DEALING WITH MISSING QUALITY OF LIFE OUTCOME DATA IN CLINICAL TRIALS: THE ROLE OF REMINDERS

A THESIS PRESENTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY AT THE UNIVERSITY OF ABERDEEN

SHONA A FIELDING

BSc (Hons) (Edinburgh) MSc (Reading)

2009

Declaration

This thesis has been written by Shona A Fielding. The work contained in this thesis has been undertaken by the candidate and has not been submitted in any previous application for a degree. Quotations have been distinguished by quotation marks and the sources of information specifically acknowledged.

Shona A Fielding

Acknowledgements

Firstly, I would like to thank my supervisors, Professor Peter Fayers and Dr. Craig Ramsay for their support and guidance over the course of this thesis work. I appreciate the valuable input and expertise Professor Luke Vale provided for the economic aspects of the work. In addition, thanks to Dr. Diane Fairclough for providing her time, hospitality and supervision during a three week visit to Denver, USA. I would like to acknowledge and thank the secretarial staff and the Medical Statistics Team in the Section of Population Health at the University of Aberdeen for their continued support.

None of the work would have been possible without the datasets provided. Therefore, thanks go to the Health Services Research Unit at the University of Aberdeen and in particular, Marion Campbell, Graeme Maclennan, Jonathan Cook, Samantha Wileman, Janice Cruden, Alison Macdonald and Gladys McPherson. Many thanks, to the TOMBOLA group and in particular Seonaidh Cotton and also to Marit Jordhøy for providing the Norwegian Palliative Care trial data.

I would also like to thank Graeme Maclennan, Jonathan Cook and Craig Ramsay for their help in identifying articles for the review of imputation use in clinical trials. I am grateful to family and friends who have provided support throughout. Particularly, my father Professor Antony Fielding for commenting on drafts and my sister Lorna, for checking the grammar. Finally, I would like to thank the Chief Scientist Office for providing the funding for the project.

Summary

Background

Randomised controlled trials (RCTs) are an important way of evaluating healthcare interventions. Missing data are a problem for any outcome, but are particularly prominent for quality of life (QoL) outcomes, as the reason why the data are missing is likely to be related to the QoL itself. Ignoring this could lead to biased results, and the publication of a wrong conclusion about a particular therapy could have disastrous consequences. It is essential that the results of RCTs can be relied upon as ultimately they influence clinical practice.

Objectives

The overall objective of the work was to evaluate different strategies available to deal with missing QoL data to ensure the analysis was optimal. A novel approach using data collected through postal reminders was undertaken. The specific aims were:

1. To provide a general overview of methods for missing data
2. To identify the current practice for dealing with missing QoL outcomes in RCTs
3. To investigate the use of available procedures to identify the missing data mechanism
4. To investigate whether imputation techniques are suitable for dealing with missing QoL data and which methods are most appropriate
5. To investigate alternative analysis strategies for the example RCTs
6. To discuss the economic benefit of different data collection strategies
7. To provide recommendations on how to deal with missing QoL outcomes in RCTs.

Methods

Seven completed trial datasets were used throughout. Each trial administered one or more QoL instruments at two or more follow-up assessments, through postal

questionnaires. A reminder system was employed if the initial questionnaire was not returned within two weeks. This recovered a proportion of data which would otherwise have been missing. This reminder-response data was used to investigate the missing data mechanism and investigate suitable methods of imputation (as the true values were in fact known). The reminder system entails a significant amount of resources. Cost-effectiveness analyses were used to determine if the reminder system and/or imputation were cost-effective in increasing the amount and quality of data available for analysis. Current literature on the subject of missing data relies heavily on simulated data or data removed in a certain way from a complete dataset. A novel approach was used here to investigate ways of dealing with missing data, with the advantage that actual, known data was used.

Results

The mechanism of missing data for the example trials is investigated in chapter five, using the methods described in chapter four. The reminder-responses were utilised in two ways: firstly, to investigate the mechanism of non-response and secondly, to investigate the mechanism behind reminder response. Missing QoL responses were generally found to be missing at random and those participants who provided missing data tended to have poorer observed QoL scores. It was found that ignoring the reminder-responses and investigating the missing data mechanism using only immediate responses may give a distorted view of this mechanism.

One option for handling missing data is imputation. The procedures for this are described in chapter six. Chapter seven carries out these methods on the example datasets. Using the reminder responses, the accuracy (bias and precision) could be assessed as the true values were known (from reminders). Simple imputation was not expected to perform well, since the mechanism of missing data was typically found to not be missing completely at random. This proved to be the case with multiple imputation, which out-performed simple imputation in most situations.

The trial analysis for the example datasets was an analysis of covariance at a single assessment, adjusting for baseline QoL and other patient characteristics. This approach does not make use of the interim responses and is a type of complete-case analysis, that assumes missing completely at random. Chapter eight describes alternatives to this and these are implemented in chapter nine. A repeated measures model which assumed data were missing at random was found to be useful for a number of trials, as a larger number of patients were included in the analysis.

Chapter ten investigates the cost of different data collection strategies, namely the reminder system and/or use of imputation. It was shown that in trials of less than 1000 participants the reminder system is cost-effective and provides a better quality result than imputation. In larger trials, the cost of reminders can be vast as the number of patients involved is larger and the cost-effectiveness of the strategy would be largely down to the proportion of missing data.

Conclusion

When QoL is a major endpoint in a trial it is essential to ensure that missing data are minimised. There is ample evidence that patients with the poorest health status or the lowest QoL are the ones most likely not to respond to questionnaires. This bias can jeopardise the results. The use of a reminder system to collect follow-up outcome data is recommended for the following reasons:

1. Each patient in a trial is costly and a reminder system is a cost-effective use of resources to maintain the sample size.
2. Using reminders to minimize the amount of missing data also reduces the threat of bias.
3. Data collected by reminders enables a more informed selection of imputation methods which again reduces the risk of bias.

The results from the work strongly support the use of reminders in trials where patient-reported outcomes are a major study-endpoint. The statistical analysis of the QoL outcomes should consider the mechanism of missing data, potential

imputation methods and whether a cross-sectional or longitudinal method is most appropriate. The responses by reminder can be used to help with this decision making. In conclusion, there is no single way of dealing with missing data that is applicable in all situations, but the approaches outlined throughout this thesis provide researchers with some tools to develop a strategy to dealing with the problem of missing QoL outcomes in clinical trials.

Table of contents

List of Tables

List of Figures

List of Abbreviations

ABB – Approximate Bayesian Bootstrap

ACMV – available case missing value

AIC – Akaike information criterion

ANCOVA – analysis of covariance

ANOVA – analysis of variance

AR - autoregressive

ASA (grade) – American Knee Society grade

ASHI – Arthritis Specific Health Index

AUC – area under the curve

BCF – baseline carried forward

BMI – body mass index

BMJ – British Medical Journal

BNA - borderline nuclear abnormalities

CBA – cost benefit analysis

CCMV – complete-case missing variable

CEA – cost effectiveness analysis

CHART - Centre for Healthcare Randomised Trials

CI – confidence interval

CLM – conditional linear model

CONSORT – Consolidated Standards of Reporting Trials

CRD – completely random dropout

CS – compound symmetry

CSO – Chief Scientist Office

CUA – cost utility analysis

DCE – discrete choice experiment

EF – emotional functioning

EORTC – European Organisation for the Research and Treatment of Cancer

EQ5D – EuroQoL EQ5D

GORD – gastro-oesophageal reflux disease

HAQ - Health Assessment Questionnaire

HPV – Human Papilloma virus

HRQoL – health related quality of life

HSRU – Health Services Research Unit

I – immediate-responder

ICER – incremental cost-effectiveness ratio

ID – informative dropout

IQR – inter-quartile range

JAMA – Journal of the American Medical Association

KAT – Knee Arthroplasty Trial

LOCF – last observation carried forward

LS – Listing and Schlittgen test

LVCF – last value carried forward

MAR – missing at random

MAVIS – Randomised trial of mineral and vitamin supplementation

MC – Monte Carlo

MCAR – missing completely at random

MCMC – Markov Chain Monte Carlo

MCS – mental component score of SF12/SF36

MI – multiple imputation

ML – maximum likelihood

MLE – maximum likelihood estimates

MW – Mann Whitney

MNAR – missing not at random

N – non-responder

NB – net benefit

NCMV – neighbouring case missing value

NE - northeast

NEJM – New England Journal of Medicine

NPC – Norwegian Palliative Care

NVCB – next value carried backward

NW – northwest

OKS – Oxford Knee Score

OR – odds ratio

PA – pain

PCS - physical component score of SF12/SF36

PF – physical functioning

PMM – predictive mean match

PRISM – Paget's disease: a Randomised Trial of Intensive versus Symptomatic Management

QALY – quality adjusted life years

QLQ – quality of life questionnaire

QoL – quality of life

R – reminder-responder

RCT – randomised controlled trial

RD – random dropout

RECORD – Randomised Evaluation of Calcium and / OR vitamin D

REFLUX – Randomised Evaluation oF Laproscopic sUrgery for refluX

REML – restricted maximum likelihood

RQLS – Reflux quality of life score

SD – standard deviation

SE – standard error

SF12– Short-form 12

SF36 – Short-form 36

SW - southwest

TOMBOLA – Trial Of Management of Borderline and Other Low-grade Abnormal smears

UN – unstructured

VAS – visual analogue scale

WTP – willingness to pay

List of Publications

Published

1. Shona Fielding, Graeme MacLennan, Jonathan A Cook, Craig R Ramsay. 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.

2. Shona Fielding, Peter M Fayers, Alison McDonald, Gladys McPherson, Marion K Campbell for the RECORD study group. Simple imputation methods were inadequate for missing not at random (MNAR) quality of life data. Health and Quality of Life Outcomes 2008, 6: 57.

3. Fielding S, Fayers PM, Loge JH, Jordhøy MS, Kaasa S. Methods for handling missing data in palliative care research. Palliative Medicine 2006; 20: 791-798.

4. Shona Fielding, Peter M Fayers, Craig R Ramsay. Investigating the missingness mechanism in quality of life data: A comparison of approaches. Health and Quality of Life Outcomes 2009; 7: 57.

Submitted

5. S Fielding, PM Fayers, CR Ramsay. Dealing with missing quality of life data: A comparison of imputation methods and modelling approaches. (Clinical Trials, January 2009)

Chapter 1 Introduction and background

1.1 Introduction, aims and objectives

The randomised controlled trial (RCT) is widely recognised as the gold standard design for the evaluation of new healthcare interventions (Pocock 1983). The information gained from trials is optimal when the trial dataset is complete or relatively few data are missing. If large proportions of data are missing then any analysis that ignores the missing data may be biased. Ensuring a dataset is as complete as possible can, however, involve significant resources for the trial team. This has the potential to delay any analysis and implementation of the conclusions into clinical practice. In the case of quality of life (QoL) studies, data are unlikely to be complete because patients are entitled to refuse to respond to questionnaires. To date, little research exists to inform whether missing data in trials do result in biased conclusions and whether the effort put in to collect all possible data, is worth the extra cost. This thesis aims to fill this gap in the current literature and discusses the issues surrounding the analysis of missing quality of life outcomes in RCTs.

Datasets arising from seven completed RCTs are utilised throughout. A feature of all these trials is the use of a reminder system for follow-up questionnaires. Thus, data that were initially classified as 'missing', but subsequently recovered after reminders, are known. Although there is extensive literature about dealing with missing data including methods of imputation and alternative analysis strategies, I am not aware of publications using recovered data to test and validate the statistical procedures. This is probably due to the fact that even where reminder systems are used, few trial organisations implement rigorous recording systems of who did and did not complete questionnaires after a reminder. Hence the example datasets used here, offer a unique opportunity to explore the true impact of missing data.

The work presented in this thesis aims to meet a number of objectives:

- To provide a general overview of methods for missing data
- To identify the current practice for dealing with missing QoL outcomes in RCTs
- To investigate the use of available procedures to identify the missing data mechanism
- To investigate whether imputation techniques are suitable for dealing with missing QoL data and which methods are most appropriate
- To investigate alternative analysis strategies for the example RCTs
- To discuss the economic benefit of different data collection strategies
- To provide recommendations on how to deal with missing QoL outcomes in RCTs

This first chapter provides an initial background to the concepts mentioned above and in particular the RCT, missing data and QoL outcomes. Further details are then provided throughout the remainder of the thesis.

1.2 Randomised controlled trials (RCTs)

Conducting an RCT is considered the most robust method of clinical research and is described as the 'gold standard' in evaluating health care interventions, whether they are drug therapies, surgical interventions or other medical procedures (Pocock 1983). An RCT is a comparative trial where new treatments are compared to a standard treatment (or placebo). Patients are randomised to one of the treatment groups. The randomisation process ensures that the treatment groups are not systematically different, with respect to known and unknown baseline characteristics, and contain a similar number of participants.

To guarantee the randomisation process is not subverted, allocation concealment must be enforced and is usually undertaken using central computer generation. Trials may be open where treatment allocation is known by everybody. Alternatively some degree of blinding can be used. For example, the doctor

knows the treatment, but the patient does not, or neither the doctor nor the patient knows which treatment. The control group may receive the standard treatment or if one is not available then a non-active placebo. For example, in a trial involving a tablet treatment, the control group would receive tablets that cannot be distinguished from the active treatment, except they do not contain the active ingredient.

An RCT collects outcome information usually at baseline and then at some pre-specified times post intervention (e.g. monthly). This may occur at a single follow-up assessment or at several, providing longitudinal information. The follow up information such as QoL data may be collected at a clinic appointment or often through a postal survey. This mode of collection will inevitably lead to a proportion of missing data. Some patients can forget to return the questionnaire, consciously chose not to fill it out or do not receive it due to a change of address. Trial centres often make use of reminders which are sent to the participant if they have not responded to the questionnaire within a particular time frame. The use of a reminder-system at follow-up can help to alleviate some of the problems of missing responses. This is discussed further in later chapters.

