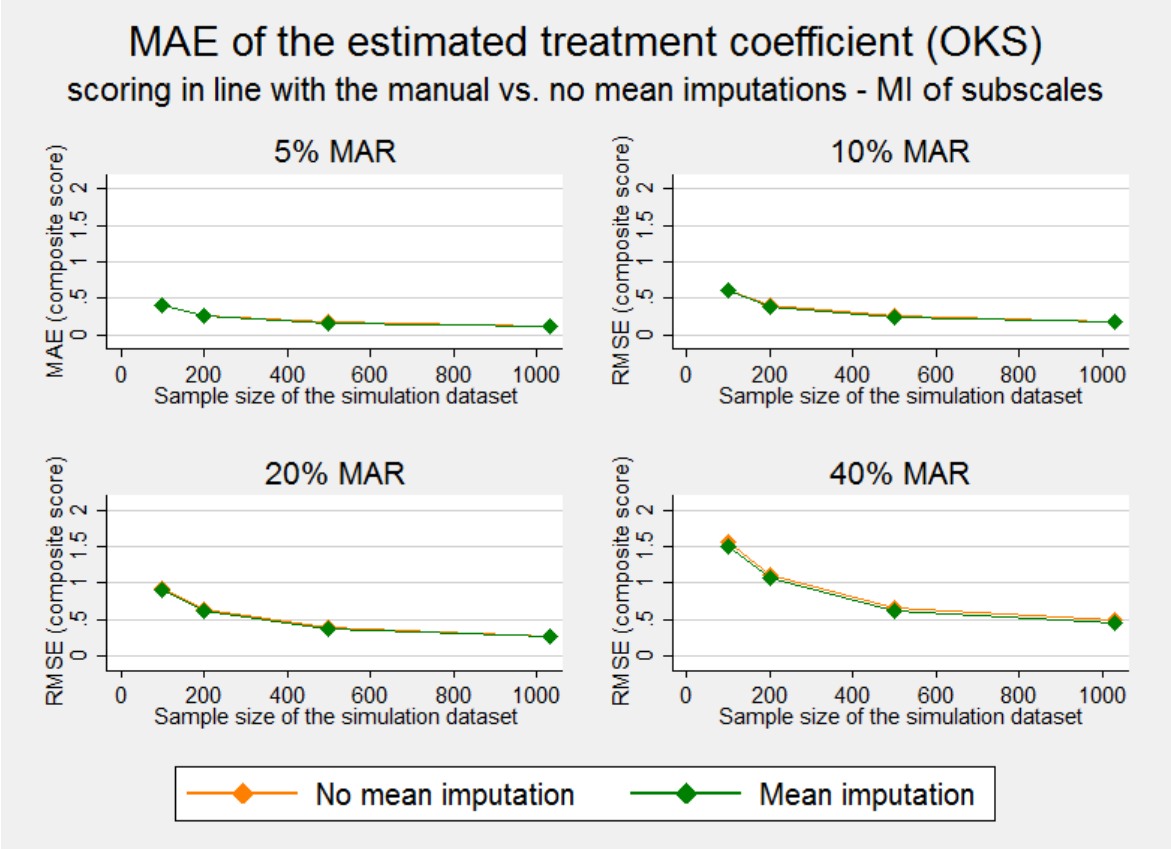


Figure 13: MAE in the treatment coefficient estimates comparing the effect of following the scoring manual vs. not using mean imputation for MI at the subscale level (using the observed missing data patterns)

8.2 Comparative performance of the different MI approaches: EQ-5D-3L

8.2.1 Results for the EQ-5D-3L simulations – simplified item-level imputations

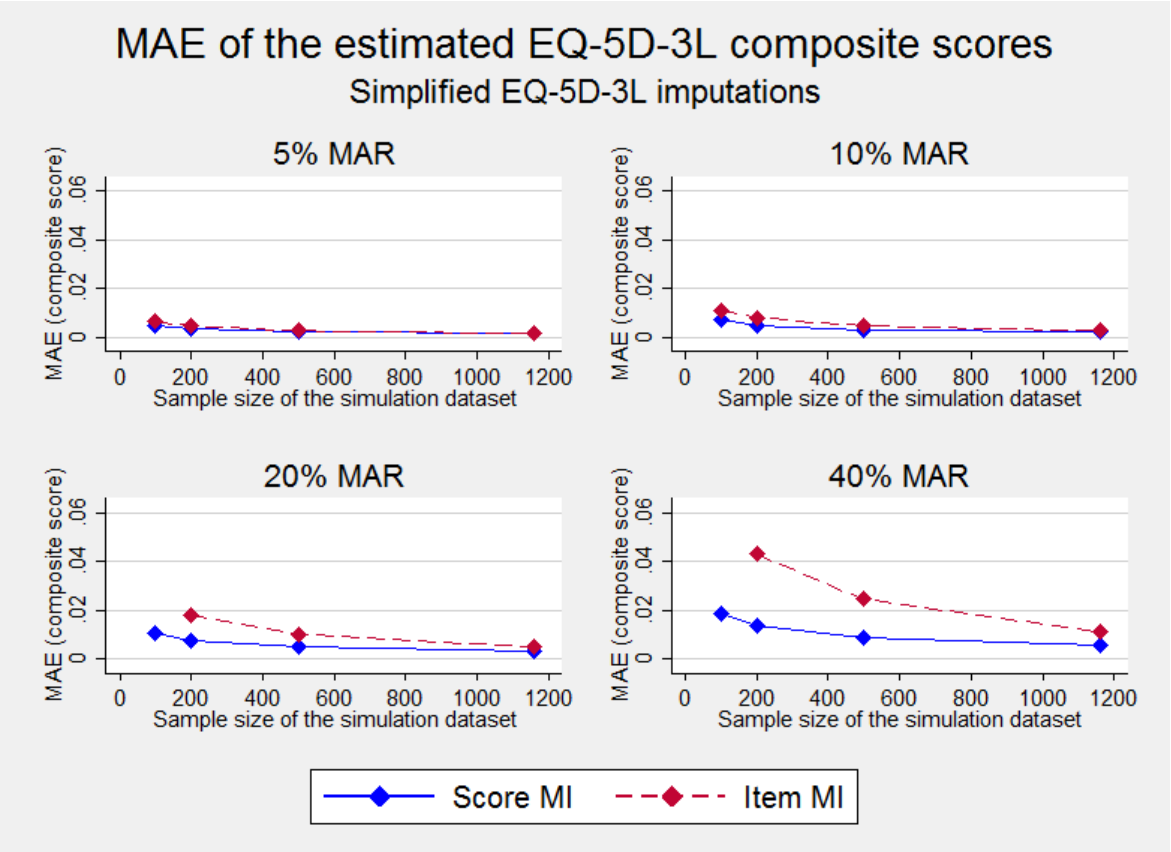


Figure 14: MAE in the EQ-5D-3L composite score estimates

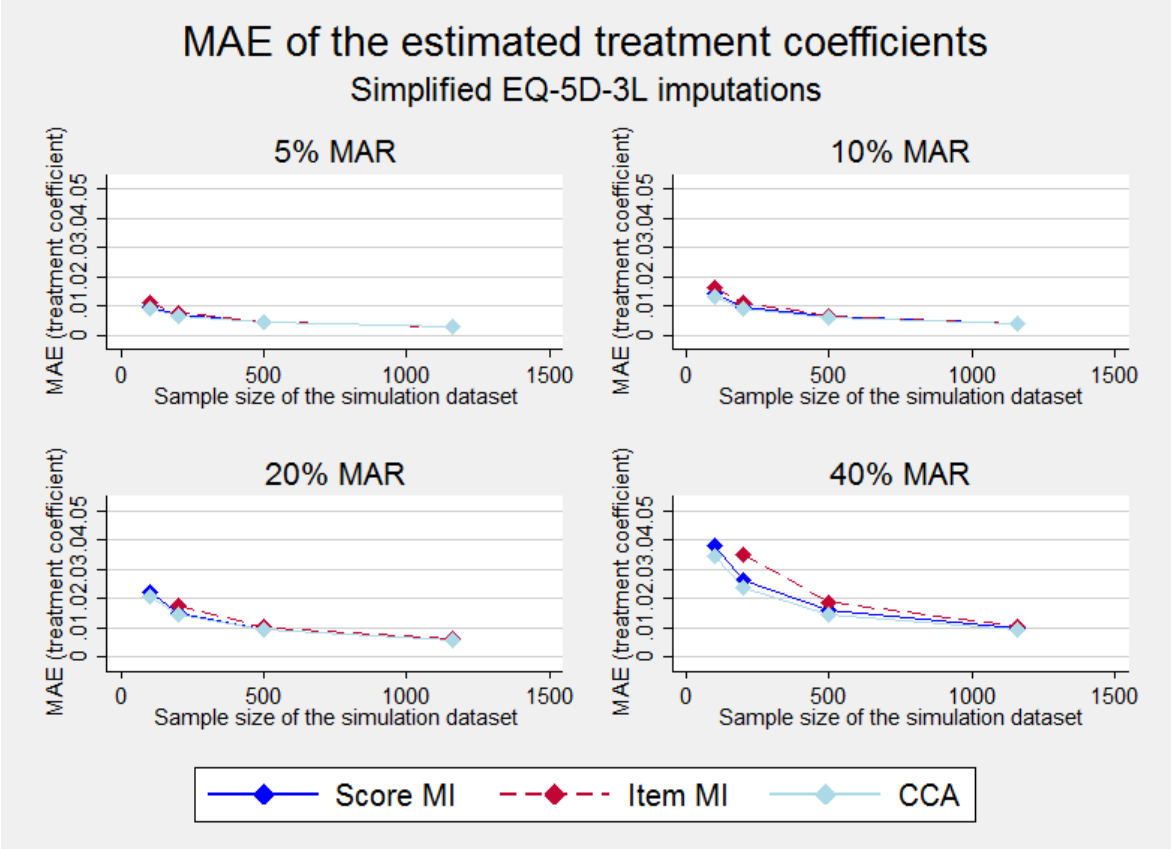


Figure 15: MAE in the treatment coefficient estimates using the imputed EQ-5D-3L as the outcome variable in the regression model