1.3 Quality of life

Quality of life (QoL) can mean different things to different people. The World Health Organization declared health to be "a state of complete physical, mental and social well-being, not merely the absence of disease" (World Health Organization 1948). The term "health-related quality of life" (HRQoL) is frequently used to distinguish between the more general QoL and the requirement of clinical trials. Throughout this thesis the abbreviation QoL will be used to imply HRQoL. Different instruments aim to measure different aspects of QoL, but often include general health, physical functioning, emotional functioning, cognitive functioning, role functioning and social well-being and functioning. Some QoL instruments will cover several aspects or dimensions, while others will focus on a single construct like 'pain'.

QoL instruments come in two main forms: generic and disease-specific. Generic instruments are intended for general use, irrespective of illness or condition. They can also be applied to healthy populations. Examples of a generic instrument are the SF-36 (Ware et al. 1993) and EuroQoL - EQ5D (Brooks, R with the EuroQoL Group 1996). The former is an instrument consisting of 36 items, measuring eight QoL dimensions and produces a physical summary and mental summary score. The EuroQoL EQ5D is a shorter five-item instrument measuring overall health status. A disease specific instrument is tailored to the condition for which its use is intended. These instruments are more sensitive to determine differences in QoL between sufferers of the same condition. An example is the European Organization for Research and Treatment of Cancer (EORTC) QLQ-C30, which is a cancer specific questionnaire (Aaronson et al. 1993). A number of QoL questionnaires have been developed to measure a particular aspect of QoL. For example, the Hospital Anxiety and Depression Scale (HADS) asks questions related to anxiety and depression (Zigmond, Snaith 1983). These types of instruments are usually used in conjunction with more general questionnaires, such as the SF-36.

QoL outcome measures are becoming increasingly used in RCTs. It is no longer sufficient to consider only the main outcome when deciding between competing alternatives. The cost of the treatment and also the overall benefit to the patients' everyday life is important. This is particularly evident in the context of a terminal disease where QoL of the patient may be just as important as the main outcome measure. A patient may prefer to have less time alive with a better quality of life, than a prolonged life, but with poorer quality. However, QoL outcomes are not limited to terminal disease and are often included as secondary outcomes. They are then considered in the overall benefit of competing treatment alternatives. If there is a large portion of missing data this has the potential to bias the results and ultimately affect clinical practice. QoL outcomes are perhaps more susceptible to missing data, as often they will be missing for reasons related to the QoL of the participant. Those feeling unwell may be less likely to fill in and return

questionnaires, thus providing an inflated view of the QoL across the treatment group. This is discussed further in section 1.4.3 and in later chapters. A brief description of each of the QoL instruments employed in the example RCTs is now provided.

1.3.1 Short Form-36 (SF36)

The SF-36 developed by Ware *et al*. evaluates general health status (Ware et al. 1993). The emphasis is placed on physical, social and emotional functioning. The SF-36 has become one of the most widely used measures for general health status. The questionnaire can either be self-administered or administered by a trained interviewer. There are 36 questions that address eight health concepts. These are summarised by two scores: physical health and mental health. The physical health domain is divided into scales for physical functioning (10 items), role-physical (4 items), bodily pain (2 items) and general health (5 items). Mental health consists of vitality (4 items), social functioning (2 items), role-emotional (3 items) and mental health (5 items). The remaining question asks about the respondents' perception of their health 'now'. The SF-12 is a shorter version of the SF-36, which covers all dimensions, but only uses 12 items to do so. Copies of the SF-36 and SF-12 instruments are found in the Appendix (1.1 and 1.2 respectively).

1.3.2 Arthritis Specific Health Index (ASHI)

The SF-36 Arthritis-Specific Health Index (ASHI) was constructed to improve the responsiveness of the SF-36 Health Survey to changes in the severity of arthritis. Arthritis-specific scoring algorithms are used to do this (Ware et al. 1999; Keller et al. 1999). A set of weights were identified to aggregate the eight scales in the generic SF-36 to achieve a single index that is more responsive to changes in arthritis severity.

1.3.3 EuroQoL EQ5D

The EuroQoL EQ5D is another general purpose instrument taking about two minutes to complete covering five dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression (Brooks, R with the EuroQoL Group 1996). Each dimension is addressed by a simple three-category response scale. A

single index is generated for all health states using the EuroQoL UK population tariff. There are 3*3*3*3*3=243 possible unique values, ranging from -0.59 (worst QoL) to 1 (best QoL). The score is treated as a continuous variable. In addition, a visual analogue scale (VAS) scored from 0 to 100 is presented on which the respondent should mark "your own health state today". This measure is frequently employed as it has the ability to be used as a 'utility' value in a cost-utility evaluation (Drummond et al. 1997). A copy of the EQ5D QoL instrument is found in Appendix 1.3.

1.3.4 EORTC QLQ-C30

The European Organisation for the Research and Treatment of Cancer (EORTC) quality of life questionnaire (QLQ) is an integrated system for assessing the health related QoL of cancer patients participating in clinical trials (Aaronson et al. 1993). The core questionnaire (QLQ-C30) is the result of more than a decade of collaborative research. The domains included are global health status (2 items), physical functioning (5 items), role functioning (2 items), emotional functioning (4 items), cognitive functioning (2 items) and social functioning (2 items). There are nine symptom scales, six of which are a single item: dyspnoea (shortness of breath), insomnia, appetite loss, constipation, diarrhoea and financial difficulties. The remaining three are fatigue (3 items), nausea and vomiting (2 items) and pain (2 items). The raw scores are linearly transformed onto a 0-100 scale. A mean change of 5 to 10 units is regarded as 'a little' subjective change to patients and 10 to 20 units a moderate change. Differences of 10 or more units are regarded as clinically significant (Osoba et al. 1998). A copy of the QLQ-C30 instrument is provided in Appendix 1.4.

In addition to the core questionnaire, there are a number of supplementary QLQ modules. These provide more detailed information to evaluate the QoL in specific patient populations. A module may be developed to assess: symptoms related to a specific tumour site, side effects associated with treatment, or additional QoL domains affected by the disease. Some of the existing modules are breast cancer, head and neck cancer, lung cancer, ovarian cancer and the oesophageal cancer module (Fayers et al. 2001).

1.3.5 Oxford Knee Score (OKS)

The Oxford Knee Score is a questionnaire containing twelve items measuring patients' perceptions of knee pain and function (Dawson et al. 1998). It contains questions such as "During the last four weeks have you felt that your knee might suddenly 'give way' or let you down?" There are five response categories to this question: rarely/never; sometimes or just at first; often, not just at first; most of the time; all of the time. Each of the twelve items is scored from one to five, from least to most difficulty or severity. These are combined to produce a single score that ranges from 12 (least difficulty) to 60 (most difficulties). A copy of this questionnaire is found in Appendix 1.5.

1.3.6 Reflux questionnaire

The Reflux questionnaire was developed as a tool to determine QoL among patients with gastro-oesophageal reflux disease (GORD) during the REFLUX trial (Grant et al. 2008; Macran et al. 2007). The final measure consists of 31 items covering seven categories: heartburn; acid reflux; wind; eating and swallowing; bowel movements, sleep; work, physical and social activities. The measure produces two outputs; a quality of life score (RQLS) and five reflux symptom scores. Both are measured on a 0-100 scale. The five symptom scores are general discomfort, wind and frequency, nausea and vomiting, activity limitation and finally, constipation and swallowing. The RQLS aims to measure the extent to which an individual participant feels their GORD symptoms, and any side effects of treatment that affect their QoL. A copy of this questionnaire is found in Appendix 1.6.

1.4 Missing Data

1.4.1 What are missing data?

Missing assessments are inevitable and in particular usually occur to some degree in longitudinal studies. The term 'missing' is used to include those that were expected (non-compliance), but can be used as a broader definition that includes unavoidable attrition such as death. Data may be missing for a variety of reasons, including being lost in the post. Alternatively, the data may be missing if the

subject was too ill to respond, or because the clinical staff thought the patient was too ill to be burdened with a questionnaire. This type of missingness is informative because it implies a high probability of low QoL, even though the exact value is unknown. Therefore, it is important to determine why the data are missing, as this will have an impact on any inferences made.

1.4.2 Types of missing data

There are two main types of missing data: missing items and missing forms. Missing items occur when a respondent to a questionnaire has answered some questions but failed to answer others. If a patient fails to complete a whole questionnaire then this is regarded as a missing form. Within some QoL instruments it is possible for one or more items to be missing, but the QoL scale score can still be calculated. For example, in the QLQ-C30, if at least half the items from the scale have been answered, it is assumed that the missing items have values equal to the average of those which are present for that respondent (Aaronson et al. 1993). This procedure is well-validated and a widely accepted method for overcoming missing items in some QoL instruments. Where appropriate, this process was implemented when calculating the QoL scores for the example datasets. Beyond this, the work presented in this thesis deals with the issue of missing forms, rather than missing items.

Table 1.1: Patterns of missing data for a study with four assessments

	Assessment			
Pattern	1	2	3	4
Complete	X	X	X	X
Monotone (terminal)	X	X		
Intermittent	X		X	X
Mixed	X		X	

X represents an observed assessment

In a longitudinal setting, the pattern of missing data can be described as either "monotone" (terminal) or "intermittent". Monotone missing data occurs when no further observations were made on a patient following a number of completed assessments. Intermittent missing data occurs when one or more observations for a patient were missing before one was observed. It is possible for a patient to have

a mixed pattern, with a period of intermittent dropout followed by monotone dropout. Table 1.1 illustrates each of these patterns for a study with four assessments with 'X' representing a completed assessment.

1.4.3 Why are missing data a problem?

Firstly, missing data are a problem due to the loss of power as a result of a reduced number of observations (Fairclough 2002). To some extent this can be overcome by increasing the planned sample size. Secondly, a major concern is that missing data may result in bias and the results of the clinical trial may not reflect the true situation (Fayers, Machin 2007). These issues are associated with missing data for any outcome, but are particularly prominent in the context of QoL outcomes. For example, if it is those patients experiencing poorer QoL that have missing data, then ignoring the presence of the missing data causes the analysis to be based on the observed data of those doing well. Therefore, any estimates of QoL will be overestimated. Alternatively, it is possible that individuals are dropping out of the study when they feel better, therefore, QoL will be underestimated. If the proportion of missing data is small, then provided the data are analysed appropriately, only a small bias will result (Fayers, Machin 2007). It is not possible to reduce bias by increasing the sample size.

1.4.4 Mechanism of missing data

The missing data mechanism (or missingness) was originally presented by Rubin (Rubin 1976). Rubin defined three mechanisms of missing data: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). In simple terms for MCAR, missingness does not depend on the variable of interest, or any others in the dataset. MAR occurs when missingness is conditional on another variable observed in the dataset, but not the variable of interest. For MNAR, missingness depends on the actual (unobserved) data value. The technical details of these definitions are presented in chapter four.

In the context of quality of life, the mechanism refers to whether missingness is somehow related to the QoL. MCAR occurs if the missingness has nothing to do with QoL (past, present or future). It could be because the patient has moved or the questionnaire was lost. Covariate dependent missingness also falls within this category. For example, if missingness varies between age groups, but within each individual age group, missingness is MCAR then there is covariate-dependent missingness. When missingness is related to the observed QoL, the data are MAR. For example, if the previous measure of QoL is predictive of missingness, then there is MAR data. MNAR data occurs if missingness is related to unobserved QoL (past, present or future). Chapter five uses the methods identified in chapter four to investigate the missing data mechanism present in the example datasets.

1.5 Approaches to deal with missing data

The ability to deal with the missing data is of paramount importance to be able to analyse the data without bias and prevent a loss of power. There are several approaches that can be used to deal with the missing data and these are investigated throughout the course of the thesis. A brief introduction to these approaches is provided here, with further detail provided in later chapters.

1.5.1 Complete case analysis

The easiest, but least desirable way to deal with missing data is to simply ignore it and carry out a complete case analysis. Clearly this has implications for the results of the trial. If the data are MAR, the resulting analysis will be severely biased. When data are MCAR, the only consequence to this approach is a loss of power due to the reduced number of subjects. By restricting the data to complete cases, the analysis method identified in the trial protocol can be undertaken.

1.5.2 Imputation

The process of imputation is based on completion rather than deletion. Imputation is a procedure whereby a reasonable alternative value replaces one that is missing. This results in an augmented complete dataset to which the standard statistical analysis techniques can be applied. Usually, the observed

values are used as a basis to impute values for the missing observations. Common simple methods include last value carried forwards, simple mean imputation and regression (Fayers, Machin 2001). The majority of methods require the MCAR assumption and will often under-estimate the variance, resulting in inappropriate standard errors and ultimately in-appropriate confidence intervals and p-values (Carpenter, Kenward 2007). More recently multiple imputation (MI) techniques have been developed which can take account of the uncertainty surrounding the missing value (Kenward, Carpenter 2007). Rather than a single value being imputed, a number of imputations are carried out, creating several augmented datasets. Standard statistical analysis is carried out on each of these datasets and the results combined using Rubin's method (Little, Rubin 2002). Advancements in computer software have allowed the emergence of MI as a more valid alternative to simple imputation. Both simple and multiple imputation procedures are discussed in more detail in chapter six and implemented on the example datasets in chapter seven.

1.5.3 Model-based strategies

Complete case analysis and most simple imputation procedures require the MCAR assumption. Often in QoL settings this assumption will be unlikely to be true and chapter five will show the data are more likely to be at least MAR if not MNAR (Fielding et al. 2008a). Clinical trial QoL outcomes are often analysed using the complete cases at the final endpoint perhaps adjusting for baseline QoL and other appropriate patient characteristics (chapter three). The data collected at the intermediate assessments are often not used. Model-based techniques can make use of this available data to overcome some of the problems that the missing data may cause. Methods such as a repeated measures model or mixed effects model, which assume MAR, can be employed.

In the case of MNAR data, more complex approaches should be considered. One option is a pattern mixture model where the portion of the model specifying the missing data mechanism does not need to be specified (Little 1994). Other possibilities include the selection model, joint mixed effects model or conditional

linear model (Fairclough 2002). All of these methods for non-ignorable missing data require strong assumptions and these assumptions cannot formally be tested. Further details on all these approaches are found in chapter eight and implemented on the trial data in chapter nine.

1.6 Example RCT datasets

To illustrate the issues surrounding missing data, a number of trial datasets will be utilised. These are described in detail in chapter three. There are seven RCTs each providing QoL assessments on at least three occasions throughout the trial. At each follow-up assessment a reminder system was employed generating an extra portion of data that would otherwise have been missing. Detailed records were kept by the trial researchers, and for each patient it is known whether the patient responded without reminder, after reminder or not at all. Using this additional information the data collected through reminders can be used to test the procedures of investigating the missingness mechanism, the accuracy of different imputation procedures and the appropriateness of the alternative model-based strategies.

1.7 Evaluating the economic benefit of different data collection strategies

The importance of reducing the amount of missing data is well known. The reminder system is one strategy that is thought to help achieve this. The current practice of employing a reminder system has obvious cost implications for the trial unit. A researcher is required to keep track of non-response, be responsible for issuing reminders or taking the time to carry out telephone interviews. Assessing whether this process is cost-effective is important for the future conduct of trials. Initially this will be assessed by conducting a cost-effectiveness analysis, to estimate the incremental cost per additional unit of data collected using the incremental cost-effectiveness ratio (ICER).

An alternative decision making tool to ICER and an elicitation of willingness to pay (WTP) is the net benefit (NB) framework. Stinnett and Mullahy provide a comprehensive account of this framework (Stinnett, Mullahy 1998). The cost-effectiveness decision rule uses the value λ that you are willing to pay for the extra information. This will be calculated for each of the example trials to determine whether the reminder strategy or imputation provided a net-benefit. Chapter ten uses these approaches to determine whether the additional effort in data collection is worth the extra cost.

1.8 Overview of thesis

The aims and objectives of the work presented in this thesis were outlined in section 1.1. The chapters that follow address these objectives. Chapter two provides a summary of the current missing data literature and a review into current imputation use for missing QoL outcomes in clinical trials. This chapter describes the rationale for the work presented in the rest of the thesis and highlights the importance of considering the missing data when analysing QoL outcomes. Chapter three provides a brief background for each of the example trial datasets. Chapter four discusses the theory behind the mechanism of missing data and the procedures to investigate this mechanism. This is followed by chapter five which implements these procedures to investigate the missingness mechanism for each of the QoL outcomes collected within the seven trial datasets (described in chapter three).

Following on from this, chapter six discusses the theory of imputation and describes the different imputation methods available and how different methods are appropriate for each of the missing data mechanisms described in chapter four. Chapter seven carries out imputation on the missing data in the trial datasets to identify which methods are most suitable. Data collected by reminders are removed, allowing the accuracy of imputation to be calculated. As described in chapter three, the QoL outcomes are often analysed using a complete case analysis on the final endpoint. Chapter eight describes alternative model-based

strategies that make use of the intermediary assessments and which can be used in the presence of missing data. Chapter nine carries out these strategies on the trial datasets. Chapter ten considers the economic benefit of the reminder system and the competing alternative of imputation. Finally, chapter eleven discusses the findings of all the work carried out and presents some recommendations for the future.

Chapter 2 Overview of the missing data literature

2.1 Background

The literature on missing data is ever increasing, as researchers realise the importance of considering missing data when analysing trial data. Simply ignoring the data and carrying out complete case analysis is now regarded as perhaps the least desirable option (Carpenter, Kenward 2007). Despite this, the complete case approach is still widely used in the analysis of quality of life (QoL) outcomes in clinical trials (section 2.2). As was introduced in chapter one QoL outcomes are increasingly being used in clinical trials as part of follow-up questionnaires. Missing QoL data can be very informative in its own right and the reason for missingness could be the QoL itself. QoL is a subjective patient reported outcome, which perhaps makes it more sensitive than other outcomes to missing data assumptions. Disregarding those patients without QoL information is likely to bias the results, affect trial conclusions and ultimately clinical practice. Therefore, it is important to make use of as much data as possible from as many patients as possible. Although not intended to be systematic, this section aims to summarise some of the current literature on missing data.

2.2 Review of the missing data literature

2.2.1 Missing data mechanism

The idea of the missing data mechanism was first introduced by Rubin (Rubin 1976). Rubin defined three mechanisms of missing data: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). The details of these are found in chapter four. Since then many authors have looked at ways of establishing this missing data mechanism (Curran et al. 1998a; Listing, Schlittgen 2003; Listing, Schlittgen 1998; Little 1988; Ridout 1991; Diggle, Kenward 1994). Little developed a test based on means under the different missing data patterns (Little 1988). Listing and Schlittgen also proposed a test based on means (Listing, Schlittgen 1998) and secondly, a non-parametric procedure which combines several Wilcoxon rank sum tests (Listing, Schlittgen

2003). Schmitz and Franz discussed a non-parametric version of the first Listing and Schlittgen test (Schmitz, Franz 2002). Diggle used an approach which tests whether the subset about to dropout are a random sample of the whole population (Diggle 1989). Ridout adopted a similar approach to Diggle, by utilising logistic regression (Ridout 1991). The response indicator distinguishes between responders at a given assessment who continue in the study and those who do not and this is their final QoL assessment. Fairclough detailed a logistic regression procedure subtly different from that of Ridout (Fairclough 2002). The missingness indicator used by Fairclough distinguishes between responders and non-responders at each assessment. The detail of these methods are provided in chapter four and application to the example datasets is undertaken in chapter five, to investigate the missing data mechanism present within the data.

Establishing the missing data mechanism is an important first step in determining how best to analyse the data (Curran et al. 1998a). Once this mechanism is known, it can affect the choice between different analysis options and whether imputation (and which method) is appropriate. As this is a critical step in deciding how to analyse data when there are missing observations, many investigators have explored tests of missingness. They have either generated artificial datasets using simulation techniques (Musil et al. 2002), or have made use of existing datasets in which missing data was then artificially created (Myers 2000). This procedure is potentially misleading: the missing patterns were predetermined and pre-specified, and usually the performance of the various tests can be anticipated through the known mechanism that was used to generate the samples. Chapter five presents an alternative approach that utilises the data collected through reminders. Since the data is ultimately known, the mechanism behind reminder-response can be identified and used to inform the actual mechanism of missing data.

2.2.2 Imputation

Imputation is a common approach to deal with missing data. Various methods can be implemented to impute the missing data and produce a more complete dataset for analysis. These methods are discussed further in chapter six. Simes *et al*. discuss that whilst imputation can lead to bias, there are instances where the imputation of missing values is desirable (Simes, Greatorex & Gebski 1998). They recommend that in cancer clinical trials the use of data imputed from auxiliary health outcomes is preferable to ignoring the missing data. This could then be used for the primary treatment comparisons of QoL or be used as a sensitivity analysis to assess the impact of the missing QoL data on the main conclusions. However, the accuracy of imputation cannot normally be determined, as the true values are not known. Thus, various authors have explored the potential accuracy of imputation methods by artificially removing data from a dataset and treating it as missing (Musil et al. 2002; Myers 2000; Twisk, de Vente 2002; Hunsberger et al. 2001). This is a circular argument, because the data is either removed at random or according to some known and pre-specified pattern. In practice, the major analytical problem is one does not know the exact missing mechanism. By removing the data in a certain way, the mechanism of missingness is inherent in the data and thus, some imputation methods will perform better than others depending on what this mechanism is.

Engels and Diehr noted the need to use data with real missing patterns, and attempted to overcome these problems by using a dataset where a value was observed after one or more missing values had occurred; the observed value was treated as the true value for the missing data at the preceding time points (Engels, Diehr 2003). Various imputation methods were applied for the missing values, and the results compared against the observed value to assess accuracy of the imputation methods. As Engels and Diehr comment,

> "this analysis hinges on the similarity of a known value following a string of missing values to other observations that are missing at that same time (Engels, Diehr 2003)."

In contrast, Musil *et al*. simulated a dataset with MAR data and imputed the missing values (Musil et al. 2002) . They concluded that all methods had their limitations, but mean substitution was least effective, whereas the expectation maximisation algorithm was reasonable. Twisk and De Vente preferred longitudinal imputation methods (those using person-specific data from other assessments) to cross-sectional methods (those based on other observations at the assessment in question) (Twisk, de Vente 2002). Hunsberger *et al*. found that a predicted regression procedure performed best (Hunsberger et al. 2001).

A number of case studies and simulation studies have shown multiple imputation to be superior to simple imputation in the presence of informative missing data (Myers 2000; Hunsberger et al. 2001; Cook 1997; Donders et al. 2006; Huson, Chung & Salgo 2007; Liu, Gould 2002; Morita et al. 2005; Patrician 2002; Tang et al. 2005). However, as stated above, in these studies the missing data have been created artificially and the missing data patterns are pre-determined or pre-specified. It is, therefore, not surprising that some imputation methods perform poorly when the mechanism is already known. For example, simple methods such as mean imputation are only useful if the data are likely MCAR. If not MCAR, this type of method is likely to produce biased results. Fielding *et al.* showed that simple imputation was inadequate in the presence of MNAR and suggested multiple imputation might be more appropriate (Fielding et al. 2008a).

Chapter seven uses the example datasets to assess accuracy of different imputation methods. The advantage of the approach used here over that used by previous authors is that the true value of the data being imputed is known. The data collected by reminders are regarded as missing; they are then imputed allowing the accuracy to be assessed. This differs from the approaches outlined above, as the data are not being removed subject to a known mechanism. The mechanism is the one inherent in the data. Chapter five identifies this mechanism and thus, informs which imputation methods are likely to be appropriate in chapter seven.

2.2.3 Model-based approaches to missing data

There are a number of different methods for analysing longitudinal data containing missing values. Troxel *et al*. summarised these issues while investigating the analysis of missing QoL outcomes from cancer clinical trials (Troxel et al. 1998). The validity of each method depends on the pattern of missing data and the missing data mechanism. As previously discussed, the easiest approach is to simply ignore the missing data using a complete-case analysis and apply standard statistical procedures. However, this strategy could result in biased estimates of treatment effect, unless non-differential missingness patterns and ignorable missingness mechanisms have been identified (Yang, Shoptaw 2005). Ignorable missingness refers to missing data that are at most missing at random and statistical analysis using observed data likelihood function could provide unbiased results, but in no circumstances does it imply that one can disregard or overlook the missing data. When fractions of missing data are low or the reasons for missingness are comparable across treatment groups, an unbiased estimation may still be possible to obtain (Yang, Shoptaw 2005).

Under MAR, available case analysis such as mixed effects models can be used whereas for MNAR data more sophisticated methods such as the selection model, joint model or pattern mixture model are appropriate (Fairclough 2002). These approaches are described in detail in chapter eight and applied to the example datasets (where appropriate) in chapter nine. Joint modelling can be used to combine a repeated measures approach with a model for the drop out mechanism. This can be called a selection model because the process of obtaining incomplete data can be viewed as a form of selection (Encrenaz et al. 2005). Diggle and Kenward also proposed a selection model which makes it possible to calculate unbiased parameters in the analysis of dependent variables (Diggle, Kenward 1994).

Encrenaz *et al.* examined the influence of drop-outs in a longitudinal study of opiate users (Encrenaz et al. 2005). A classic data analysis (linear mixed effects model for repeated measures) was compared to a selection model, which in this

case consisted of a joint model of the addiction index score and of the drop-out probability, in order to reduce bias induced by dropouts. The missing data were found to be informative and associated with low drug use. They found that not taking the informative drop-outs into account could lead to false conclusions. This conclusion was replicated in Fairclough *et al.* where analyses assuming random patterns of data tended to underestimate the decline in QoL over time (Fairclough et al. 2003). However, the sensitivity analyses showed that the significant difference in QoL between treatment groups was still apparent even when the non-random data loss was accounted for (Fairclough, Gagnon & Zagari 2005; Fairclough et al. 2003).

Molenberghs *et al*. investigated the analysis of an incomplete longitudinal clinical trial using various simple and more complex models including last observation carried forward (LOCF) imputation (Molenberghs et al. 2004). They raised the point that caution ought to be used, since no modelling approach whether MAR or MNAR can recover the lack of information due to incompleteness of the data. MNAR models are more general and explicitly incorporate the dropout mechanism. The inferences they produce are typically dependent on un-testable and often implicit assumptions regarding the distribution of the unobserved measurements, given the observed measurements. Molenberghs *et al*. recommend that a compromise between shifting to MNAR models, or ignoring them altogether, is to make them a component of a sensitivity analysis (Molenberghs et al. 2004).

Houck *et al*. agree with this and concluded that sensitivity analysis addressing the impact of alternative assumptions or models on the MNAR data should be conducted before drawing conclusions from the study (Houck et al. 2004). As does Fairclough *et al.* who suggest that to explore the impact of deviations from the MAR assumption on study conclusions, a sensitivity analysis should ideally include MNAR models (Fairclough et al. 2008). Methods that utilize auxiliary information such as multiple imputation (MI) and joint models figure prominently in these models (Fairclough et al. 2008).

The analysis of repeated measures with missing data is not trivial because of the difficulties in identifying the missing data mechanism (Troxel et al. 1998). This is particularly true in the analysis of QoL data, where data may be missing for several reasons. Therefore, it is important to collect as much information as to why QoL questionnaires are not completed to enable researchers to inform the missing data models. The theory behind these various model-based strategies introduced here is discussed in chapter eight and are applied to the example data in chapter nine.

2.3 A review to assess current use of imputation in RCTs

2.3.1 Background

Imputation is one way of dealing with missing data to provide an augmented complete dataset on which standard statistical analysis procedures can be carried out. As highlighted above different methods are appropriate in different situations and their accuracy will depend on the underlying missing data mechanism. It is important to consider this mechanism when deciding whether to carry out imputation and then to decide which procedure is most appropriate. This consideration is explored further in chapter five and imputation of missing data in the example datasets is carried out in chapter seven.