8.2.2 Results for the EQ-5D-3L simulations – simplified item-level imputations

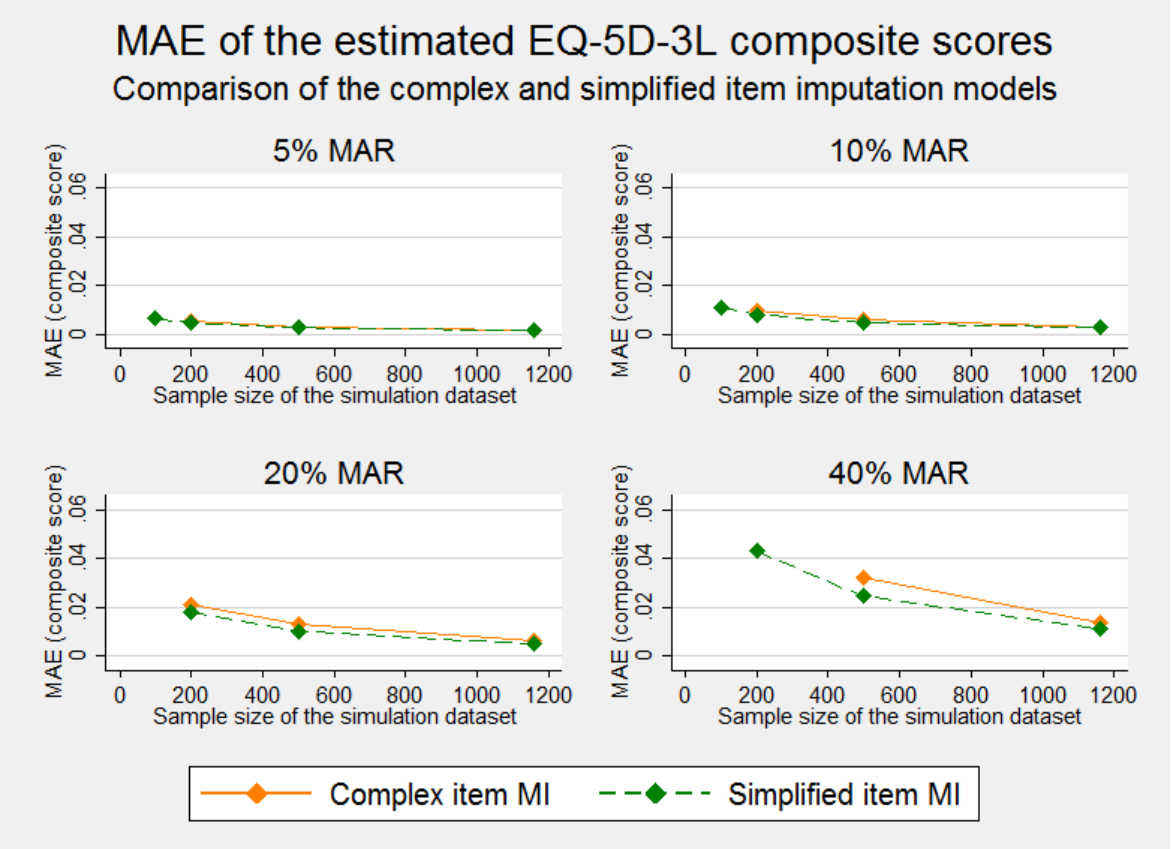


Figure 16: MAE in the EQ-5D-3L composite score estimates – comparing the complex and simplified item imputation model

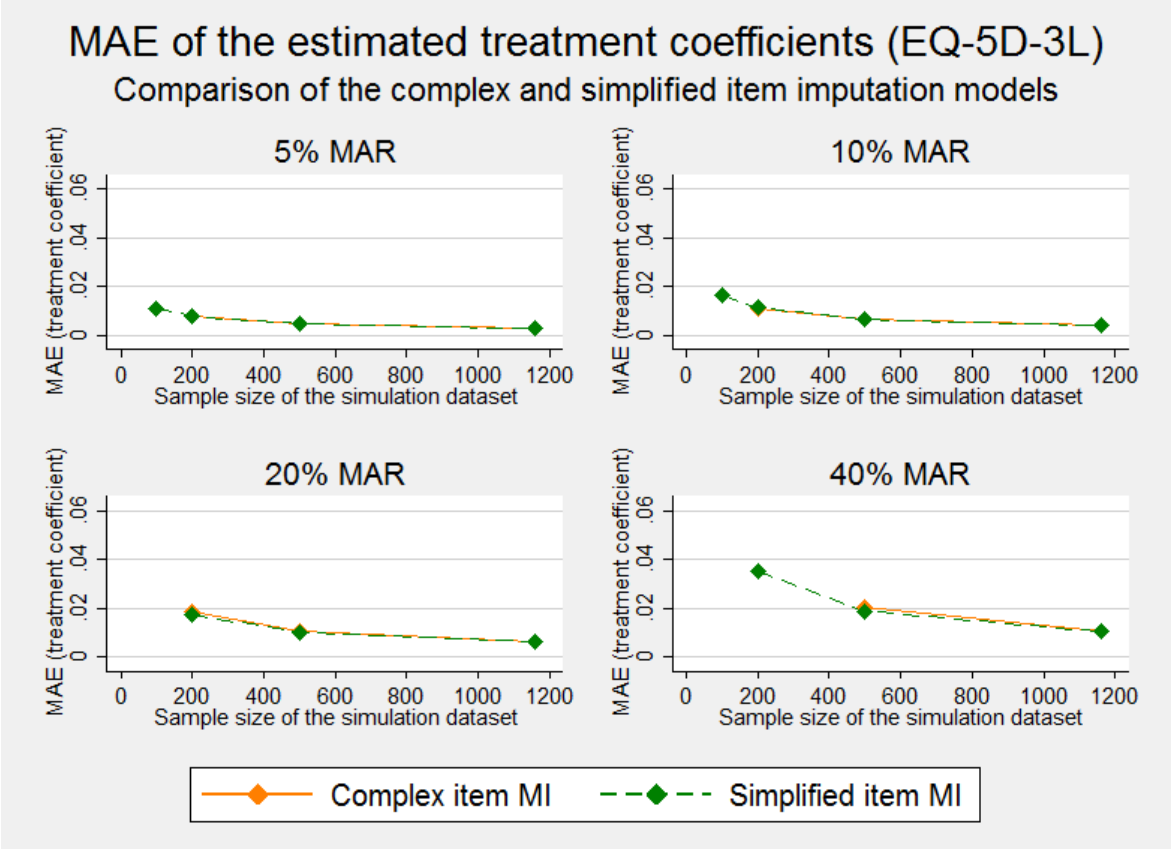


Figure 17: MAE in the treatment coefficient estimates using the imputed EQ-5D-3L as the outcome variable in the regression model – comparing the complex and simplified item imputation model

8.3 Comparative performance of the different MI approaches: SF-12

8.3.1 Results for the EQ-5D-3L simulations – simplified item-level imputations

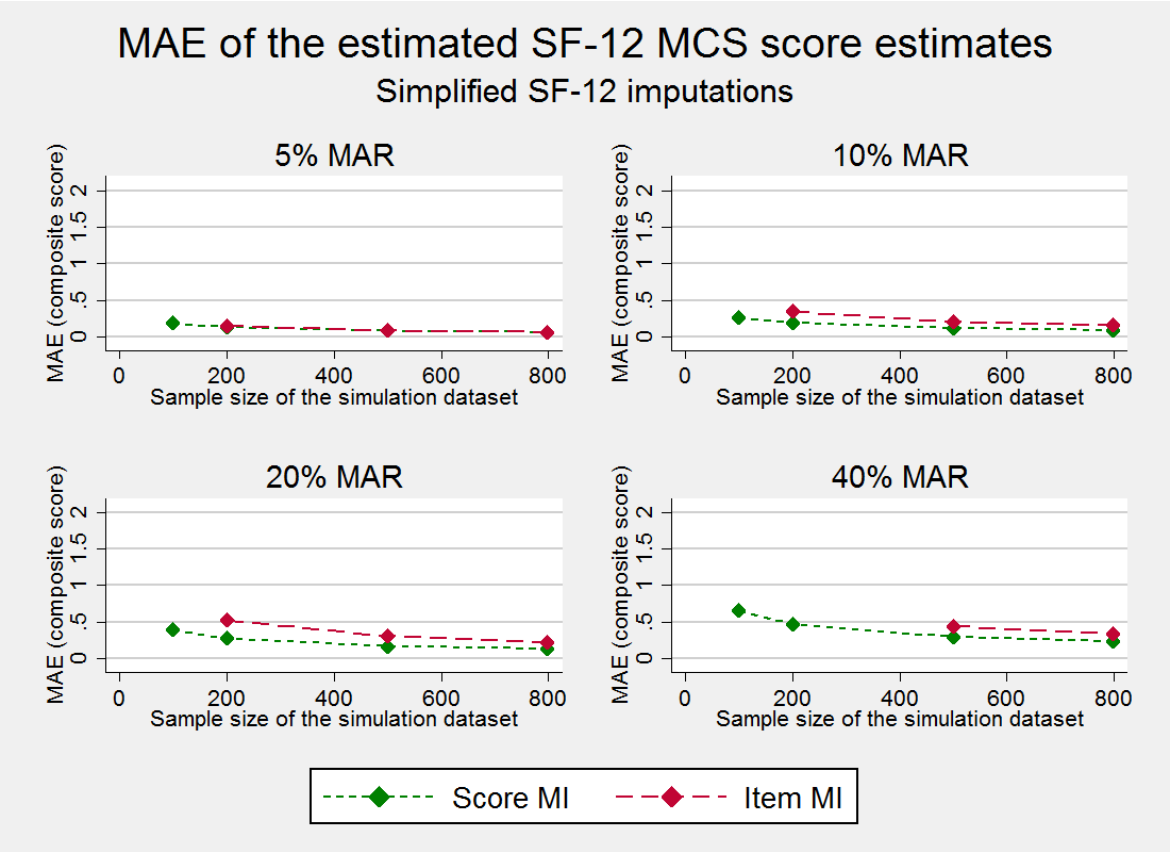


Figure 18: RMSE in the SF-12 MCS score estimates

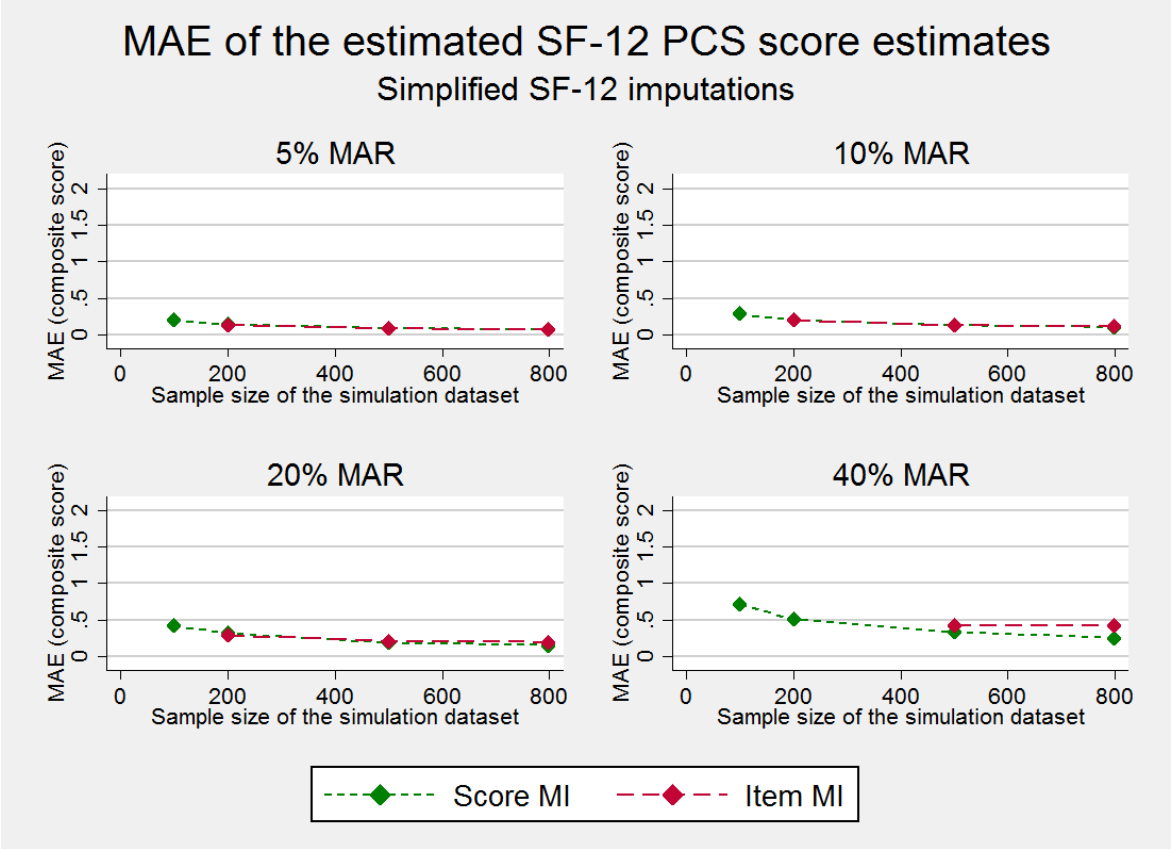


Figure 19: RMSE in the SF-12 PCS score estimates

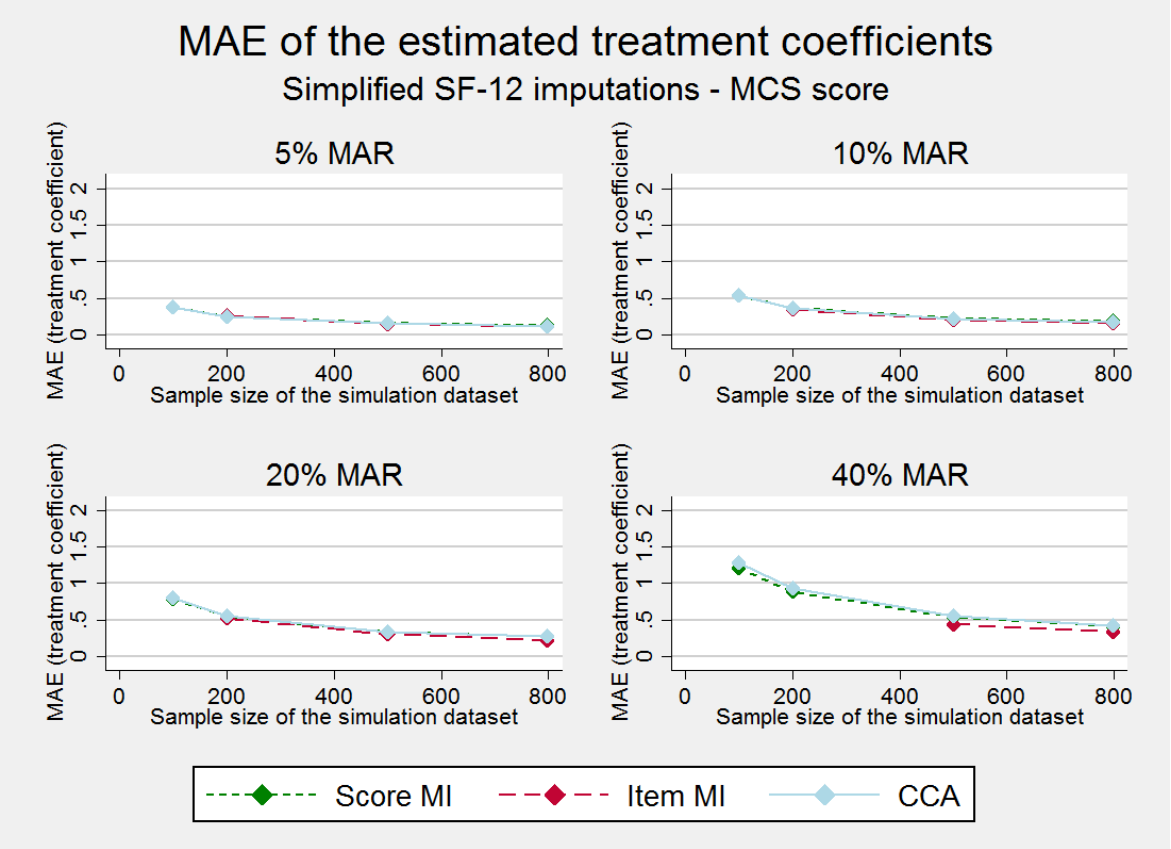


Figure 20: RMSE in the treatment coefficient estimates using the imputed SF-12 MCS score as the outcome variable in the regression model

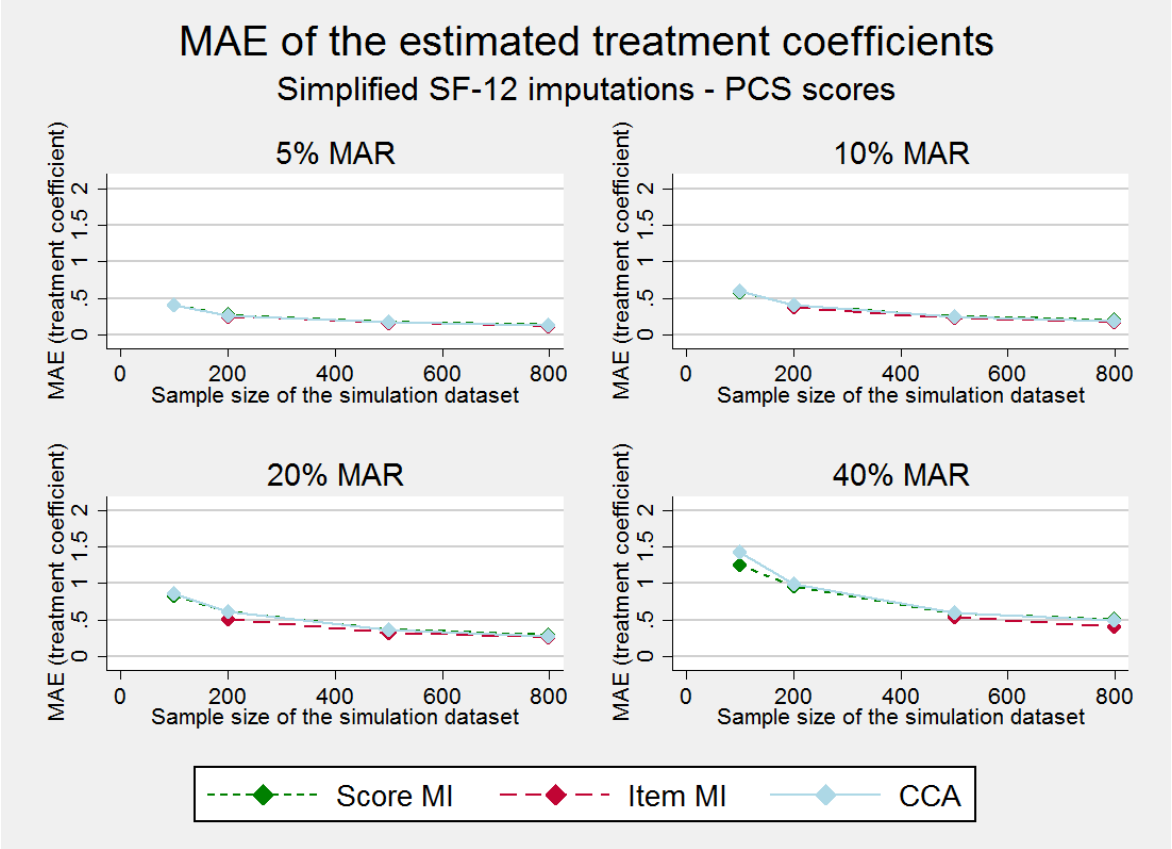


Figure 21: MAE in the treatment coefficient estimates using the imputed SF-12 PCS score as the outcome variable in the regression model

8.3.2 Comparison of the complex and simplified item imputation model

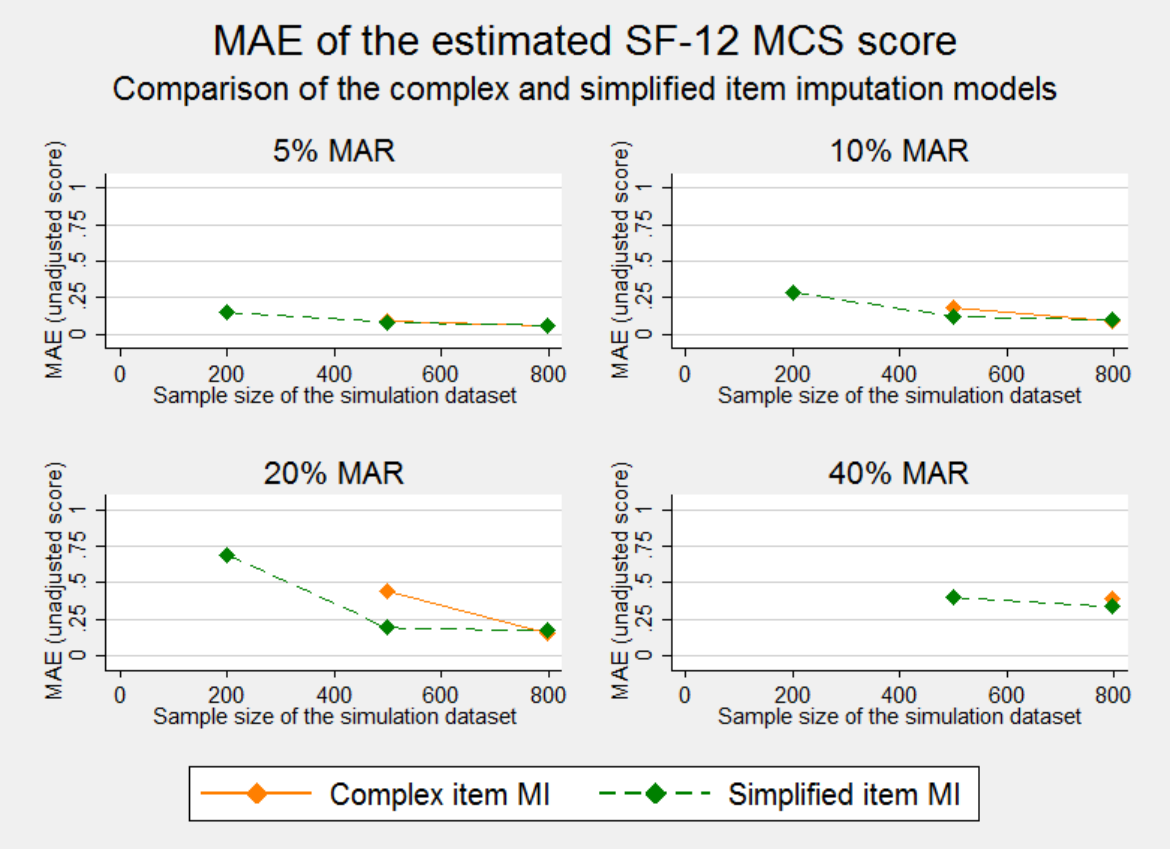


Figure 22: MAE in the SF-12 MCS score estimates – comparing the complex and simplified item imputation model