In 2004, Wood *et al*. carried out a review of imputation use for missing outcome data (Wood, White & Thompson 2004). This review considered all outcomes from clinical trials both continuous and categorical. They found that despite advances in software, simple imputation procedures were more commonly used by researchers. As the focus of this thesis is restricted to QoL outcomes, it was thought a similar review specifically for QoL outcomes would be appropriate. The review presented here considered RCTs published during 2005 and 2006 in four leading medical journals. The aim was to determine whether imputation is currently being used by trial researchers to overcome missing QoL data and if so

which methods they adopt. In addition, was the mechanism of missingness discussed and the rationale for the choice of imputation method presented? As already mentioned it is important to consider the missing data mechanism when choosing between imputation methods, but it was thought that this was often ignored. The methods and results of this review are outlined below and have also been published in the peer-reviewed journal Trials (Fielding et al. 2008b).

2.3.2 Methods

A PubMed search was carried out to identify RCTs published during 2005 and 2006 in the four leading medical journals: British Medical Journal **(**BMJ), Journal of the American Medical Association (JAMA), Lancet and New England Journal of Medicine (NEJM). Of those articles identified a random selection of a half was obtained. The focus of this review was those studies which included QoL outcomes (both primary and secondary). Data extraction was carried out as a two stage process, firstly to identify those studies which included a QoL outcome. Once these had been identified a more detailed data abstraction was undertaken to obtain details about the trial, its outcomes, use of imputation and results. The details of the trials recorded during data abstraction were:

- outcome and type (primary and QoL)
- single or repeated endpoints
- amount of missing information
- was imputation used and if so, which method
- was the mechanism of missingness discussed
- were reasons for missingness presented
- number of participants
- patient demographics (age, gender etc)
- description of treatments
- method of primary analysis
- was imputation part of primary analysis or sensitivity analysis

The initial data extraction was carried out by one of four reviewers with queries resolved by consulting a second reviewer. To assess consistency between

reviewer's two papers were doubly abstracted. No inconsistencies were shown. The second stage was undertaken by one reviewer (Shona Fielding), with a second reviewer consulted where necessary.

2.3.3 Description of the missing data

The search strategy described above produced 568 articles for potential inclusion. Following a process of random selection, 285 articles were identified for this review. A QoL outcome (primary or secondary) was reported in 61 trials (21%) and these form the basis of the review presented here. The majority of these 61 articles were published in the BMJ (n=27, 44%) with 17 (28%) published in NEJM, ten (16%) in the Lancet and the remaining seven (12%) in JAMA.

Table 2.1: Number of studies with different proportions of missing data

Proportion of missing data*	No Imputation	Imputation	Total
None	6	0	6
<10%	16	5	21
11-20%	6	5	11
>20%	8	3	11
Unclear	6	6	12
Total	42	19	61

* For the primary QoL endpoint

Table 2.1 describes the proportion of missing data split between studies that did and did not employ imputation. There were 42 studies that did not perform imputation techniques; of which six did not provide enough information to determine the proportion of missing data. Of the remaining 36 studies, 16 had less than 10% missing data in the primary QoL endpoint. This is in contrast to the 5 out of 19 studies that used an imputation method. The proportion of missing data was unclear for six of the 19 studies which used imputation. Across all 61 studies with a QoL outcome, 12 (20%) were unclear on the proportion of missing data.

Current CONSORT (Consolidated Standards of Reporting Trials) guidelines for the reporting of RCTs require authors to provide a flow diagram of participants in the trial (Moher, Schulz & Altman 2001). This diagram should detail the withdrawals and reasons for withdrawal. The majority of trials (n=50, 82%)

contained within this review did provide the flow diagram and reasons for missingness such as withdrawal, death or other medical problems. However, there was no detailed discussion of these reasons and the impact they may have had on the analysis. In only one study the mechanism of missingness was discussed explicitly and it was found to be non-ignorable (Petersen et al. 2005b). Thus, ignoring the missing data is liable to provide biased results as the QoL is related to missingness.

2.3.4 Quality of life instruments

There was a range of QoL instruments used within the identified trials. Nine different generic QoL instruments were utilised: General Health Questionnaire – seven trials; SF12/SF36 – 14 trials; World Health Organisation Quality of Life questionnaire – one trial; Global assessment of functioning – one trial; EuroQoL EQ5D - five trials. There were a large number of disease specific measures used due to the differing disease areas the identified trials covered. Some examples were the asthma QoL score, dermatology life index, rhinoconjunctivitis QoL measure, irritable bowel disease questionnaire, Alzheimer's disease assessment scale and the Oswestry pain score. The post treatment follow up using these QoL measures ranged from one to five assessments, with the majority being collected within twelve months.

2.3.5 Imputation

Nineteen (31%) of the 61 trials used some form of imputation. Descriptions of these 19 studies are found in Appendix 2.1. Thirteen studies undertook imputation in the primary analysis (Petersen et al. 2005b; Ballard et al. 2005; Berry et al. 2006; Blumenthal et al. 2005; Buszewicz et al. 2006; Feagan et al. 2005; Hsieh et al. 2006; Kaplan et al. 2006; Kennedy et al. 2005; Korzenik et al. 2005; Nair et al. 2006; Winblad et al. 2006; Wright et al. 2005). Seven of these studies employed the imputation method last value carried forward (LVCF) or equivalently last observation carried forward (LOCF) (Ballard et al. 2005; Blumenthal et al. 2005; Feagan et al. 2005; Kaplan et al. 2006; Korzenik et al. 2005; Nair et al. 2006;

Winblad et al. 2006). Berry *et al*. used a combination of worst value imputation and LVCF (Berry et al. 2006). The worst value observed in the sample was imputed if missingness was known to be due to asthma. If missingness was unrelated to asthma, LVCF was used. Hseih *et al.* carried forward the baseline QoL value to the post treatment and six month follow up assessment (Hsieh et al. 2006). Buszewicz *et al*. employed hotdeck imputation (random selection from observed scores for each missing value) for missing baseline values and multiple imputation (using a predictive model) for the missing follow up scores (Buszewicz et al. 2006). Kennedy *et al*. imputed missing scores based on changes in other items when at least 75% of those items were present (such as irritable bowel syndrome severity scale) (Kennedy et al. 2005). Petersen *et al*. used a projection method appropriate for assessing responses among subjects with neurodegenerative disease (Petersen et al. 2005b). In the remaining study that employed imputation as part of the primary analysis, the imputation method was not specified, only that imputation was undertaken (Wright et al. 2005).

Six trials reported results after imputation as a sensitivity analysis. LVCF was used by four studies (McManus et al. 2005; Meggitt, Gray & Reynolds 2006; Petersen et al. 2005a; Thomas et al. 2006) and of these, Peterson *et al.* also considered imputing a zero value for those that were missing and McManus *et al*. evaluated the use of the mean of the series. Fairbank *et al.* employed multiple imputation using a regression model as part of a sensitivity analysis (Fairbank et al. 2005). Hunkeler *et al*. mentioned that they used imputation in a sensitivity analysis, but did not specify which method (Hunkeler et al. 2006).

The imputation process described above relates to missing form imputation, namely the whole QoL measure. The QoL instruments are made up of items which contribute to the score. In the case of the EQ5D measure, if one of the five items is missing, the overall health status score cannot be calculated. However, in the SF36, if at least half the items in a scale are provided, the mean of the observed items is imputed for the missing items, allowing the scale score to be calculated. In the trials contained within this review, item imputation was not discussed, so it

is possible this process was carried out, reducing the amount of missing data reported.

2.3.6 Analysis methods

In those studies where imputation was not used (42 trials), the method of analysis was unclear in three cases (7%). A complete case analysis was undertaken in 30 of the 42 trials (71%). The methods of complete-case analysis were: t-test (11 trials); analysis of variance (ANOVA) (one trial); analysis of covariance (ANCOVA) (13 trials); general linear model (one trial); Mann-Whitney test (four trials). For those studies not undertaking a complete-case analysis, a repeated measures approach was used by nine trials (22%). Eight of these used a linear mixed model for those patients with at least baseline data, thus allowing for some missing values in follow up data. In the ninth trial, area under the curve (AUC) was used for analysis.

Following imputation, 10 of 19 trials used ANCOVA, two used regression, two trials used a general linear model, two a t-test, one used generalised estimating equations, one a stratified rank test and finally one used a repeated measures model. All of the studies, which employed imputation, collected data for repeated assessments, but only four of the 19 trials used a repeated measures analysis.

2.3.7 Conclusion

Curran *et al*. discuss the importance of the missing data mechanism and how that can impact on the choice of imputation of missing QoL outcomes (Curran et al. 1998b). The short review presented here highlighted that this consideration is often ignored, or at least not reported in the peer-reviewed publications. The mechanism of missing data was only explicitly discussed by one study. However, the majority of trials did identify the number of patients used in analysis by use of a flow diagram, as required by the CONSORT guidelines (Moher, Schulz & Altman 2001). Half of the trials with a QoL outcome performed complete case analysis. There was no detailed discussion of the impact this had on the analysis and bias contained within the reported results. Complete case analysis is easy to

perform, but has the potential to remove a large portion of the patients from the analysis. This method has two major disadvantages. Firstly, it reduces the sample size (and thus the power of study) and secondly, may produce biased results unless the data are MCAR. The standard mixed effects model assumes MAR data. However, this assumption was not discussed by any of the authors which undertook this method of analysis.

A fifth of the articles sampled contained a QoL outcome, with 31% of these performing an imputation procedure. The rationale behind the choice of imputation method was not discussed by any of the authors. The review showed that of those choosing to carry out imputation, LVCF was popular. However, this method makes the assumption that the outcome is unchanged with time and in QoL situations this is unlikely. In some studies, you might carry forward an off-treatment score to an on-treatment missing value which is not likely to be that reflective the truth. As Gadbury, Coffey and Allison state,

> "…although intuitively appealing, LOCF [LVCF] requires restrictive assumptions to produce valid statistical conclusions (Gadbury, Coffey & Allison 2003)."

Carpenter and Kenward agree with this conclusion and provide a thorough critique of LOCF [LVCF], warning against its use (Carpenter, Kenward 2007).

The results of this review support the conclusions of the previously mentioned review by Wood *et al.*, which considered all primary outcomes irrespective of type (Wood, White & Thompson 2004). They found that LVCF was popular with only one trial using multiple imputation. Despite the advances in available software (Yu, Burton & Rivero-Arias 2007) and recommendations against LVCF, the researchers involved in RCTs tended to prefer this simple method.

2.4 Summary

Despite there being numerous articles on determining the missingness mechanism and determining the accuracy of imputation, to my knowledge they are all based

upon simulated data or data that has been removed in a particular way. Thus, the performance of some methods can be anticipated because the mechanism of missingness is known. The work presented in this thesis has the advantage that real 'known' data are used. By utilising the data collected through the reminder system the mechanism of missingness can be identified and accuracy of imputation assessed. Identification of the correct mode of missingness and most appropriate method of imputation can make a large impact on the analysis of clinical trials. The sensitivity of different analyses depends on the proportion of missing assessments and the strength of the underlying causes for missing data (Fairclough, Peterson & Chang 1998). It is crucial to identify the mode of missingness and thus, the most appropriate method for valid analysis and unbiased results. In general, the undesirable effect of missingness on bias and power increases with the severity of non-randomness, as well as the proportion of missingness (Curran et al. 1998a).

As previously mentioned, missing QoL data can be informative and identifying the mechanism of missingness is important in determining which (if any) imputation method is suitable. None of the authors contained in the review of section 2.3 gave proper attention to this. In chapters four and five, this missingness mechanism is discussed in more detail and the methods are applied to the example trial datasets. The conclusions from this can be used to inform which imputation strategies are likely to perform best in chapter seven. The impact of the imputation on calculated treatment effects is also determined. The review also showed that despite repeated measures being taken, the chosen analysis method was often based on the final endpoint only. Chapter eight describes alternative model-based strategies (including repeated measures techniques) and chapter nine applies these to the datasets.

The current literature is extensive on the issue of missing data. However, the approaches used in this thesis are novel, in the sense that missing data later recovered are utilised. The problem of never knowing the data required to test the missing data assumptions or the accuracy of imputation is overcome. Using data

recovered through reminders allows the assessment of the impact of the missing data on trial conclusions and determines whether different analysis strategies give consistent conclusions. The emphasis of the whole thesis is on an empirical investigation of methods for missing data, rather than a theoretical one. The responses collected by reminder form the basis of this investigation.

Chapter 3 Description of the datasets

3.1 Introduction

To illustrate the issues surrounding the analysis of clinical trials with missing QoL data, seven data sources will be used. All seven datasets are from completed randomised controlled trials (RCTs) and have been provided by secondary sources. Five datasets have been obtained from trials undertaken by the Health Services Research Unit (HSRU) at the University of Aberdeen. HSRU has a national remit, to research the best ways to provide health care, and to train those working in the health services in research methods. Most research projects aim to find out whether developments within the health service really are effective, efficient and appropriate. The unit is funded by the Chief Scientist Office (CSO) of the Scottish Government Health Directorate. The five trial datasets provided by this unit are:

1. REFLUX – The place of minimal access surgery amongst people with gastro-oesophageal reflux disease
2. MAVIS – Effect of multivitamin and multimineral supplements on morbidity from infections in older people
3. RECORD – Daily oral vitamin D and calcium in the secondary prevention of osteoporosis related fractures in older people
4. KAT - A randomised trial of different knee prostheses
5. PRISM – Multi-centre randomised trial of symptomatic versus intensive bisphosphonate therapy for Paget's disease

Two further trial datasets were obtained from other sources:

6. TOMBOLA – Trial of management of borderline and other low-grade abnormal smears
7. Norwegian Palliative Care (NPC) Trial – A cluster randomised trial comparing standard palliative care to comprehensive palliative care on advanced cancer patients quality of life, location of care and place of death

The TOMBOLA dataset was obtained from the Epidemiology Group, which forms part of the Section of Population Health at the University of Aberdeen. The Norwegian Palliative Care Trial was conducted within Norwegian Public Health Care at the University Hospital, Trondheim, Norway.