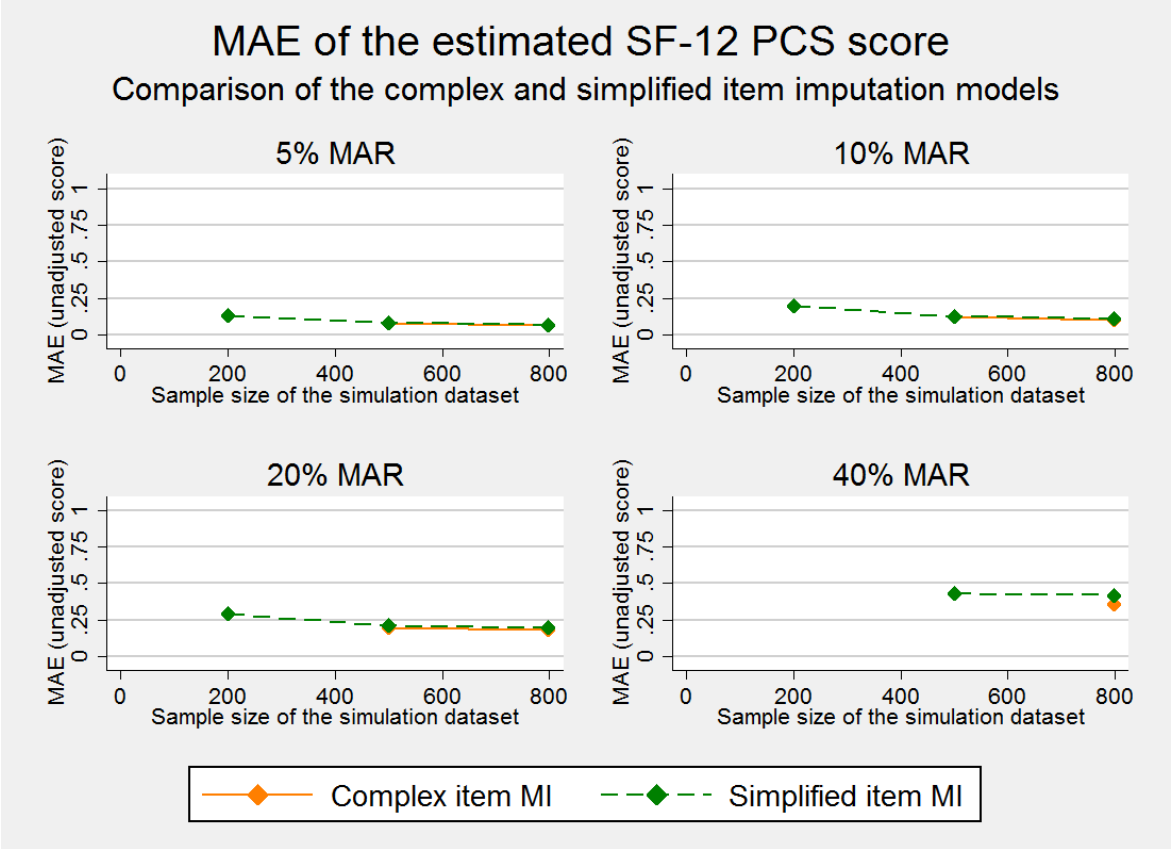


Figure 23: MAE in the SF-12 PCS score estimates – comparing the complex and simplified item imputation model

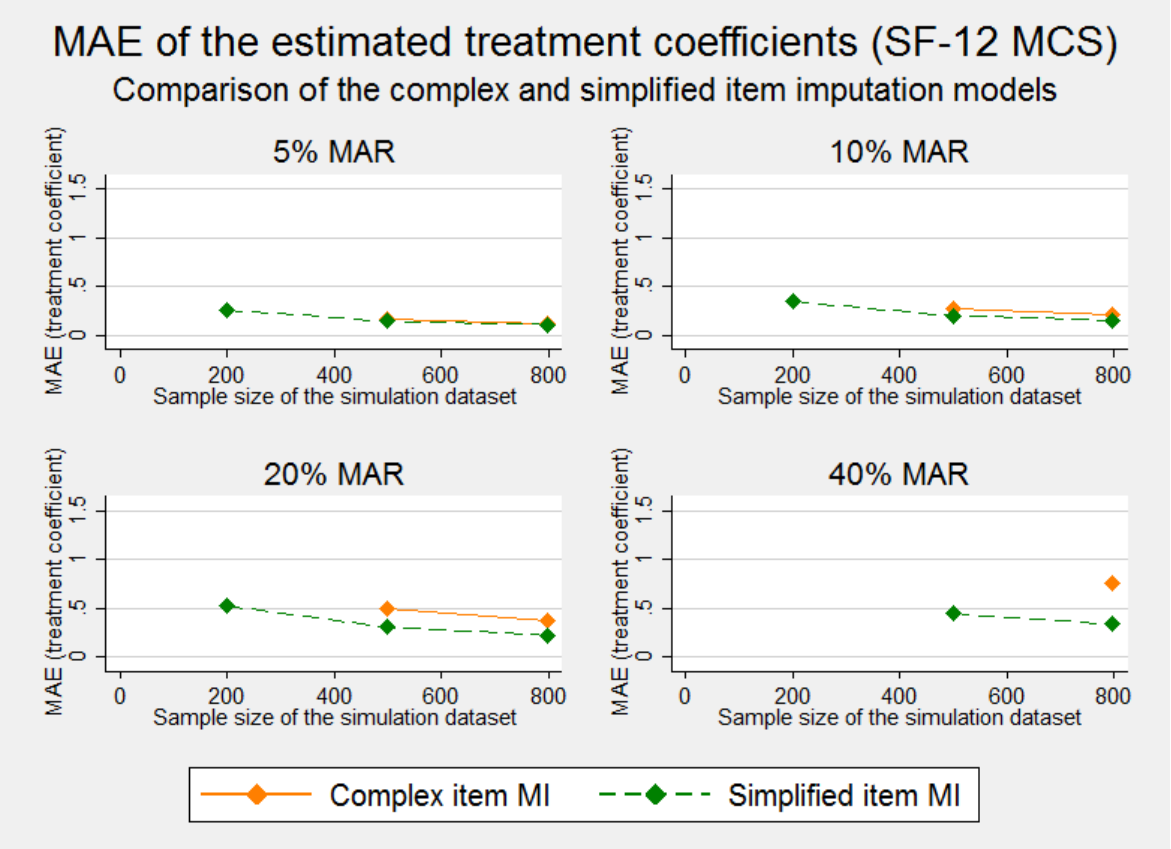


Figure 24: MAE in the treatment coefficient estimates using the imputed SF-12 MCS score as the outcome variable in the regression model – comparing the complex and simplified item imputation model

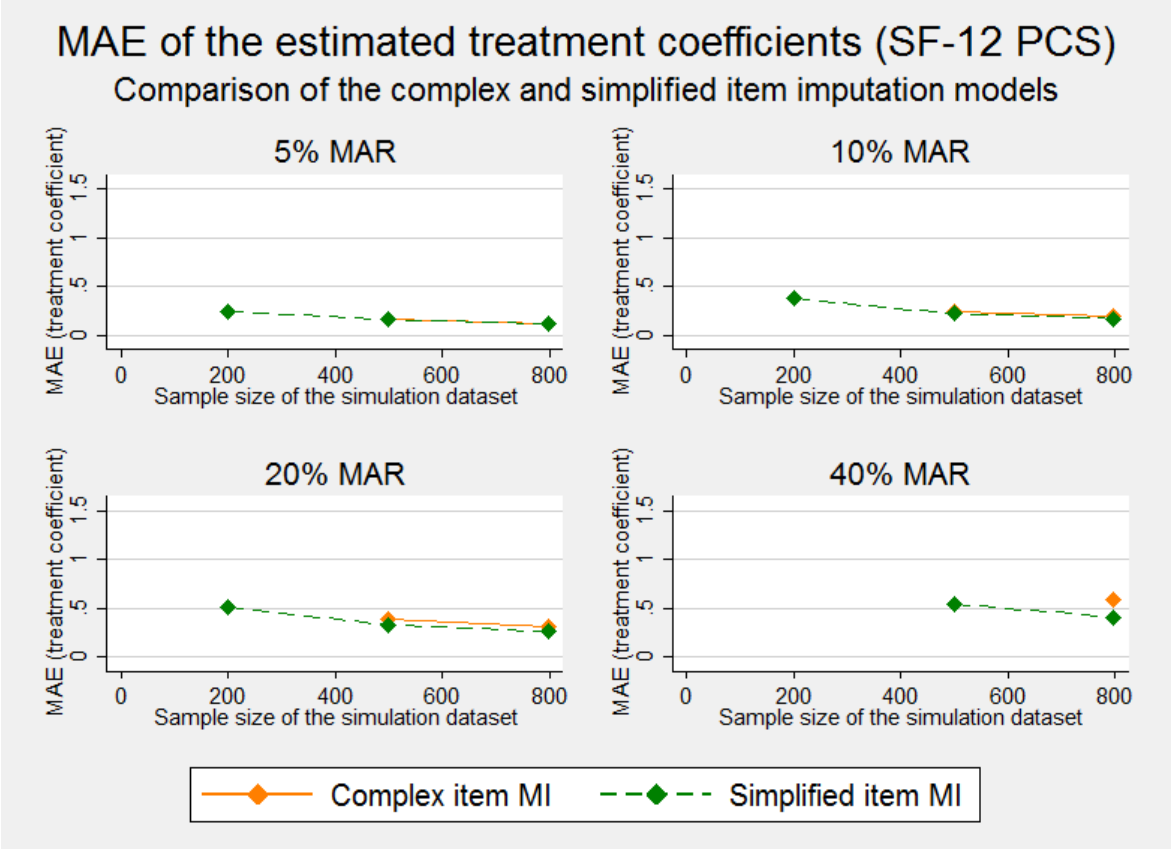


Figure 25: MAE in the treatment coefficient estimates using the imputed SF-12 PCS score as the outcome variable in the regression model – comparing the complex and simplified item imputation model

Appendix 9: Search strategy used to identify relevant literature on approaches to handle longitudinal missing data

PubMed was searched to identify publications that considered the handling of missing data in longitudinal datasets. The following search strategy was used, and rerun in October 2016:

PubMed Search strategy used:

((((missing*[Title/Abstract] OR incomplete[Title/Abstract] OR attrition[Title/Abstract] OR drop-out[Title/Abstract] OR drop-out[Title/Abstract])) AND (longitudinal*[Title/Abstract] OR repeated[Title/Abstract]))) AND ((baye*[Title/Abstract] OR imputation[Title/Abstract] OR "maximum likelihood"[Title/Abstract] OR "pattern mixture"[Title/Abstract] OR "selection model"[Title/Abstract]))

The search identified a total of 584 results. 54% of the identified studies were deemed to be not relevant to this chapter, as they were study reports, rather than methodological evaluations or assessments of statistical approaches for handling missing longitudinal data, or referred to unrelated methodology. A further 5% of publications focussed on Bayesian statistics, which were not further discussed in Chapter 5.

The remaining studies (41%) often matched more than one of the following categories, but were categorised as follows, in line with the most prominent feature of the publications:

- Simulation studies (8%)
- Papers considering sensitivity analysis or MNAR data (9%)
- Reviews (3%)
- Case studies (21%)

Many of the identified papers considered single imputation techniques, including LOCF, which are not considered in this chapter, as they are known to be inferior to the approaches discussed in the chapter, i.e. ML, MI and IPW.

Publications that were most representative of the research on statistical methods for handling missing longitudinal data are referenced in Chapter 5. These were chosen because they assessed the comparative performance of different approaches, or summarised existing research.

Google Scholar was used to identify publications that were cited by, as well as cited the relevant papers to ensure that a thorough overview of the current methodological background on the handling of missing longitudinal data could be provided in this chapter.

Appendix 10: Stata code used in the 'motivating example' case study

This appendix shows the Stata code used to obtain the results presented in the motivating example. Global macros are used to link to the location where the data is stored; the data set used in this analysis has been described in the main text.

```
*set up global macros to link to stored data
global temp "P:\Research datasets\KAT\Stata data\Temporary"

*******************************************************************************
*** ML repeated measures approach                                           ***
*******************************************************************************
use "$temp/LongMethExplore_analysis", clear

*check numbers in trial population:
misstable patterns oks25 oks100 oks200 oks300 oks400 oks500, freq
misstable patterns oks25 oks100 oks200 oks300 oks400 oks500
*14 individuals without any follow-up data - this explains the difference in
* "number of groups"

*adjust for same variables as the analysis model used in the simulations
reshape long oks, i(id) j(time)

*Analaysis model:
mixed oks i.comp_b_alloc oks_bl b100i.time i.sex age  || id: time , ///
 mle  cov(unstructured) stddev


*******************************************************************************
*** MI approach                                                             ***
*******************************************************************************
use "$temp/LongMethExplore_analysis", clear

*Set data to be mi*
*mi set flong
mi set wide
*need to double-check later if any of them need to be transformed
mi register regular comp_b_alloc age sex diseaseplace bmi ASAGrade ///
  oks_bl site_size OpComps AlloProc

*Register Imputed values
mi register imp oks25 oks100 oks200 oks300 oks400 oks500
mi impute chained (pmm, knn(1)) ///
 oks25 oks100 oks200 oks300 oks400 oks500 = ///
 i.comp_b_alloc oks_bl age i.sex ///
 bmi i.ASAGrade i.site_size ///
 i.OpComps i.AlloProc, ///
 add(25) by(comp_b_alloc)

sort id _mi_m

mi reshape long oks, i(id) j(time)
mi estimate: mixed oks i.comp_b_alloc oks_bl b100i.time i.sex age ///
 || id: time , mle  cov(unstructured)

*******************************************************************************
*** IPW approach                                                            ***
*******************************************************************************
```

```
use "$temp/LongMethExplore_analysis", clear

*firstly, generate probability of participants having complete fup data
gen full_fup = 1 if oks_calc_3m == 1 & oks_calc_1yr == 1 & oks_calc_2yr == 1 ///
 & oks_calc_3yr == 1 & oks_calc_4yr == 1 & oks_calc_5yr == 1
replace full_fup = 0 if oks_calc_3m == 0 | oks_calc_1yr == 0 |  ///
 oks_calc_2yr == 0 | oks_calc_3yr == 0 | oks_calc_4yr == 0 | oks_calc_5yr == 0

*compare treatment allocation in full data set and IPW subset:
tab comp_b_alloc
tab comp_b_alloc if full_fup == 1

logit full_fup oks_bl age i.sex ///
      bmi i.ASAGrade i.site_size ///
      i.OpComps i.AlloProc
predict pr

*calculate the inverse of the probability of having follow-up data:
gen ipw = 1/pr

reshape long oks, i(id) j(time)

*now apply to the complete cases only - i.e those with complete follow-up data
drop if full_fup == 0
mixed oks i.comp_b_alloc oks_bl b100i.time i.sex age  || id: time , ///
 mle cov(unstructured) stddev pweight(ipw) vce(robust)
```

Appendix 11: Stata code for the generation of missing data within the longitudinal OKS follow-up data

This appendix shows the Stata code used to simulate the probability of data being missing to each of the observations. The code is shown for the simulation of MAR data following the observed missing data mechanism, as well as a 'stronger' MAR mechanism, which gives more weight to variables outside the analysis model. The remainder of the code is in line with the information presented in Chapter 4, and not reproduced here. Global macros are used to link to the location where the data is stored, and local macros are used to store the newly generated datasets intermittently.

11.1 Stata code for the generation of missing data following the observed missing data mechanism

This appendix shows the Stata code used to impose the observed missing data pattern onto the longitudinal follow-up of participants with completely observed data.

```
use "$work/full_oks_sf&eq_`ss_loop'`miss_loop'", replace

** IMPOSING MISSING DATA ON FULL DATASET
// 1. Generate a random variable, X, from the uniform distribution (i.e. U[0,1]).
// Then assign the patients to 1 of the observed missing patterns by
// comparing X with the cumulative probabilities
// observed for each MD pattern:
/* simplify use the following pattern for missing data at the different
   follow-up time points
Missingness pattern              Total  True %    %used in sim     Cumulative%
No follow-up data available      62     13.51%       22.06%        22.06%
Only three month data missing    49     10.68%       17.44%        39.50%
Only five year data missing      46     10.02%       16.37%        55.87%
Data available to year one       34     7.41%        12.10%        67.97%
Data available to year two       26     5.66%        9.25%         77.22%
Only four year data missing      23     5.01%        8.19%         85.41%
Only three year data missing     22     4.79%        7.83%         93.24%
Data available to year three     19     4.14%        6.76%         100.00%
*/

local seed2 =`c'*(`c'+ 35211)
set seed `seed2'
gen x = uniform()
gen mpattern = 1 if x>=0 & x<=0.2206
```

```
replace mpattern = 2 if x>0.2206 & x<=0.3950
replace mpattern = 3 if x>0.3950 & x<=0.5587
replace mpattern = 4 if x>0.5587 & x<=0.6797
replace mpattern = 5 if x>0.6797 & x<=0.7722
replace mpattern = 6 if x>0.7722 & x<=0.8541
replace mpattern = 7 if x>0.8541 & x<=0.9324
replace mpattern = 8 if x>0.9324 & x<=1
*percentages of MD pattern when taking into account only participants with the
* 8 most common MD pattern

// 2.   Create a linear score for all patients using the beta coefficients in
// Table 2 according to the missing pattern assigned

*build logistic regression models to decide on the factors (and their magnitude)
* to be included into the MAR mechanism
*HERE: COMPARE THOSE WHO FALL IN THE RELEVANT MD PATTERN ONLY TO THOSE WITH
* NO MISSING FUP DATA
gen lscore =       -3.853986 -.0956438*comp_b_alloc -.0965764*oks_bl  ///
+.0351526*age -.4247348*sex -.0138832*bmi -.1847268*asa1 -.6939582*asa3 ///
+.5949436*site_size_m +       0*site_size_l +.9036195*OpComp_sim ///
- 1.083149*AlloProc if mpattern==1
 replace lscore =  -6.084773 +.311389 *comp_b_alloc +.0297232*oks_bl  ///
+.0264918*age +.4605087*sex +.0696749*bmi +.5884773*asa1 -.2168361*asa3 ///
-.905378 *site_size_m -2.089013*site_size_l -.4155457*OpComp_sim ///
-1.386952*AlloProc if mpattern==2
 replace lscore =  -4.834719 -.0181203*comp_b_alloc -.019513 *oks_bl  ///
+.0011325*age +.1987929*sex +.0429858*bmi -.4421613*asa1 +.2135848*asa3 ///
+.921554 *site_size_m +1.013083*site_size_l +.1958116*OpComp_sim ///
-.1832821*AlloProc if mpattern==3
 replace lscore =  -.9763687 +.1170064*comp_b_alloc -.0824726*oks_bl  ///
+.0170368*age -.3370247*sex -.0910559*bmi +.6720494*asa1 +1.175718*asa3 ///
-.0243828*site_size_m -.5187048*site_size_l +.4715631*OpComp_sim ///
-.0907714*AlloProc if mpattern==4
 replace lscore =  -9.54153  -.5000946*comp_b_alloc -.122736 *oks_bl ///
+.0927622*age -.3297777*sex +.0704717*bmi -.405749 *asa1 -.4241179*asa3 ///
-.4246207*site_size_m -.554603 *site_size_l +.1991558*OpComp_sim ///
+.1689976*AlloProc if mpattern==5
 replace lscore =  -5.966279 -.2545011*comp_b_alloc +.0350849*oks_bl  ///
+.0068547*age -.126535 *sex +.0308315*bmi +1.890293*asa1 +     ///
  0*asa3 +.0676068*site_size_m -.7328085*site_size_l +.8068883*OpComp_sim ///
-1.461439*AlloProc if mpattern==6
 replace lscore =  -5.848963 -.4876504*comp_b_alloc -.0353161*oks_bl  ///
+.0207439*age +.1186497*sex +.0260797*bmi +.6771494*asa1 +.9572947*asa3 ///
-.0789474*site_size_m +       0*site_size_l +.5359108*OpComp_sim ///
-.1063515*AlloProc if mpattern==7
 replace lscore =  -9.155526 -.0802481*comp_b_alloc -.0004742*oks_bl  ///
+.0915319*age -1.003634*sex -.0317177*bmi -.0287499*asa1 -.3722173*asa3 ///
-1.692288*site_size_m -.7686751*site_size_l +.1089205*OpComp_sim ///
+.887177* AlloProc if mpattern==8
 *updated based on logistic regression
```

11.2 Stata code for the generation of missing data following a 'stronger' missing data mechanism

This appendix shows the regression model determining the probability of participants having missing data under the 'stronger' missing data mechanism.

```
*for this scenario, devide predictors of missingness by 2 if in analysis model,
*and times by 3 if not in analysis model
gen lscore =       0.5*-3.853986 -0.5*.0956438*comp_b_alloc ///
-0.5*.0965764*oks_bl  +0.5*.0351526*age -0.5*.4247348*sex -3*.0138832*bmi ///
-3*.1847268*asa1 -3*.6939582*asa3 +3*.5949436*site_size_m +      ///
 0*site_size_l   +3*.9036195*OpComp_sim -3*1.083149*AlloProc if mpattern==1
 replace lscore =  0.5*-6.084773 +0.5*.311389 *comp_b_alloc ///
+0.5*.0297232*oks_bl +0.5*.0264918*age +0.5*.4605087*sex ///
+3*.0696749*bmi +3*.5884773*asa1 ///
-3*.2168361*asa3 -3*.905378 *site_size_m -3*2.089013*site_size_l ///
-3*.4155457*OpComp_sim -3*1.386952*AlloProc if mpattern==2
 replace lscore =  0.5*-4.834719 -0.5*.0181203*comp_b_alloc ///
-0.5*.019513 *oks_bl  +0.5*.0011325*age +0.5*.1987929*sex ///
+3*.0429858*bmi -3*.4421613*asa1 +3*.2135848*asa3 ///
+3*.921554 *site_size_m +3*1.013083*site_size_l ///
+3*.1958116*OpComp_sim -3*.1832821*AlloProc if mpattern==3
 replace lscore =  0.5*-.9763687 +0.5*.1170064*comp_b_alloc ///
-0.5*.0824726*oks_bl  +0.5*.0170368*age -0.5*.3370247*sex ///
-3*.0910559*bmi +3*.6720494*asa1 +3*1.175718*asa3 -3*.0243828*site_size_m ///
-3*.5187048*site_size_l +3*.4715631*OpComp_sim -3*.0907714*AlloProc ///
if mpattern==4
 replace lscore =  0.5*-9.54153  -0.5*.5000946*comp_b_alloc ///
-0.5*.122736 *oks_bl  +0.5*.0927622*age -0.5*.3297777*sex ///
+3*.0704717*bmi -3*.405749 *asa1 -3*.4241179*asa3 -3*.4246207*site_size_m ///
-3*.554603 *site_size_l +3*.1991558*OpComp_sim +3*.1689976*AlloProc ///
if mpattern==5
 replace lscore =  0.5*-5.966279 -0.5*.2545011*comp_b_alloc ///
+0.5*.0350849*oks_bl  +0.5*.0068547*age -0.5*.126535 *sex +3*.0308315*bmi ///
+3*1.890293*asa1 +3*       0*asa3 +3*.0676068*site_size_m ///
-3*.7328085*site_size_l +3*.8068883*OpComp_sim -3*1.461439*AlloProc ///
if mpattern==6
 replace lscore =  0.5*-5.848963 -0.5*.4876504*comp_b_alloc ///
-0.5*.0353161*oks_bl  +0.5*.0207439*age +0.5*.1186497*sex ///
+3*.0260797*bmi +3*.6771494*asa1 +3*.9572947*asa3 ///
-3*.0789474*site_size_m +       0*site_size_l +3*.5359108*OpComp_sim ///
-3*.1063515*AlloProc if mpattern==7
 replace lscore =  0.5*-9.155526 -0.5*.0802481*comp_b_alloc ///
-0.5*.0004742*oks_bl  +0.5*.0915319*age -0.5*1.003634*sex -3*.0317177*bmi - ///
3*.0287499*asa1 -3*.3722173*asa3 -3*1.692288*site_size_m ///
-3*.7686751*site_size_l +3*.1089205*OpComp_sim +3*.887177* AlloProc ///
if mpattern==8
 *updated based on logistic regression
```

11.3 Stata code for the assigning participants to their drop-out pattern

This appendix shows the Stata code used to assign participants to their drop-out patterns in the simulation scenario considering monotone missingness only.

```
gen mpattern = 1 if x>=0 & x<=0.1667
replace mpattern = 2 if x>0.1667 & x<=0.3334
replace mpattern = 3 if x>0.3334 & x<=0.5001
replace mpattern = 4 if x>0.5001 & x<=0.6668
replace mpattern = 5 if x>0.6668 & x<=0.8335
replace mpattern = 6 if x>0.8335 & x<=1
```

Appendix 12: MI and IPW models used in Chapter 5

This appendix shows the Stata code for the MI models and generation of the weights for the IPW models. Identical coding is used for the simulations using the observed MAR mechanism, a 'stronger' MAR mechanism, and the simulation using data with a simulated treatment effect. The code is extended for the simulation study that also utilises SF-12 (MCS and PCS scores) and EQ-5D-3L data.