In each of these trials a number of follow-up assessments were made through the use of a postal questionnaire. QoL instruments were included on this questionnaire and for three of the trials (REFLUX, KAT and NPC trial) this was the main (primary) trial outcome. In the remaining four trials, QoL instruments were used as secondary outcomes. A common feature of all seven trials was the use of a reminder system for follow up questionnaires. Once a follow-up questionnaire was issued, if it had not been returned within two weeks a reminder including an additional questionnaire was sent to the participant. In some cases, a second reminder was sent a further two weeks later. The seven datasets were chosen for illustration as the trial researchers had kept good records of which patients completed the follow up questionnaires after being issued a reminder. This feature of knowing which patients required prompting to return their questionnaires forms the basis of the work presented. Patients who responded without the need for a reminder are termed 'immediate-responders' throughout. Those who required prompting either with one or two reminders are termed 'reminder-responders'.

In the five HSRU trials, the QoL outcomes were analysed using an analysis of covariance (ANCOVA) to assess the difference in quality of life (QoL) (at the final endpoint) between treatment groups. This difference was adjusted for baseline QoL and a number of other baseline characteristics specific to each trial. This was carried out as a complete case analysis and incorporated only those patients for which baseline and final QoL assessments were available. As discussed in chapter one, this complete-case analysis strategy has implications for bias and loss of power. One aim of the work presented in this thesis is to determine whether this strategy was the most appropriate. The remainder of this chapter aims to describe

each of the seven trial datasets in turn. Information on the patient characteristics, follow-up assessments, including questionnaire completion, and the results from the reported trial analysis are provided.

3.2 REFLUX

The aim of REFLUX was to evaluate the clinical- and cost-effectiveness of early laparoscopic surgery compared with continued medical management amongst people with gastro-oesophageal reflux disease (GORD) (Grant et al. 2008).

Table 3.1: REFLUX - Description of randomised groups at trial entry N (%)

	Total (N=357)	Surgical (N=178)	Medical (N=179)
Baseline questionnaire returned – N (%)	349 (98)	175 (98)	174 (97)
Age – mean (SD)	46.3 (11.1)	46.7 (10.3)	45.9 (11.9)
Male – N (%)	236 (66)	116 (65)	120 (67)
BMI – mean (SD)	28.4 (4.2)	28.5 (4.3)	28.4 (4.0)
Duration in months of prescribed medication for GORD - median(IQR)	32 (15,76)	33 (15,83)	31 (16,71)
Employment status - N (%)			
Full-time	226 (63)	116 (65)	110 (61)
Part-time	29 (8)	13 (7)	16 (9)
Retired	34 (10)	12 (7)	22 (12)
Other	68 (19)	37 (21)	31 (17)
Age left full-time education – N (%)			
16 and under	218 (62)	110 (63)	108 (61)
17-19 years	78 (22)	38 (22)	40 (23)
20 years +	58 (16)	28 (16)	30 (22)
Current Smoker – N (%)	86 (24)	46 (26)	40 (22)
Erosive oesophagitis – N (%)	182 (59)	85 (55)	97 (62)
Co-morbidity - H. Pylori status – N (%)			
Positive (subsequently treated)	26 (10)	12 (9)	14 (10)
Negative (subsequently untreated)	4 (2)	1 (1)	3 (2)
Negative	148 (55)	75 (56)	73 (54)
Uncertain	90 (34)	45 (34)	45 (33)
Hiatus Hernia present – N (%)	196 (59)	94 (57)	102 (60)
Asthma – N (%)	42 (12)	21 (12)	21 (12)
Source of recruitment – N (%)			
Retrospective	167 (49)	84 (49)	83 (48)
Prospective	176 (51)	87 (51)	89 (52)

GORD – gastro-oesophageal reflux disease; H.Pylori – helicobacter Pylori.

At three and twelve months after surgery (or equivalent for those managed medically), all patients were sent postal questionnaires which included the EQ5D, SF12 and a newly developed Reflux quality of life score (RQLS) (Macran et al. 2007). Although follow up may have continued for longer, the 12 month scores were considered the primary endpoint. The trial was spilt into two parts: a randomised trial (N=357) and preference trial (N=453). Those patients who did not consent to randomisation had the opportunity to enter the preference trial. The trial publications deal with each of these two sets of participants separately (Grant et al. 2008; Grant et al. 2009). The focus of this thesis is randomised trials. Therefore, only the data from those participants in the randomised part of the trial are utilised here (N=357). Baseline characteristics of the patients in the randomised trial are shown in Table 3.1.

Table 3.2 shows the number of participants responding at each follow-up assessment. At the 3-month assessment of those participants sent the questionnaire (91%), there were 43% immediate-responders, 51% were reminder-responders and 6% did not respond at all. At 12 months, 95% of participants were sent the questionnaire with 40% of these responding straight away. Again, only 6% of those sent the questionnaire did not respond (Table 3.2). In this trial, the use of reminders substantially increased the response rate. Without the reminder-response data, the trial would have achieved a response rate of less than 40% for the primary outcome. It is unlikely that a trial result based on this would have been accepted for publication.

Table 3.2: REFLUX - Number (%) of each responder type

	Follow-up N(%)	
	3 months	12 months
Immediate responder	141 (40)	136 (38)
Reminder responder	167 (47)	183 (51)
Sent - not returned	18 (5)	19 (5)
Not sent	31 (9)	19 (5)
Total	357 (100)	357 (100)

The mean (SD) of the four QoL scores are shown in Table 3.3 split by treatment group. The EQ5D and SF12 component scores were reasonably stable in the medical group, but there was slightly more variation in the surgical group. The

greatest change was shown in the surgical group for the RQLS from baseline to 12 months (63.6 to 84.6 units). There was an improvement in the medical group, but of less magnitude. Across the whole sample there was a slight improvement in the mean QoL from baseline to 12 months for the EQ5D and SF12 component scores, but this improvement was much greater for the RQLS.

Table 3.3: REFLUX - Mean (SD) QoL scores

	Surgical		Medical		Total	
	N	Mean (SD)	N	Mean (SD)	N	Mean (SD)
EQ5D						
Baseline	171	0.71 (0.26)	173	0.72 (0.25)	344	0.72 (0.26)
3 months	149	0.79 (0.23)	153	0.69 (0.30)	302	0.74 (0.27)
12 months	152	0.75 (0.25)	164	0.71 (0.27)	316	0.73 (0.26)
SF12 physical component score (PCS)						
Baseline	167	45.0 (9.8)	169	45.3 (9.1)	336	45.2 (9.4)
3 months	143	48.5 (8.9)	149	45.5 (10.6)	292	46.9 (9.9)
12 months	150	48.0 (10.2)	161	45.1 (9.7)	311	46.5 (10.0)
SF12 mental component score (MCS)						
Baseline	167	45.4 (11.7)	169	45.2 (12.0)	336	45.3 (11.8)
3 months	143	47.9 (12.5)	149	44.0 (12.9)	292	45.9 (12.8)
12 months	150	46.6 (12.8)	161	45.1 (13.0)	311	45.8 (13.0)
RQLS						
Baseline	164	63.6 (24.1)	163	66.8 (25.5)	327	65.2 (24.4)
3 months	141	83.9 (19.4)	145	70.6 (24.6)	286	77.1 (23.1)
12 months	145	84.6 (17.9)	154	73.4 (23.3)	299	78.8 (21.6)

The primary trial analysis for the quality of life outcomes was an ANCOVA of the 12 month scores adjusting for sex, age, BMI and baseline QoL (Grant et al. 2008). The analysis method dictated that only those patients who provided baseline and 12 month scores could be included (EQ5D – 86%, SF12 – 83%, RQLS – 77%). The results of the ANCOVA for the 12 month treatment difference are shown in Table 3.4.

Table 3.4: REFLUX – reported trial results

QoL Measure	N	Mean Difference at 12 months	95% CI	p-value
EQ5D	308	0.047	(-0.004, 0.097)	0.07
PCS	298	3.45	(1.75, 5.23)	<0.001
MCS	298	1.56	(-0.76, 3.89)	0.19
RQLS	276	14.1	(9.6, 18.6)	<0.001

There was evidence of a difference in the PCS between surgical and medical patients ($p<0.001$), with those receiving surgery, on average reporting 3.5 units higher on the scale (better physical functioning). The MCS was not found to differ between treatment groups ($p=0.19$). Although, the group mean was slightly higher at 12 months in the surgical group. The EQ5D score was borderline significantly different ($p=0.07$) with the surgical patients reporting slightly higher scores. The greatest impact of treatment was seen on the RQLS giving a mean difference of 14.1 (9.6, 18.6). Those receiving surgery reported the better reflux specific QoL ($p<0.001$). Hence, overall there appeared to be better QoL at 12 months in the surgical group. However, this difference was only statistically significant for the PCS and the RQLS.

3.3 MAVIS

The MAVIS trial was a randomised controlled trial of multi-vitamin and mineral supplementation (Avenell et al. 2005). The aim of the study was to determine whether, in persons aged 65 and over, dietary supplementation reduced infection rates and antibiotic usage, reduced general practice consultation rates, improved quality of life, reduced the number of infection related hospitalisations and improved memory. There were 910 participants recruited aged 65 or over, who did not already take vitamins or minerals. Trial participants were asked to take one tablet daily for one year and were randomly allocated to receive either a multi-vitamin and mineral supplement (n=456) or a placebo (n=454). Randomisation was stratified by general practice and minimised by age (65-74, 75-84, ≥85), sex and type of accommodation (community or in care). Information was collected on health status, quality of life, nutritional risk and cognitive function (Table 2.5). Participants kept a simple diary, recording any infections they contracted during the trial period. QoL was assessed at baseline, six and twelve months with the EQ5D and the SF12 questionnaires.

The groups were evenly balanced for the patient characteristics (Table 3.5). The primary trial analysis of the QoL data was an ANCOVA to estimate the mean

difference in follow-up QoL scores between treatment groups. This was adjusted for baseline values, age group, sex and type of accommodation. There was no statistically significant effect from supplementation on QoL at either six or 12 months (Table 3.6).

Table 3.5: MAVIS - Baseline participant characteristics N (%)

Characteristic	Supplement Group (N=456)	Placebo Group (N=454)
Aged >=85	19 (4)	16 (4)
Women	217 (48)	214 (47)
BMI (mean (SD))	28.2 (4.2)	27.9 (4.1)
Smoking status	(N=456)	(N=453)
Current smoker	57 (13)	63 (14)
Current number of drugs taken	(N=455)	(N=453)
0-2	205 (45)	234 (52)
3-6	198 (44)	164 (36)
>6	52 (11)	55 (12)
Past and present chronic conditions	(N=456)	(N=454)
Hypertension	188 (41)	172 (38)
Heart disorders	137 (30)	130 (29)
Chest disorders	86 (19)	87 (19)
Diabetes	37 (8)	42 (9)
Cancer	46 (10)	46 (10)
Cerebrovascular disease	31 (7)	22 (5)
Chronic infection present at recruitment	42 (9)	38 (8)
Flu jab in last year	432 (95)	423 (93)
Place of residence		
Community	440 (97)	439 (97)
Nursing home	16 (3)	15 (3)
Housing tenure		
Owner occupier	340 (75)	332 (73)
Public sector tenant	88 (19)	92 (20)
Other	28 (6)	30 (7)
Nutrient at high risk of being deficient		
Iron	73 (16)	37 (8)
Folate	25 (6)	21 (5)
Vitamin C	58 (13)	59 (13)
Vitamin D	70 (15)	49 (11)
At risk for any of above	145 (32)	117 (26)

The majority of patients in the MAVIS trial responded to the initial mailing of the follow-up questionnaire and did not need prompting by reminder (Table 3.7). At six months an additional 4% responded by reminder, increasing to 11% at twelve months. The response rate was high initially, but the use of reminders increased

the response rate to over 90%. Some questionnaires were not sent due to withdrawal and in a few cases, due to death (N=12).

Table 3.6: MAVIS - QoL scores by treatment group

	Supplement Group		Placebo Group		Mean difference	
	N	Mean (SD)	N	Mean (SD)	(95% CI)	p-value
EQ5D						
Baseline	455	0.75 (0.2)	453	0.78 (0.2)		
6 months	433	0.77 (0.2)	422	0.80 (0.2)	-0.013 (-0.034,0.006)	0.18
12 months	421	0.77 (0.2)	409	0.80 (0.2)	-0.019 (-0.040, 0.002)	0.08
SF12 physical component score (PCS)						
Baseline	454	42.8 (11.4)	452	43.8 (10.7)		
6 months	431	43.5 (11.0)	419	44.9 (9.9)	-0.40 (-1.33,0.53)	0.40
12 months	415	43.7 (11.1)	412	44.3 (10.5)	0.07 (-0.90,1.03)	0.89
SF12 mental component score (MCS)						
Baseline	454	53.4 (8.7)	452	54.0 (8.7)		
6 months	431	52.8 (9.3)	419	53.8 (8.6)	-0.64 (-1.66,0.40)	0.23
12 months	415	53.2 (9.1)	412	53.6 (9.2)	-0.03 (-1.11,1.05)	0.96

Table 3.7: MAVIS – Number (%) of each responder type

	Follow up	
	6 months	12 months
Immediate responders	830 (91)	738 (81)
Reminder responders	34 (4)	101 (11)
Sent –not returned	6 (1)	6 (1)
Not sent	40 (4)	65 (7)
Total	910 (100)	910 (100)
% to respond by reminder	4%	12%

3.4 RECORD

The RECORD trial was a placebo-controlled trial of daily oral vitamin D and calcium supplementation in the secondary prevention of osteoporosis related fractures in older people (The RECORD Trial Group 2005). Quality of life was assessed at 4, 12, 24, 36 and 48 months using the EQ5D and the SF12 instruments. A number of demographic characteristics collected at baseline are presented in Table 3.8. There were 5,292 patients recruited to the trial. The majority (85%) were female, 67% were less than 80 years old, 94% could walk outdoors unaccompanied and 62% had arm fractures. It had been less than 90 days since

the recruiting fracture for 82% and 87% of participants were resident in their own home prior to fracture. There were four treatment groups (placebo, calcium only, vitamin D only, calcium and vitamin D). This resulted in two treatment comparisons of interest: vitamin D versus no vitamin D; calcium versus no calcium.

Table 3.8: RECORD - Baseline characteristics of participants N (%)

	Calcium and Vitamin D (N=1306)	Vitamin D only (N=1343)	Calcium only (N=1311)	Placebo (N=1332)	Total
Age group					
70-74	448 (34)	499 (37)	468 (36)	502 (38)	1917 (36)
75-79	440 (34)	403 (30)	423 (32)	399 (30)	1665 (31)
80-84	258 (20)	259 (19)	254 (19)	259 (19)	1030 (19)
85+	160 (12)	182 (14)	166 (13)	172 (13)	680 (13)
Sex					
Female	1104 (85)	1136 (85)	1113 (85)	1127 (85)	4480 (85)
Marital Status					
Single	81 (6)	97 (7)	102 (8)	68 (5)	348 (7)
Married	529 (41)	523 (39)	501 (38)	516 (39)	2069 (39)
Divorced	54 (4)	48 (4)	59 (5)	61 (5)	222 (4)
Widow(er)	639 (49)	671 (50)	643 (49)	681 (51)	2634 (50)
Locomotor ability					
Can walk unaccompanied	1221 (94)	1271 (95)	1232 (94)	1255 (94)	4979 (94)
Recruiting Fracture					
Proximal femur	228 (17)	231 (17)	222 (17)	223 (17)	904 (17)
Other leg and pelvic	285 (22)	255 (19)	308 (24)	282 (21)	1130 (21)
Distal arm	452 (35)	472 (35)	460 (35)	462 (35)	1846 (35)
Other arm	339 (26)	383 (29)	319 (24)	362 (27)	1403 (27)
Other	2 (<1)	2 (<1)	2 (<1)	3 (<1)	9 (<1)
Time since recruiting fracture					
Up to 90 days	1072 (82)	1099 (82)	1070 (82)	1090 (82)	4431 (82)
Residence type prior to fracture					
Own home	1137 (87)	1182 (88)	1148 (88)	1161 (87)	4628 (87)
Sheltered Housing	137 (11)	133 (10)	133 (10)	135 (10)	538 (10)
Other	32 (2)	28 (2)	30 (2)	36 (3)	126 (2)
Residence type after fracture					
Own home	1118 (86)	1159 (86)	1138 (87)	1140 (86)	4555 (86)
Sheltered Housing	134 (10)	131 (10)	132 (10)	134 (10)	531 (10)
Other	54 (4)	53 (4)	41 (3)	58 (4)	206 (4)

Each patient was followed up for at least two years. In addition to this, those patients recruited early on in the trial were followed up at three and four years. For this reason, in chapter four onwards only the first three assessments in the

RECORD trial will be utilised. Some patients were not sent follow-up questionnaires, as they had either died or had withdrawn from the trial. Table 3.9 details the number of patients responding at the five assessments. Of those patients sent the questionnaire, the proportion of patients responding increases with the length of follow up. The proportion of immediate responders also increases with time from 63% at four months to 74% at four years. At four months, 25% of the responders were reminder responders, reducing to 19% at four years.

Table 3.9: RECORD - Number (%) of each responder type

	Month of assessment				
	4	12	24	36	48
Immediate responders	3081 (63)	2854 (67)	2695 (72)	1793 (74)	716 (74)
Reminder responders	1051 (21)	898 (21)	741 (20)	440 (18)	169 (18)
Sent – not returned	817 (17)	499 (12)	329 (8)	181 (8)	76 (8)
Total sent	4949 (100)	4251 (100)	3765 (100)	2414 (100)	961 (100)
Total available for assessment	5292	5292	5292	3663	1629
Not sent	343 (6)	1041 (20)	1527 (29)	2878 (54)	4331 (82)

The primary trial analysis for the QoL outcomes in RECORD was an ANCOVA of the 24-month treatment difference (calcium versus no calcium or vitamin D versus no vitamin D). This difference was adjusted for the baseline QoL score, sex, age, time since recruiting fracture and whether or not the recruiting fracture was of the proximal femur or of the distal forearm (both as binary variables).

Table 3.10: RECORD - Mean (SD) and treatment difference (calcium)

	With calcium	Without calcium	Treatment difference (95% CI)	p-value
EQ5D				
4 months	N=1917	N=1990		
	0.7 (0.2)	0.7 (0.3)		
2 years	N=1546	N=1658		
	0.7 (0.3)	0.7 (0.3)	0.015 (0.00, 0.03)	0.05
SF12 Physical score component score (PCS)				
4 months	N=1791	N=1848		
	41.0 (11.0)	40.4 (11.2)		
2 years	N=1537	N=1612		
	41.1 (10.7)	41.3 (11.5)	0.44 (-0.17,1.05)	0.16
SF12 Mental component score (MCS)				
4 months	N=1791	N=1848		
	50.1 (10.7)	50.5 (10.2)		
2 years	n=1537	n=1612		
	50.3 (10.7)	50.5 (10.2)	0.03 (-0.63,0.68)	0.94

Table 3.10 displays the calculated treatment difference (95% confidence interval) along with the mean (SD) QoL scores for the calcium treatment comparison. Table 3.11 shows the same details for the Vitamin D treatment comparison. There was no statistical difference in QoL outcomes observed between treatment groups at two years (The RECORD Trial Group 2005).

Table 3.11: RECORD - Mean (SD) and treatment difference (Vitamin D)

	With Vitamin D	Without Vitamin D	Treatment difference (95% CI)	p-value
EQ5D				
4 months	N=1984	N=1923		
	0.7 (0.3)	0.7 (0.3)		
2 years	N=1264	N=1226		
	0.7 (0.3)	0.7 (0.3)	-0.002 (-0.017,0.013)	0.81
SF12 Physical component score (PCS)				
4 months	N=1833	N=1806		
	40.8 (11.0)	40.6 (11.3)		
2 years	n=1588	n=1561		
	41.2 (11.2)	41.1 (11.1)	0.135 (-0.48,0.74)	0.91
SF12 Mental component score (MCS)				
4 months	N=1833	N=1806		
	50.5 (10.5)	50.1 (10.4)		
2 years	n=1588	n=1561		
	50.6 (10.4)	50.2 (10.4)	-0.03 (-0.69, 0.62)	0.64

3.5 KAT

The KAT trial was a multi-centre study involving 117 surgeons in 34 centres. The trial was designed to measure the long-term clinical and cost effectiveness of different types of knee replacement (The KAT trial group 2009). There were overlapping trials evaluating four developments in knee surgery. Individual patients could participate in a maximum of two comparisons. The four comparisons were:

1. Is a metal backing plate for tibial component of the total knee replacement better than a single high density polyethylene component? (N=262)
2. Is it better to resurface the patella as part of a knee replacement or not? (N=1715)

3. Does a polyethylene moving component (bearing) between the tibia and femur have a better outcome than standard designs without a moving bearing? (N=539)
4. Is a unicompartmental arthroplasty better than a total knee arthroplasty? (N=34).

Functional status (Oxford Knee Score, OKS) and quality of life (SF12 and EQ5D) were measured at baseline, three months and annually after the operation. There were 2356 patients involved in the trial. However, some patients were involved in more than one treatment comparison.

Table 3.12: KAT – Baseline participant information (N=2356)

Characteristic	Mean (SD)	
Age (years)	69.9 (8.4)	
Weight (kg)	80.9 (16.6)	
Height (cm)	164.4 (13.4)	
Primary type of knee arthritis (N=2267)	**N**	**(%)**
Osteoarthritis	2157	(95)
Rheumatoid	106	(5)
Both	4	(<1)
Extent knee arthritis (N=2352)		
One knee	600	(25)
Both knees	946	(40)
General	806	(34)
Gender (N=2353)		
Male	1029	(44)
Female	1323	(56)
ASA Grade (N=2176)		
Completely Fit and Healthy	372	(17)
Some illness but has no effect on normal activity	1345	(62)
Symptomatic illness present but minimal restriction	444	(20)
Symptomatic illness causing severe restriction.	15	(1)
Any post operative complications (N=2240)		
Yes	330	(15)
No	1910	(85)
Any readmissions		
Yes	182	(8)
No	2174	(92)

ASA – American Knee Society grade

The treatment related variables collected at baseline included: primary type of knee arthritis, extent of knee arthritis, conditions affecting mobility, previous knee

surgery, American Knee Society (ASA) grade, post-operative complications and re-admissions (Table 3.12). The majority of patients (95%) had osteoarthritis of the knee, with 40% displaying arthritis in both knees. Most patients had no post-operative complications (85%) or re-admissions (92%).

At baseline assessment 2242 (95%) questionnaires were returned. EQ5D scores were available at baseline for 2195 patients (98% of those returned). For the SF12 component scores, 2162 (96% of those returned) were available. Table 3.13 details the number (%) of participants returning the questionnaire at each of the follow up assessments. The proportion of questionnaires not sent was between 7-9%. Only a small proportion of those questionnaires sent were not returned (<10%).

Table 3.13: KAT - Number (%) of each responder type

	3 months	1 year	2 years
Immediate responders	1848 (78)	1746 (74)	1618 (69)
Reminder responders	222 (9)	314 (13)	349 (15)
Not returned	78 (3)	136 (6)	219 (9)
Not sent	208 (9)	160 (7)	170 (7)
Total	2356	2356	2356

The proportion of responding patients who did so after reminder was 11% at three months, 15% at one year and 18% at two years. The primary trial analysis for treatment difference in the QoL outcomes was assessed at one and two years. An ANCOVA adjusting for baseline QoL, site {one knee, both knees, general}, age group {<60, 60-80, 80+} and sex was undertaken. Appendix 3.1 describes these results split by treatment comparison. No significant differences between treatment groups were found for each of the four QoL scores at either one or two years follow up.

3.6 PRISM

The PRISM trial was designed to evaluate the clinical effectiveness and cost-effectiveness of symptomatic versus intensive bisphosphonate therapy for the management of Paget's disease (Ralston et al. 2006). Paget's disease of the bone is a common skeletal disorder. It is characterised by focal increases in bone

turnover, Paget's disease is a cause of substantial morbidity, casing diverse symptoms such as bone pain, pathological fracture, deafness, bone deformity and secondary osteoarthritis (Siris 1998). The philosophy of the symptomatic treatment is that no deliberate attempt is made to restore alkaline phosphatase to within the normal range. The aim of the intensive treatment is to keep alkaline phosphatase within the normal range. The treatment was administered for a period of between two and five years.

Table 3.14: PRISM - Baseline characteristics (N=1324)

	Yes N (%)	No N (%)	Not sure N (%)
Bone Scan (N=1320)	822 (62)	498 (38)	
Bones Involved in Paget's disease			
Skull (N=1320)	313 (24)	1007 (76)	
Pelvis (N=1320)	885 (67)	435 (33)	
Tibia (N=1319)	248 (19)	1071 (81)	
Spine (N=1319)	517 (39)	802 (61)	
Femur (N=1320)	430 (33)	890 (67)	
Hearing Aid? (N=1317)	296 (23)	1021 (77)	
Bone Pain?			
Is there bone pain? (N=1323)	911 (69)	359 (27)	53 (4)
Patient reported (N=911)	906 (99)	4 (<1)	1 (<1)
Clinician reported (N=904)	618 (67)	155 (17)	131 (14)
Fractures			
Any? (N=1322)	518 (39)	804 (61)	
in pagetic bone (N=618)	133 (26)	385 (74)	
in non-pagetic bone (N=524)	417 (79)	107 (21)	
clinical vertebrae fractures (N=518)	51 (10)	467 (90)	

The trial involved people aged over 18 with symptomatic or asymptomatic Paget's disease. At recruitment, some patient characteristics (e.g. age and gender) and some baseline clinical information (e.g. family history, age at diagnosis, presence of bone pain) were collected (Table 3.14 and Table 3.15). There were 1331 patients recruited to the trial, but seven were excluded post-randomisation leaving 1324 patients for follow up – with 695 (52%) males and 629 (47%) females. The majority were married (64%), with 26% widowed. The remaining patients were single (5%), divorced (4%) or living with partner (1%). There were 849 (64%) patients with polystotic Paget's disease and 472 (36%) with monostotic Paget's disease.

QoL was assessed through postal questionnaires which included the EuroQoL EQ5D and SF36 instruments. Disease specific QoL was measured by the Arthritis Specific Health Index (ASHI) which is calculated from the SF36 instrument using different item weightings. QoL data was collected at baseline, 12 and 24 months and then annually up to five years as appropriate.

Table 3.15: PRISM - Level of deformity for bones of Paget's disease N (%)

Level of deformity	No deformity	Mild/moderate	Severe
Skull (N=1320)	1220 (92)	82 (6)	18 (1)
Femur (N=1320)	1155 (87)	140 (11)	25 (2)
Tibia (N=1321)	1129 (85)	138 (10)	54 (4)
Other bones (N=1318)	1141 (86)	150 (11)	27 (2)

Table 3.16 details the numbers (%) of responders and non-responders to the yearly follow-up questionnaires. The number of patients that were actually sent the follow-up questionnaire decreases from 95% at year one to only 5% at year four. The proportion of non-response from those patients sent the questionnaire was small (4% -7%). Reminder response as a proportion of the responders was 6% at year one, 19% at year two, 29% at year three and 20% at year four.