Code for the MI model excluding SF-12 and EQ-5-3L

```
cap:  mi impute chained (pmm, knn(1)) ///
 miss_oks25 miss_oks100 miss_oks200 miss_oks300 miss_oks400 miss_oks500 = ///
 oks_bl age i.sex ///
 bmi i.ASAGrade i.site_size ///
 i.OpComp_sim i.AlloProc, ///
 add(`miss_perc') by(comp_b_alloc)
```

Code for the MI model excluding SF-12 and EQ-5-3L

```
cap:  mi impute chained (pmm, knn(1)) ///
 miss_oks25 miss_oks100 miss_oks200 miss_oks300 miss_oks400 miss_oks500 = ///
 oks_bl age i.sex ///
 bmi i.ASAGrade i.site_size ///
 i.OpComp_sim i.AlloProc ///
 AGG_PHYS0_full AGG_PHYS25_full AGG_PHYS100_full AGG_PHYS200_full ///
 AGG_PHYS300_full AGG_PHYS400_full AGG_PHYS500_full eq_ind0_full ///
 eq_ind25_full eq_ind100_full eq_ind200_full eq_ind300_full ///
 eq_ind400_full eq_ind500_full, ///
 add(`miss_perc') by(comp_b_alloc)
```

Code for the IPW weights excluding SF-12 and EQ-5-3L

```
use "$work/miss_oks`ss_loop'`miss_loop'", clear
* for testing: use "$work/miss_oks_test", clear
*firstly, generate probability of participants having complete fup data
gen full_fup = 1 if miss_oks25 != . & miss_oks100 != . & miss_oks200 != . ///
 & miss_oks300 != . & miss_oks400 != . & miss_oks500 != .
replace full_fup = 0 if miss_oks25 == . | miss_oks100 == . |  ///
 miss_oks200 == . | miss_oks300 == . | miss_oks400 == . | miss_oks500 == .

tab full_fu

*sometimes, not even this model converges:
cap: logit full_fup oks_bl age i.sex ///
      bmi i.ASAGrade i.site_size ///
         i.comp_b_alloc ///
      i.OpComp_sim i.AlloProc
predict pr
gen ipw = 1/pr
```

Code for the IPW weights including SF-12 and EQ-5-3L

```
*IPW:
```

```
use "$work/miss_oks_sf&eq_`ss_loop'`miss_loop'", clear
* for testing: use "$work/miss_oks_test", clear
*firstly, generate probability of participants having complete fup data
gen full_fup = 1 if miss_oks25 != . & miss_oks100 != . & miss_oks200 != . ///
 & miss_oks300 != . & miss_oks400 != . & miss_oks500 != .
replace full_fup = 0 if miss_oks25 == . | miss_oks100 == . |  ///
 miss_oks200 == . | miss_oks300 == . | miss_oks400 == . | miss_oks500 == .

cap: logit full_fup oks_bl age i.sex ///
      bmi i.ASAGrade i.site_size ///
      i.OpComp_sim i.AlloProc   ///
         i.comp_b_alloc ///
         AGG_PHYS500_full eq_ind500_full
predict pr
gen ipw = 1/pr
```

Appendix 13: Feasibility of the MI and IPW approaches – instances where no valid results could be obtained

This appendix provides additional information on the feasibility of the MI and IPW models.

13.1 Feasibility of the MI models

In this appendix, the percentages of MI models that did not converge are summarised for the different simulation scenarios.

Table 10: Percentage of simulations for which no valid results for the MI approach could be obtained – simulation of a observed MAR mechanism

Percentage of participants with simulated missing data	Sample size				
	100	250	500	750	983
10%	0%	0%	0%	0%	0%
20%	0%	0%	0%	0%	0%
30%	0%	0%	0%	0%	0%
40%	0%	0%	0%	0%	0%
50%	0%	0.1%	0%	0%	0%
60%	0%	0%	0%	0%	0%

Table 11: Percentage of simulations for which no valid results for the MI approach could be obtained – simulation utilising the observed MAR mechanism and data with a simulated five point treatment difference

Percentage of participants with simulated missing data	Sample size				
	100	250	500	750	983
10%	0%	0%	0%	0%	0%
20%	0%	0%	0%	0%	0%
30%	0%	0%	0%	0%	0%
40%	0%	0%	0%	0%	0%
50%	0%	0%	0%	0%	0%
60%	0%	0%	0%	0%	0%

Table 12: Percentage of simulations for which no valid results for the MI approach could be obtained – simulation of a stronger MAR mechanism

Percentage of participants with simulated missing data	Sample size				
	100	250	500	750	983
10%	0%	0%	0.4%	2.7%	0%
20%	0%	0%	5.5%	4.0%	0%
30%	0%	0%	0%	0%	0%
40%	0%	0%	0.1%	0%	0%
50%	0%	0%	0%	0%	0%
60%	0%	0%	0.1%	0.8%	0%

Table 13: Percentage of simulations for which no valid results for the MI approach could be obtained – simulation utilising information from the SF-12 and EQ-5D-3L

Percentage of participants with simulated missing data	Sample size				
	100*	250	500	750	983
10%	0%	0%	0.9%	1.3%	0%
20%	0%	0%	1.3%	0%	0%
30%	0%	0.1%	0%	0.1%	0%
40%	0%	0%	0%	2.3%	1.3%
50%	0%	0%	0%	4.6%	2.6%
60%	0%	0%	6.2%	2.9%	0%

**these MI models were simplified, i.e. randomised treatment was included as a covariate (instead of running separate models by treatment arm)*

Table 14: Percentage of simulations for which no valid results for the MI approach could be obtained – simulations considering drop-outs only and including additional data from the SF-12 and EQ-5D

Percentage of participants with simulated missing data	Sample size				
	100*	250	500	750	983
10%	0%	0%	0%	0%	0%
20%	0%	0%	0%	0%	0%
30%	0%	0%	0%	0%	0%
40%	0%	0%	0.1%	0%	0.1%
50%	0%	0%	0%	0.2%	0.2%
60%	0%	0%	0.1%	0%	0%

**these MI models were simplified, i.e. randomised treatment was included as a covariate (instead of running separate models by treatment arm)*

13.2 Feasibility of the IPW models

The main text of this thesis, Chapter 5, tabulates the cases in which no valid results could be obtained for the IPW models for the simulation using the observed data and observed missing data mechanism. Similar summaries for the other simulation studies are shown below.

Table 15: Percentage of simulations for which no valid results for the IPW approach could be obtained – simulation utilising the observed MAR mechanism and data with a simulated five point treatment difference

Percentage of participants with simulated missing data	Sample size				
	100	250	500	750	983
10%	20.8%	1.5%	0%	0%	0%
20%	3.2%	0%	0%	0%	0%
30%	0.8%	0%	0%	0%	0%
40%	0.1%	0%	0%	0%	0%
50%	0.1%	0%	0%	0%	0%
60%	0.5%	0%	0%	0%	0%

Table 16: Percentage of simulations for which no valid results for the IPW approach could be obtained – simulation of a stronger MAR mechanism

Percentage of participants with simulated missing data	Sample size				
	100	250	500	750	983
10%	19.7%	1.3%	0%	0%	0%
20%	3.2%	0%	0%	0.1%	0%
30%	0.5%	0%	0%	0%	0.1%
40%	0%	0%	0%	0%	0%
50%	0.1%	0%	0%	0%	0%
60%	0.9%	0%	0%	0%	0%

Table 17: Percentage of simulations for which no valid results for the IPW approach could be obtained – simulations including additional data from the SF-12 and EQ-5D-3L

Percentage of participants with simulated missing data	Sample size				
	100	250	500	750	983
10%	21.3%	1.5%	0.1%	0%	0%
20%	3.2%	0%	0%	0%	0%
30%	0.8%	0%	0%	0%	0%
40%	0.2%	0%	0%	0%	0%
50%	0.2%	0%	0%	0%	0%
60%	0.3%	0%	0.1%	0.1%	0%

Table 18: Percentage of simulations for which no valid results for the IPW approach could be obtained – simulations considering drop-outs only and including additional data from the SF-12 and EQ-5D-3L

Percentage of participants with simulated missing data	Sample size				
	100	250	500	750	983
10%	15.8%	0.8%	0%	0%	0%
20%	2.4%	0.1%	0%	0%	0%
30%	0.3%	0%	0%	0%	0%
40%	0.1%	0%	0%	0%	0%
50%	0.5%	0%	0%	0%	0%
60%	1.8%	0.1%	0%	0%	0%

Appendix 14: MAE plots for the comparison of statistical methods to handle missing PROMs data in a longitudinal setting

This appendix contains graphs showing the MAE of the different analysis approaches to handle missing PROMs outcome data in a longitudinal setting, i.e. ML, MI and IPW. The graphs aim to supplement the assessment of the approaches using the RMSE, which was presented in Chapter 5.