Table 3.16: PRISM – Number (%) of each responder type

	Year 1	**Year 2**	**Year 3**	**Year 4**
Immediate responders	1131 (90)	837 (76)	414 (66)	55 (76)
Reminder responders	77 (6)	190 (17)	168 (27)	14 (19)
Not returned	48 (4)	80 (7)	41 (7)	3 (4)
Total Sent	1255 (100)	1107 (100)	623 (100)	72 (100)
Not sent (% of total)	69 (5)	217 (16)	701 (53)	1252 (95)

Table 3.17 displays the mean (SD) QoL scores for each assessment. QoL seemed to remain fairly stable, especially within the first 3 years.

Table 3.17: PRISM - Mean (SD) QoL scores

		SF36 scores				
		PCS	**MCS**	**ASHI**	**EQ5D**	
	N (%)	Mean (SD)	Mean (SD)	Mean (SD)	N (%)	Mean (SD)
Baseline	1203 (91)	36.3 (11.3)	48.7 (11.8)	35.8 (12.6)	1296 (98)	0.58 (0.30)
1 year	954 (72)	35.9 (10.8)	47.4 (12.0)	36.0 (12.0)	1091 (82)	0.59 (0.29)
2 year	873 (66)	36.1 (11.1)	47.1 (12.0)	35.9 (12.5)	992 (75)	0.60 (0.30)
3 year	504 (38)	35.8 (10.8)	46.8 (12.5)	35.2 (12.3)	560 (42)	0.59 (0.31)
4 year	63 (5)	37.1 (11.0)	47.4 (11.5)	36.9 (12.0)	67 (5)	0.62 (0.31)

PCS: physical component score; MCS – mental component score; ASHI – Arthritis Index

Physical functioning was about 36 units, mental functioning 47 to 48 units and the EQ5D overall health status was 0.58 to 0.60 units. There was a slight rise in this at four years, probably due to the much reduced sample size, which most likely includes those who were most well (as indicated by the slightly inflated scores).

The primary trial analysis for the QoL outcomes consisted of an ANCOVA for treatment difference in the two year QoL scores adjusting for baseline QoL and the minimisation variables (as indicator variables). The minimisation variables were: previous treatment with bisphosphonates; Paget's in a weight bearing limb; deformity due to Paget's; Paget's in the skull; presence of pain perceived to be due to Paget's; serum alkaline phosphatase level (indicators for normal elevated and greatly elevated). The results for the four QoL outcomes are shown in Table 3.18. No significant difference in QoL were found between the symptomatic and intensive treatment groups ($p>0.05$).

Table 3.18: PRISM - ANCOVA analysis results

	Symptomatic			Intensive			Treatment	
	N	**Mean**	**(SD)**	**N**	**Mean**	**(SD)**	**Difference (95% CI)**	**p-value**
SF36 Physical component score (PCS)								
Baseline	595	35.8	(11.3)	603	36.7	(11.3)		
2 year	429	35.6	(11.0)	437	36.5	(11.0)	0.137 (-0.90,1.17)	0.80
SF36 Mental component score (MCS)								
Baseline	595	48.6	(11.9)	603	48.7	(11.8)		
2 year	429	46.8	(12.1)	437	47.4	(11.9)	0.695 (-0.65,2.04)	0.31
ASHI								
Baseline	595	35.4	(12.8)	603	36.2	(12.3)		
2 year	429	35.4	(12.2)	437	36.3	(12.5)	-0.037 (-1.22,1.14)	0.95
EQ5D								
Baseline	628	0.57	(0.3)	622	0.60	(0.3)		
2 year	449	0.57	(0.3)	440	0.60	(0.3)	0.015 (-0.02,0.05)	0.38

3.7 TOMBOLA

The TOMBOLA trial compared different strategies for the management of borderline and other low-grade abnormal smears (Whynes et al. 2008).

Table 3.19: TOMBOLA - Baseline characteristics of participants (N=3399)

Characteristic	N (%)
Trial Centre (N=3399)	
Grampian	1103 (32)
Tayside	849 (25)
Nottingham	1447 (43)
Age group (N=3381)	
20-29	1457 (43)
30-39	902 (27)
40-49	718 (21)
50-59	304 (9)
Ethnic Group (N=3377)	
White	3228 (96)
Training (N=3375)	
None	906 (27)
Through work	667 (20)
University/college degree	971 (29)
Qualification other than degree	831 (25)
Marital Status (N=3356)	
Married/Living as married	1870 (56)
Divorced/Separated	414 (12)
Single	1036 (31)
Widowed	36 (1)
Employment Status (N=3382)	
Full time	1679 (50)
Part time	789 (23)
Student	313 (9)
Not in paid employment	601 (18)
Smoking Status (N=3969)	
Never smoked	1608 (41)
Ex-smoker	576 (15)
Current smoker	1777 (45)
Smoked at some point	8 (0)
Activity level (N=3347)	
Never	469 (14)
Rarely	693 (21)
<once a week	169 (5)
once a week	230 (7)
2-3times a week	558 (17)
>3 times a week	1228 (37)
Eligible smear (N=3399)	
Mild	910 (27)
borderline nuclear abnormalities (BNA)	2489 (73)
Human Papilloma virus (HPV) status (N=3399)	
Negative	1785 (53)
Positive	1236 (36)
No sample	378 (11)

The trial compared repeat smears with colposcopy examinations. There were two levels of randomisation in this trial: the first between cytological surveillance in primary care (repeated smears) and hospital based colposcopy examination. Within the colposcopy arm, a second randomisation occurred between biopsy and selective recall or immediate treatment. There were 4439 women recruited to the trial. However, QoL data was available for only 3399 women. This was because the QoL aspects of the data were not collected during the first year of the trial.

There were 1103 (32%) patients recruited in Grampian, 849 (25%) in Tayside and 1447 (43%) in Nottingham. Patient characteristics collected at baseline are shown in Table 3.19. The QoL measure administered in this trial was the EuroQoL EQ5D. Data was collected at the recruitment appointment (baseline), 6 weeks (to a subset of patients) and then, 12, 18, 24 and 30 months after recruitment. Of the 3399 patients, 31 (1%) provided no QoL data at all. They provided only demographic information. The remaining 3368 provided at least one measure of QoL, with 1442 (42%) providing all five scheduled assessments. The only deaths in the sample occurred after the 30 month assessment.

Table 3.20 displays the completion information for the TOMBOLA trial at each follow-up assessment. The proportion of patients not completing questionnaires increases to 45% at the final 30 month assessment. Of those who provided QoL scores, the proportion of reminder-responders varies with 24% at 6 weeks, 26% at 12m, 27% at 18m and 26% at 24m and 30m.

Table 3.20: TOMBOLA – Number (%) of each responder type

	Completed		
Assessment	**No reminder**	**Reminder**	**Not Completed**
Baseline	3300 (97)	0 (0)	99 (3)
6 weeks*	1335 (39)	417 (12)	1647 (48)
12m	1725 (51)	612 (18)	1062 (31)
18m	1543 (45)	562 (16)	1294 (38)
24m	1492 (44)	513 (15)	1394 (41)
30m	1372 (40)	488 (14)	1539 (45)

*The 6-week questionnaire was only issued to a subset of patients

The EQ5D scores split by responder type are shown in Table 3.21. The mean EQ5D score tended to be lower for reminder-responders than for the immediate-responders.

Table 3.21: TOMBOLA – Mean (SD) EQ5D scores

	Immediate responders		Reminder-responders	
	N (%)	Mean (SD)	N (%)	Mean (SD)
Baseline	3300	0.621 (0.14)	n/a	n/a
12 months	1725 (74)	0.630 (0.15)	612 (26)	0.626 (0.14)
18 months	1543 (73)	0.631 (0.15)	562 (27)	0.622 (0.15)
24 months	1492 (74)	0.634 (0.14)	513 (26)	0.623 (0.15)
30 months	1372 (74)	0.630 (0.14)	488 (26)	0.623 (0.15)

The TOMBOLA trial group collected the EQ5D data to enable calculation of quality adjusted life years (QALYs) for use in their cost-utility analysis (not yet published). The EQ5D data was not collected with analysis of treatment difference in mind and only a simplistic method was carried out (Whynes et al. 2008). For the thesis purposes, an ANCOVA model adjusting for baseline EQ5D score and the minimisation variables (age group, trial centre, eligible smear status and HPV status) was carried out. Results are presented for the 12 month and 30 month endpoints in Table 3.22.

Table 3.22: TOMBOLA - ANCOVA results EQ5D score

Comparison	Endpoint	N	Treatment Difference (95% CI)	p-value
R1	12 months	2294	-0.001 (-0.011, 0.008)	0.81
R1	30 months	1825	0.004 (-0.007, 0.015)	0.49
R2	12 months	709	0.019 (0.002, 0.035)	0.03
R2	30 months	584	0.006 (-0.014, 0.026)	0.56

R1: Cytological surveillance versus colposcopy
R2: Biopsy and selected recall versus immediate treatment

The first comparison (R1) looked at colposcopy versus cytological surveillance. No treatment difference was found at either 12 or 30 months ($p>0.05$). The second comparison looked at biopsy and selective recall versus immediate treatment following the patients' colposcopy examination. A difference of 0.02 (0.002, 0.035) units was found to be significant ($p=0.03$) between the two arms, with the biopsy and selective recall arm displaying slightly better QoL than the immediate treatment arm.

3.8 Norwegian Palliative Care Trial (NPC Trial)

The aim of the Norwegian Palliative care (NPC) trial was to evaluate the impact of a comprehensive palliative care intervention on advanced cancer patient's quality of life, location of care and place of death (Jordhøy et al. 2001).

Table 3.23: NPC Trial - Baseline characteristics of patients (N=434)

Characteristic	Total	Control	Intervention
	N (%)	N (%)	N (%)
Sex - Male	230 (53)	98 (49)	132 (56)
Marital Status			
Alone	31 (7)	20 (10)	11(5)
Married	275 (63)	117 (59)	158 (67)
Divorced	30 (7)	12 (6)	18 (8)
Widowed	98 (23)	50 (25)	48 (2)
Education years			
<=7	160 (37)	67 (34)	93 (40)
8-10	149 (34)	77 (39)	72 (31)
11-12	62 (14)	22 (11)	40 (17)
13+	63 (15)	33 (17)	30 (13)
Working Status			
Working	13 (3)	7 (3)	6 (2)
Sick Leave	82 (19)	31 (16)	51 (22)
Pension	339 (78)	161 (81)	178 (76)
Baseline Karnofsky score			
<=40	7 (2)	3 (1)	4 (1)
40-70	163 (37)	79 (40)	84 (36)
>70	264 (61)	117 (59)	147 (63)
Place of death (N=391))			
Hospital	257 (66)	113 (65)	144 (66)
Nursing Home	55 (14)	36 (21)	19 (9)
Own home	79 (20)	25 (14)	54 (25)
Randomisation pair			
Byheimsesau	250 (57)	116 (58)	134 (57)
Narstri	142 (33)	65 (33)	77 (33)
Malmel	42 (10)	18 (9)	24 (10)
Use alternative medicine? – Yes	38 (9)	21 (11)	17 (7)
Chemotherapy now? - Yes	59 (14)	32 (16)	27 (11)
Radiation therapy now? - Yes	49 (11)	27 (14)	22 (9)
Hormone therapy now? - Yes	55 (13)	29 915)	26 (11)
Stage of disease			
Local	67 (15)	36(18)	31 (13)
Regional	24 (6)	13 (7)	11 (5)
Metastasis	343 (79)	150 (75)	193 (82)

The trial was conducted within the Norwegian Public Health Care at the University Hospital Trondheim. It was a clustered trial with two rural and two urban areas. The comparison was between conventional care and a more comprehensive care package offered by the Palliative Medicine Unit. The baseline characteristics of the patients are shown in Table 3.23. There were 434 patients recruited to the trial with just over half the patients male (55%). The majority of patients were receiving a pension (78%), were married (63%), had a baseline Karnofsky performance status score of over 70 (61%) and their stage of disease was metastasis (79%).

QoL was assessed on the monthly questionnaire, which included the European Organisation for the Research and Treatment of Cancer (EORTC) quality of life questionnaire (QLQ), the QLQ-C30. The dimensions of pain, physical functioning and emotional functioning were the main QoL endpoints for the trial. All follow-up questionnaires (apart from baseline) were administered by post with a single reminder issued, if they had not responded within two weeks. If no answer was obtained, the patients received no further questionnaires and were regarded as dropouts. This differed to the previous trials where questionnaires at a subsequent assessment were still sent following a missed assessment, as long as the patient had not officially withdrawn from the trial. In the NPC trial an intermittent missing pattern occurred for 28 participants. This was due to missing items, rather than missing forms.

Table 3.24: NPC Trial – Number (%) of each responder type

Assessment	Immediate responders	Reminder responders	Sent but not returned	Not Sent
Baseline	434 (100)	0	0	0
1m	203 (47)	66(15)	68 (16)	97 (22)
2m	149 (34)	51 (12)	41 (9)	193 (44)
3m	130 (30)	35 (8)	22 (5)	247 (57)
4m	98 (23)	38 (9)	17 (4)	281 (65)
5m	100 (23)	24 (6)	14 (3)	296 (68)
6m	80 (18)	27 (6)	13 (3)	314 (72)

A large proportion of patients died during the trial, by nature of the disease group. The proportion of patients being sent questionnaires reduced to 43% at 3

months and further still to 28% at 6 months. The published trial analysis used data collected up to 4 months, since after this time a large proportion of patients had died (Jordhøy et al. 2001). Table 3.24 summarises the completion information for the NPC trial. The proportion of reminder-responders from all the responders was 25% at 1m, 26% at 2m, 21% at 3m and 28% at 4m.

Table 3.25 shows the mean (SD) for each of the three QoL dimensions at each assessment split by type of responder (immediate or reminder). The trial was designed as a clustered trial and analysed using area under the curve (AUC) for each QoL scale as a summary measure, to avoid multiple comparisons and to evaluate both early and continuous effects. This was based on changes from baseline and if data from one assessment were missing, the mean of the two adjacent ones was used. QoL was assumed to be zero after death. For patients who withdrew or dropped out during the first four months, the last value carried forward (LVCF) was used to impute missing data.