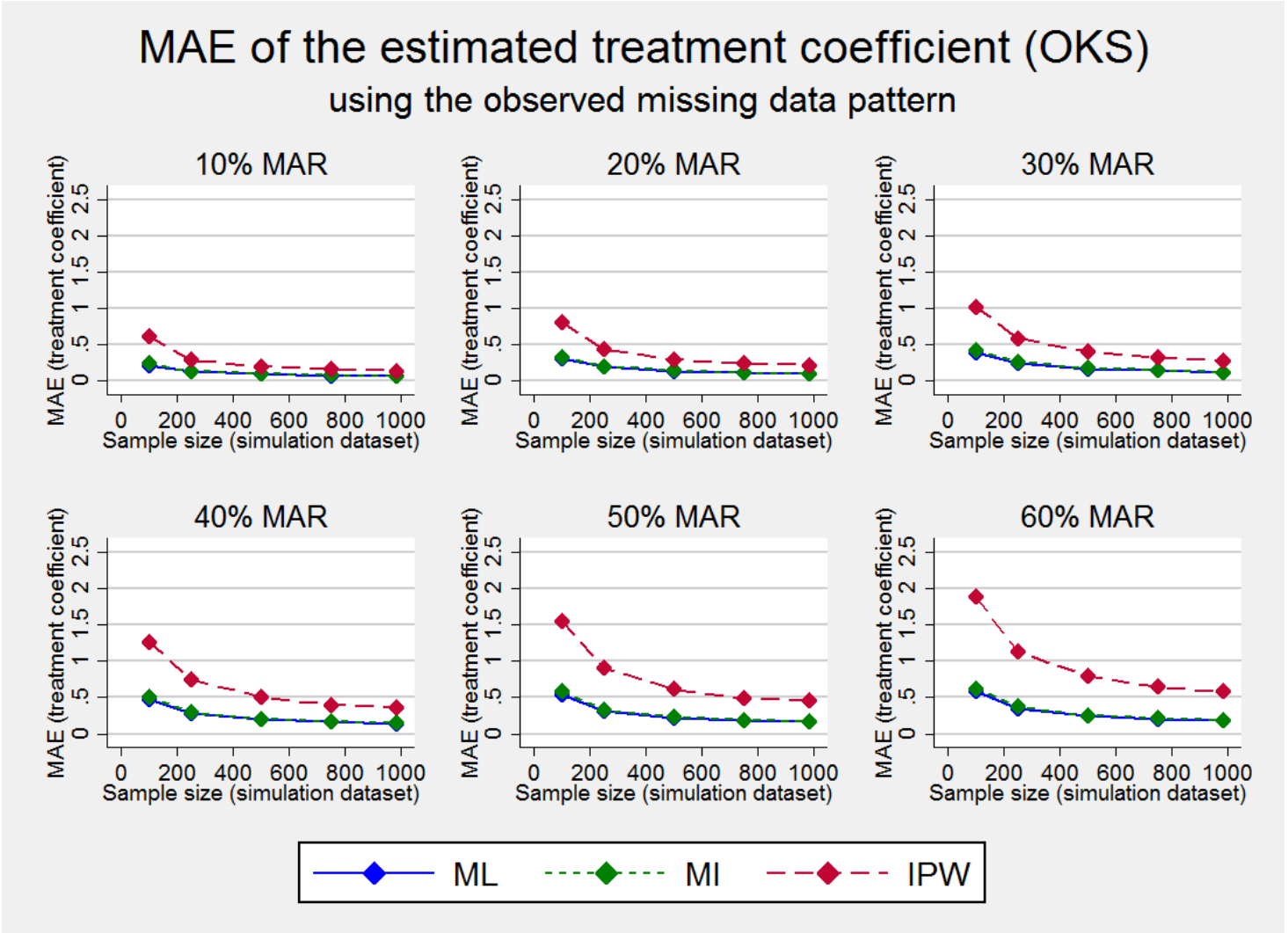


Figure 26: MAE of the estimated treatment coefficient – simulations using the observed missing data pattern

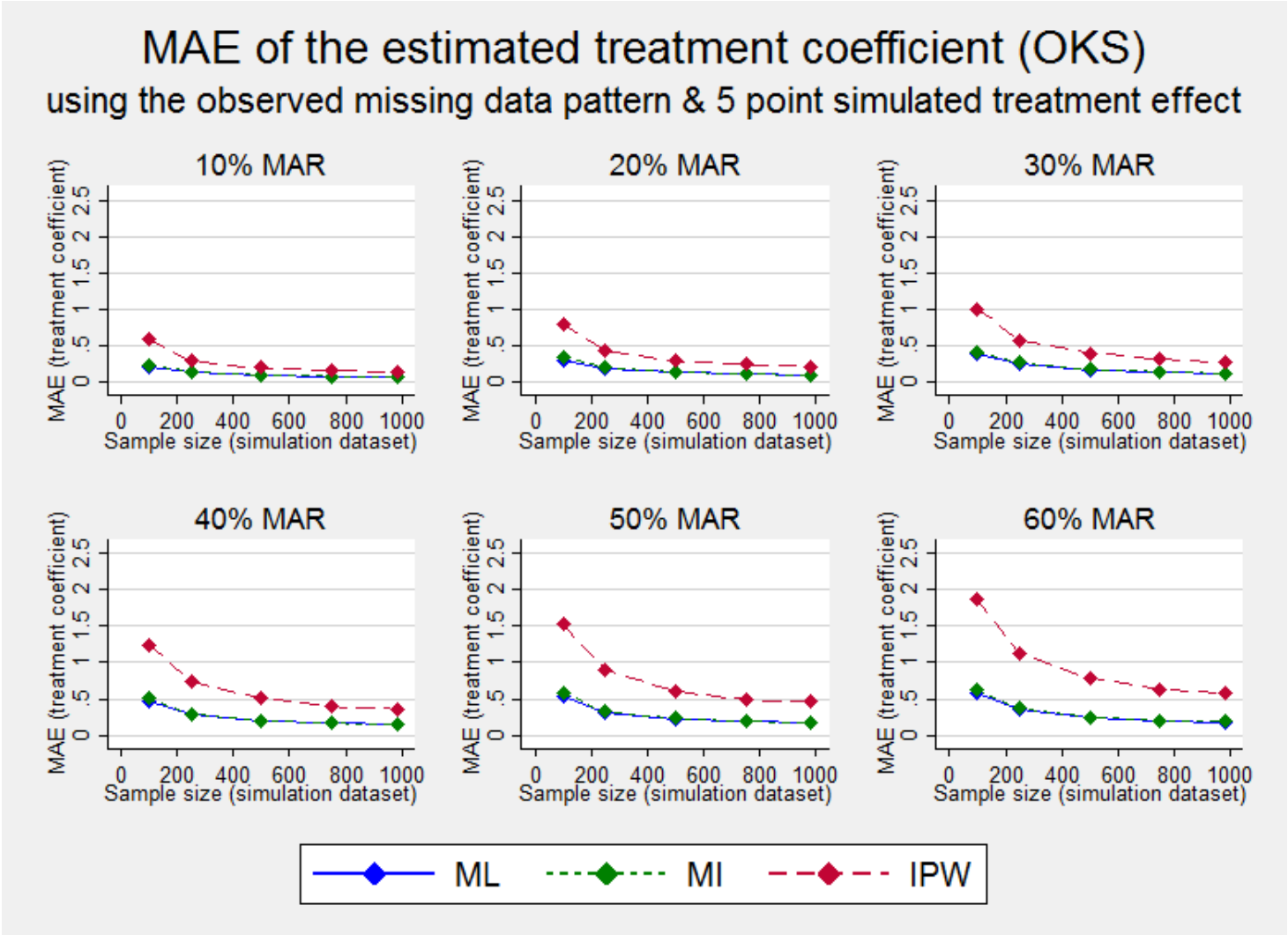


Figure 27: MAE of the estimated treatment coefficient – simulations using the observed missing data pattern and a five point treatment effect

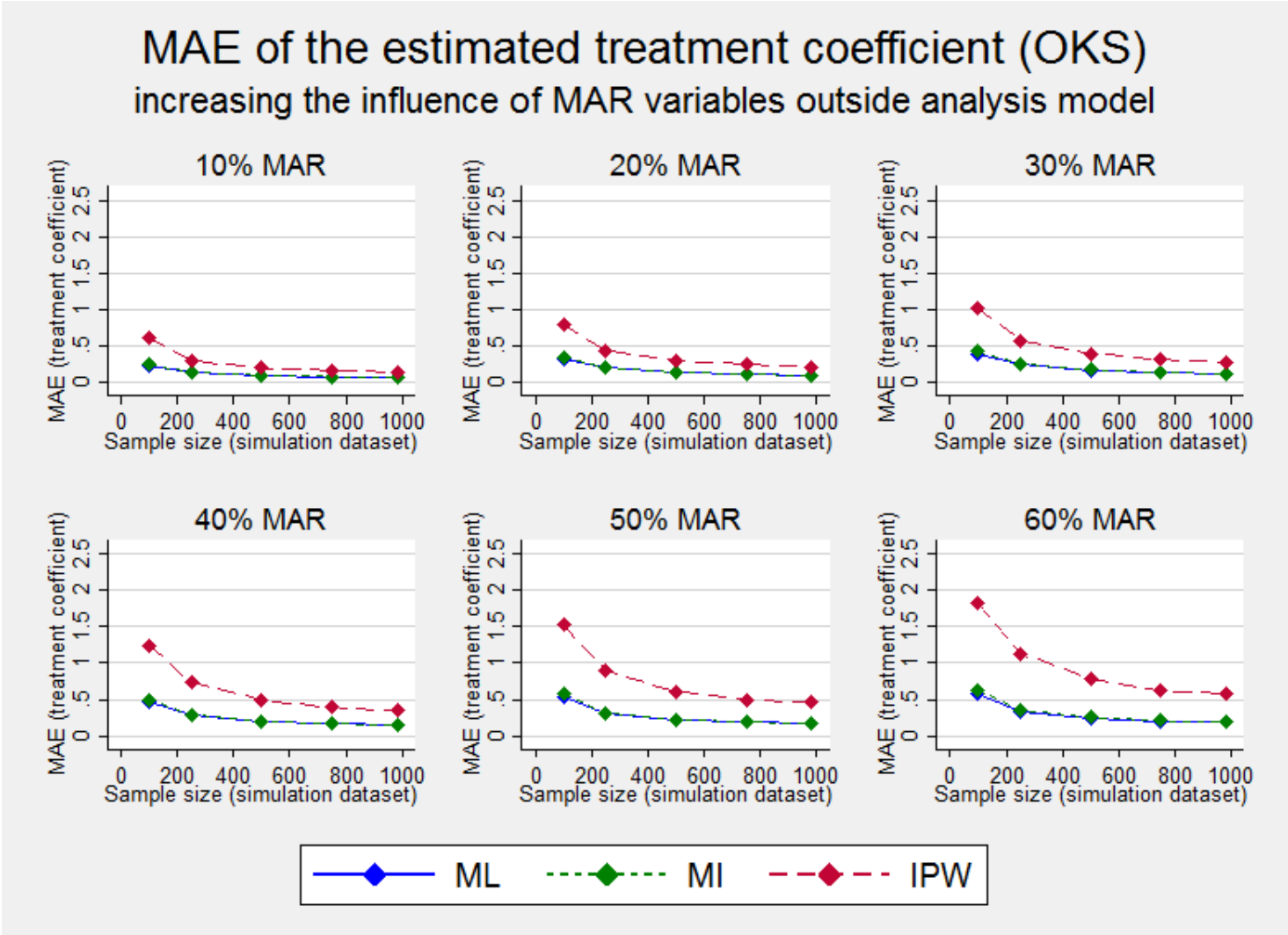


Figure 28: MAE of the estimated treatment coefficient – simulations using the observed missing data pattern and a five point treatment effect

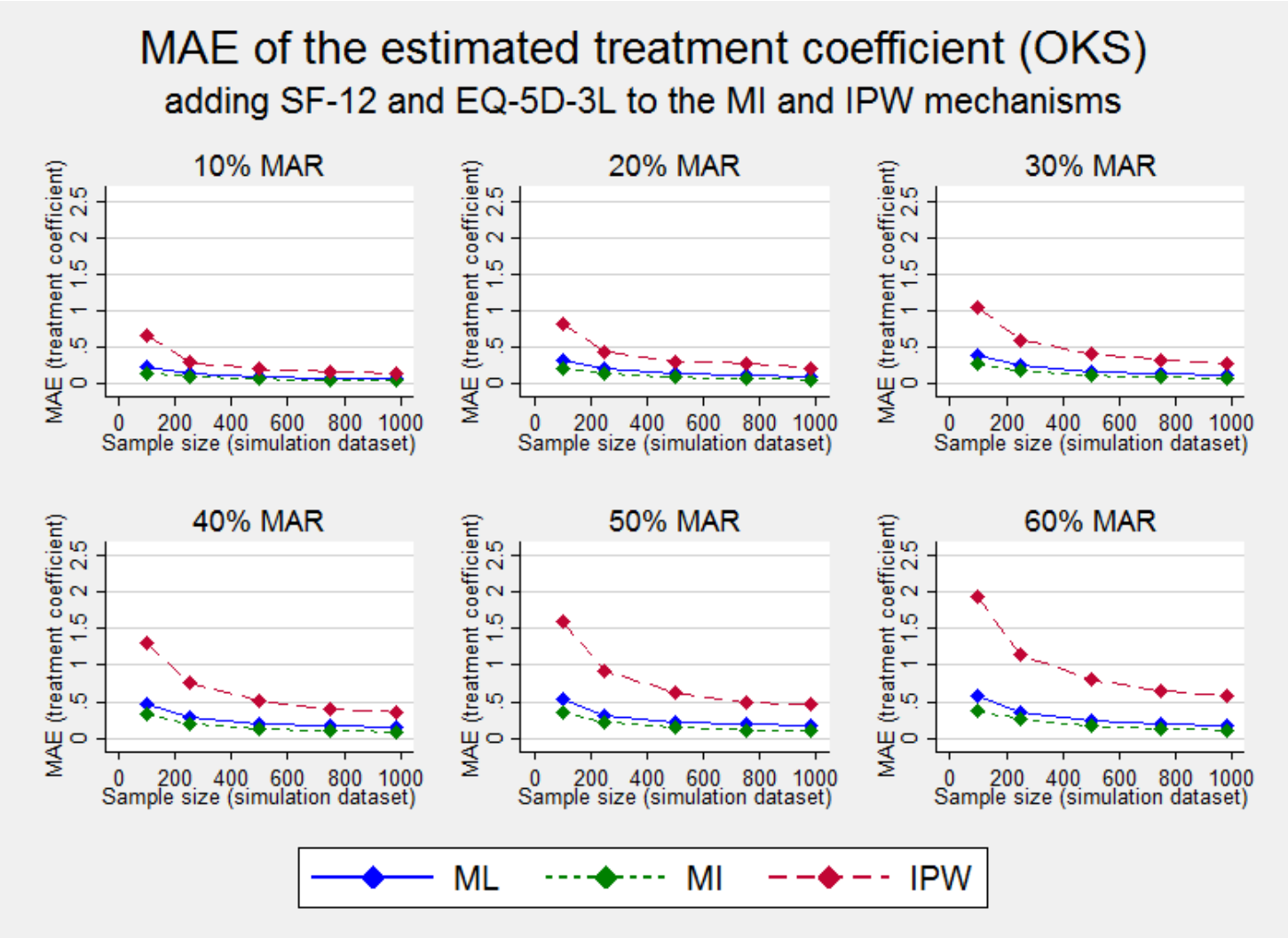


Figure 29: MAE of the estimated treatment coefficient – simulations adding the SF-12 and EQ-5D-3L to the MI and IPW mechanisms

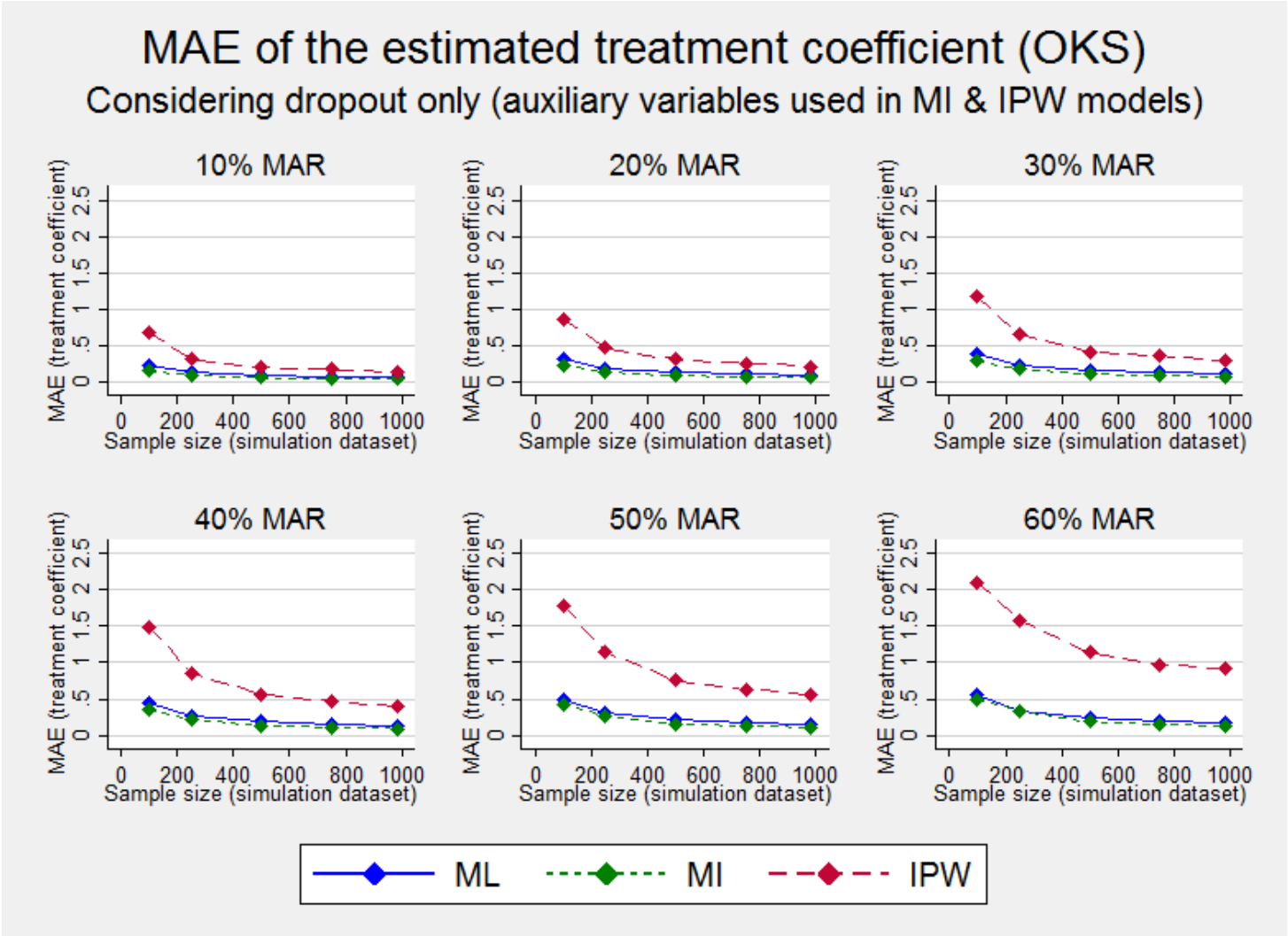


Figure 30: MAE of the estimated treatment coefficient – simulations adding the SF-12 and EQ-5D-3L to the MI and IPW mechanisms while considering dropout only

Appendix 15: Search strategy used to identify relevant literature on sensitivity analysis

PubMed was searched to identify publications that considered sensitivity analysis for missing/incomplete data for RCTs. Broad search terms were used to avoid missing potentially relevant articles, although this resulted in a large number of papers being identified that were not relevant in this context. The following search strategy was used, and rerun in October 2016:

> **PubMed Search strategy used:**
> ((((missing[Title/Abstract] OR incomplete[Title/Abstract])) AND sensitivity[Title/Abstract])) AND trial*[Title/Abstract]

The search identified a total of 620 results. Of these, 81% were excluded as they did not focus on the methodological aspects of sensitivity analyses, instead comprising RCT publications including sensitivity analyses (71%), focussing on meta-analyses (6%), or were spuriously picked up in the literature search due to the use of sensitivity in a different context or not in-human studies (4%).

The remaining 19% of identified publications were classified as follows:

- 4% were reviews, either assessing current practice in the handling of missing data, or comparing different analysis approaches for handling missing data. Sensitivity analysis was mentioned as an important component of such analyses, but no detailed discussions of such analyses were provided
- 10% were classified as case studies, demonstrating the application of approaches to sensitivity analysis

- 1% of publications considered binary endpoints or time to event outcomes, which were not relevant for Chapter 6.
- 4% of papers were classed as methodological work on sensitivity analysis, including MI, pattern mixture models, weighting approaches, Bayesian methods, tipping point analyses, as well as comparisons between different approaches.

Publications that were most representative of the research on sensitivity analysis were referenced in Chapter 5. These were chosen because they assessed the comparative performance of different approaches, were considered key papers for this research, or because they picked up on previously published publications.

Google Scholar was used to identify publications that were cited by, as well as cited the relevant papers to ensure that a thorough overview of the current methodological background on the handling of missing longitudinal data could be provided in this chapter.

Appendix 16: Stata code for the generation of missing data within the longitudinal OKS follow-up data

This appendix shows the Stata code used to implement the sensitivity analyses discussed in Chapter 6.

```
*setting global directories
global data "P:\Research datasets\KAT\Stata data\Temporary"
global store "P:\Research datasets\KAT\Output\Simulations"
global graphs "P:\Chapters\Chapter 6 -Sensitivity analyses\Graphs"

*these are the two datasets used in the case study
* use them in turn to run the sensitivity analysis
use "$data/SensAnalysisData_ss200_miss20", clear
*use "$data/SensAnalysisData_ss1000_miss20", clear

*check treatment allocations:
tab comp_b_alloc

*check missing data
gen miss = 1 if oks_miss == .
 replace miss = 0 if oks_miss != .

tab miss comp_b_alloc, col

*CCA
regress oks_miss i.comp_b_alloc oks_bl i.sex age

*change labelling
label define rand 0 "No resurfacing" 1 "resurfacing"
label define rand1 0 "No PR" 1 "PR"
label values comp_b_alloc rand1

rename comp_b_alloc treatment_allocation
label values treatment_allocation rand1

*RCT miss analysis
xi: rctmiss , sens(treatment_allocation) pmmdelta(-5/5) stagger(0.1) ///
 list listopt(sepby(delta)) ///
 color(blue red green) senstype(one): ///
 reg oks_miss treatment_allocation oks_bl i.sex age

graph export "$graphs/SensAn_rctmiss_ss1000.png", replace
graph export "$graphs/SensAn_rctmiss_ss200.png", replace


*******************************************************************************
*******************************************************************************
*Manipulate imputations
use "$data/SensAnalysisData_ss200_miss20", clear
use "$data/SensAnalysisData_ss1000_miss20", clear

*Set data to be mi*
mi set wide
```

```
*need to double-check later if any of them need to be transformed
mi register regular comp_b_alloc age sex diseaseplace height ASAGrade oks_bl ///
 site_size
mi register imp oks_miss
mi impute chained (pmm, knn(1)) oks_miss = ///
 age i.sex i.ASAGrade oks_bl height i.site_size, add(20) by(comp_b_alloc)

mi estimate: regress oks_miss i.comp_b_alloc oks_bl i.sex age
regress oks_miss i.comp_b_alloc oks_bl i.sex age


*now add 5 points to imputation estimates in patella resurfacing group where
* data is missing
forvalues i = 1(1)20 {
*generate the maximum score that can be added to outcomes
gen max`i' = 48 - _`i'_oks_miss
replace max`i' = 5 if max`i' > 5

replace _`i'_oks_miss = _`i'_oks_miss + max`i' ///
 if comp_b_alloc == 1 & _mi_miss == 1

 drop max`i'
}

*analyse manipulated MI results:
mi estimate: regress oks_miss i.comp_b_alloc oks_bl i.sex age
```