Table 3.25: NPC Trial -Mean (SD) sores for three QLQ-C30 dimensions

Type of responder		Month of follow-up				
		Baseline	**1**	**2**	**3**	**4**
Pain (PA)						
Immediate	N	434	203	149	130	98
	Mean (SD)	47.3 (35.5)	37.4 (32.9)	37.1 (31.0)	37.9 (33.2)	41.8 (33.8)
Reminder	N	N/A	66	51	35	38
	Mean (SD)	N/A	32.3 (28.9)	37.3 (32.6)	31.4 (31.0)	31.6 (28.9)
Physical Functioning (PF)						
Immediate	N	431	203	148	130	97
	Mean (SD)	46.8 (30.6)	48.4 (31.2)	50.9 (31.2)	53.3 (32.1)	53.8 (32.2)
Reminder	N	N/A	66	51	35	38
	Mean (SD)	N/A	45.1 (32.0)	52.6 (32.1)	49.1 (33.0)	45.8 (33.8)
Emotional Functioning (EF)						
Immediate	N	433	201	148	130	96
	Mean (SD)	66.2 (25.7)	70.9 (24.3)	73.3 (23.7)	74.1 (22.7)	74.7 (23.5)
Reminder	N	N/A	66	50	33	37
	Mean (SD)	N/A	75.4 (22.0)	66.0 (28.3)	72.6 (24.4)	68.8 (21.9)

Since the reported trial analysis contains some imputation already, it was decided not to use this method here. Instead, the reference analysis used was ANCOVA at one and four months in order to test treatment difference, adjusting for baseline QoL, sex, age group, randomisation cluster and Karnofsky performance status. The time points of one and four months were chosen as a reasonable sample size

was maintained up to four months and one month was considered consistent with an intervention period likely to show a clinically significant effect (Jordhøy et al. 1999).

Table 3.26: NPC Trial - ANCOVA results for QLQ-C30 dimensions

Assessment	N	Treatment Difference	95% CI	p-value
Pain				
1 month	269	2.73	(-4.04, 9.50)	0.43
4 months	136	3.26	(-6.69, 13.2)	0.52
Physical functioning				
1 month	268	-1.10	(-6.84, 4.63)	0.71
4 months	135	-5.68	(-15.1, 3.73)	0.23
Emotional functioning				
1 month	266	3.08	(-1.88, 8.03)	0.22
4 months	132	1.35	(-4.97, 7.66)	0.67

Table 3.26 shows that there was no difference between treatments for the pain, physical functioning and emotional functioning dimensions of the EORTC QLQ-C30 ($p>0.05$ all cases). This was consistent with the findings by Jordhøy *et al*., when the AUC approach was used (Jordhøy et al. 2001).

3.9 Overview

The seven example trials varied in sample size, QoL measures used, number of assessments and response rates (Table 3.27). It is easily seen that the use of reminders increased the overall response rate at the main endpoint and the impact was greatest in REFLUX. The analysis strategy for each of these trials was to carry out an ANCOVA at a single endpoint adjusting for baseline QoL and some other patient characteristics. This process was undertaken as a complete case analysis and thereby ignored the missing data. The implications of this approach were discussed in chapter one and will be discussed further in later chapters. An assumption of the complete-case analysis is that the data are missing completely at random (MCAR). The appropriateness of this assumption is investigated in chapter five. The amount of missing data has an important role to play in deciding on the most appropriate strategy to deal with the missing data. For those trials in which there is not much missing data, for example MAVIS and REFLUX,

the complete-case analysis strategy might be more appropriate. However, in RECORD, the amount of missing data at the final endpoint reaches 40% and therefore, ignoring all these participants has the potential to considerably bias the results. The impact of the amount of missing data is considered in the chapters that follow.

Table 3.27: Summary of the trial datasets

Trial	N	QoL measures	Number of assessments*	Main endpoint	Response rate at main endpoint	
					Immediate	Reminder
REFLUX	357	EQ5D RQLS SF12	3	1 year	38%	51%
MAVIS	910	EQ5D SF12	3	1 year	81%	11%
RECORD	5292	EQ5D SF12	5	2 years	51%	14%
KAT	2356	EQ5D SF12 OKS	4	2 years	69%	15%
PRISM	1324	EQ5D SF36	5	2 years	63%	14%
TOMBOLA	3399	EQ5D	6	12 months 30 months	51% 40%	18% 14%
NPC Trial	434	QLQ-C30	5	1 month 4 months	47% 23%	15% 9%

RQLS – Reflux quality of life score; OKS – Oxford Knee Score; * including baseline

The seven trial datasets described here will be utilised in subsequent chapters. The main feature of these datasets that will be important throughout this work is whether or not a patient completed the follow up questionnaire after receiving a reminder. The mechanism of missingness will be identified for each trial in chapter five using several methods presented in chapter four. Accuracy of both simple and multiple imputation is investigated in chapter seven making use of the responses obtained through reminders. Chapter nine considers alternative analysis strategies to the ANCOVA presented for each of the trials. Chapter ten makes use of the information on the proportion of reminder responses in each of the trials to determine whether the extra effort involved in reminder questionnaires is cost-effective in terms of the additional data it generates.

Chapter 4 Methods to investigate the mechanism of missing data

4.1 Introduction

Chapter one introduced the concept of the missing data mechanism and stated that it is the underlying reason why the data are missing. The data could be missing simply because the respondent has moved, or it could be that a particular treatment is affecting QoL and thus, increasing the chance of non-response. The review in chapter two highlighted that in published clinical trials only one of the 61 studies using imputation had discussed the missing data mechanism. This was despite the fact that knowing why the data are missing can be helpful in identifying the most suitable analysis strategy (Curran et al. 1998a).

In chapter three, the reported trial results using analysis of covariance (ANCOVA) were presented. This analysis strategy was based on the complete cases and makes the assumption that the missing data are MCAR. The next two chapters aim to investigate whether this assumption was appropriate in each of the example trials. The current chapter firstly defines the missing data mechanism and details a number of methods which can be used to identify the mechanism. Chapter five then utilises the methods identified and aims to investigate the missing data mechanism in each of the seven trials. Some notation to describe the methods to investigate the missing data mechanism is now outlined.

4.1.1 Notation

Firstly, consider a study with J measurements of the outcome (e.g. QoL score). The complete data matrix Y is defined as $Y = (y_{ij})$ where y_{ij} is the value of variable Y_j for the i^{th} subject. Now let Y_{obs} denote the observed components of Y, and Y_{mis} the missing components. The matrix R defines the pattern of missing data or "missingness" and is defined as $R = (r_{ij})$ where $r_{ij} = 0$ if y_{ij} is missing and $r_{ij} = 1$ if y_{ij} is observed. It follows that R_i is the vector of indicators of the missing data pattern for the i^{th} individual. For example, in a study of three assessments where the QoL

of the i^{th} subject was missing at the second assessment, but observed as 67 units at the first assessment and 54 units at the third assessment, then the subject's data would be

$$Y_i = \begin{bmatrix} 67 \\ . \\ 54 \end{bmatrix}, \; R_i = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}.$$

The missing data pattern for each patient can be identified. Let P be the number of distinct missing data patterns and $J^{\{p\}}$ is the number of observed variables in pattern p. For example, in a study with J = 3 assessments and all were observed in the first pattern then $J^{\{1\}} = 3$. In the example above, assuming this is the second pattern then $J^{\{2\}} = 2$, as two of the three assessments were observed. The number of cases with the p^{th} pattern is $n^{\{p\}}$ and $\sum n^{\{p\}} = N$. Let $M^{\{p\}}$ be a $J^{\{p\}} \; x \; J$ matrix of indicators of the observed variables in pattern p. The matrix has one row for each measure present consisting of (J-1) zeroes and a one identifying the observed measure. For example, in a study with three assessments where the first and third observation were obtained in the second pattern then

$$M^{\{2\}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The $J^{\{p\}} \; x \; 1$ vector of means of the observed variables for pattern p is then $\bar{Y}^{\{p\}}$. Using the available data (i.e. assuming that the missing data mechanism is ignorable), the maximum likelihood (ML) estimates of the mean and covariance of Y_i can be obtained. It follows from this that vector of ML estimates corresponding to the p^{th} pattern is $\hat{\mu}^{\{p\}} = M^{\{p\}}\hat{\mu}$ and $\tilde{\Sigma}^{\{p\}} = \frac{N}{N-1} M^{\{p\}} \hat{\Sigma} M^{\{p\}'}$ is the corresponding $J^{\{p\}} \; x \; J^{\{p\}}$ covariance matrix with a correction for degrees of freedom (Fairclough 2002, Little 1988).

4.1.2 Definition of the missing data mechanism

The standard definition of the missing data mechanism was originally proposed by Rubin (Rubin 1976). Three mechanisms were defined: missing completely at random (MCAR); missing at random (MAR); missing not at random (MNAR).

These mechanisms are now described using the notation defined above. The missing data mechanism is given by the conditional distribution of R given Y, say $f(R \mid Y, \phi)$, where ϕ denotes unknown parameters. If missingness does not depend on the values of the data Y, missing or observed, that is

$$f(R \mid Y, \phi) = f(R \mid \phi) \text{ for all } Y, \phi$$

then the data are said to be MCAR. For MAR data, missingness depends only on the observed components (Y_{obs}) of Y and not on the components that are missing (Y_{mis}). That is

$$f(R \mid Y, \phi) = f(R \mid Y_{obs}, \phi) \text{ for all } Y_{mis}, \phi.$$

Finally, MNAR occurs if the distribution of the missing data matrix R depends on the missing values in the data matrix Y, as well as, the observed components.

In summary, and in lay terms for MCAR, missingness does not depend on the variable of interest, or any others in the dataset. MAR is when missingness is conditional on another observed variable, but not on the outcome of interest. For MNAR, missingness depends on the actual (unobserved) data value. In the context of quality of life, the mechanism refers to whether missingness is somehow related to the quality of life. MCAR occurs if the missingness has nothing to do with QoL (past, present or future). It could be because the patient moved or the questionnaire was lost. Covariate dependent missingness also falls within this category. This occurs if say missingness varies between age groups, yet within each age group missingness is MCAR. When missingness is related to observed QoL, the mechanism is MAR. MNAR data occurs if missingness is related to unobserved QoL (past, present or future).

4.1.3 Pattern of missing data

The pattern of missing data is usually classified into one of two categories: monotone (terminal) or intermittent. Monotone missingness occurs when a patient is observed at every assessment (including baseline) until a time when they dropout out of the study and provide no further assessments. Intermittent missingness occurs if a missing observation occurs between one or more observed assessments. It is possible to have a mixed pattern, where a period of intermittent

missingness is followed by monotone missingness. Examples of these patterns were provided in Table 1.1.

4.2 Methods to investigate the missing data mechanism

A number of authors have proposed ways to investigate the missing data mechanism. Little developed a test based on the means of the variable of interest under the different missing data patterns (Little 1988). The null hypothesis is that data are MCAR. The test is appropriate for both monotone and intermittent missing data patterns. Diggle proposed an approach which tests whether the subset of patients about to dropout are a random sample of the whole population (Diggle 1989). This test requires monotone missingness. Ridout adopted a similar approach to Diggle by utilising logistic regression (Ridout 1991). Listing and Schlittgen proposed a test based on means, requiring monotone missingness (Listing, Schlittgen 1998). Following this, in 2003 they suggested a non-parametric procedure which combines several Wilcoxon rank sum tests (Listing, Schlittgen 2003). Schmitz and Franz discussed a non-parametric version of Listing and Schlittgen's test (Schmitz, Franz 2002). Fairclough adopts a logistic regression procedure similar to that of Ridout, but with an alternative response variable (Fairclough 2002). The difference between the two methods is discussed further in the sections that follow.

4.2.1 Little's test: test of missing completely at random

Little proposed a test of MCAR for multivariate data with missing values (Little 1988). He proposed a single global test statistic for MCAR that uses all available data. This test is based on the premise that if the data are MCAR the calculated means of the observed data should be the same for each pattern. If the data are not MCAR, it would be expected that the means would vary across the patterns.

Using the notation presented in section 4.1.1, the ML estimates of the mean vector and the covariance matrix for the p^{th} pattern are

$$\hat{\mu}^{\{p\}} = M^{\{p\}}\hat{\mu} \text{ and } \tilde{\Sigma}^{\{p\}} = \frac{N}{N-1} M^{\{p\}}\hat{\Sigma}M^{\{p\}'}.$$

These estimates are calculated on the complete data and assume the missing data mechanism is ignorable. Little's proposed test statistic takes the form

$$\chi^2 = \sum_{p=1}^{P} n^{\{p\}}\left(\bar{Y}^{\{p\}} - \hat{\mu}^{\{p\}}\right)'\tilde{\Sigma}^{\{p\}-1}\left(\bar{Y}^{\{p\}} - \hat{\mu}^{\{p\}}\right).$$

This test statistic was shown to be asymptotically χ^2 distributed with $\Sigma J^{\{p\}} - J$ degrees of freedom (Little 1988).

The first stage in calculating this test statistic is to obtain the ML estimates of the mean and covariance of the available data. Secondly, the data needs to be split into the P patterns and the value of $J^{\{p\}}$ obtained for each pattern. The vector of means of the observed variables for pattern p, $\bar{Y}^{\{p\}}$ should be calculated. For each of the p patterns $\hat{\mu}^{\{p\}}$ and $\tilde{\Sigma}^{\{p\}}$ are then calculated using the formulae above. Combining all these terms allows the calculation of Little's test statistic. Computation of this test statistic is not currently available in standard statistical software. Syntax originally provided by Fairclough (Fairclough 2002) was adapted to conform to that needed for SAS version 9.1 (SAS Institute Inc. 2004). This syntax is provided in Appendix 4.1.

4.2.2 Listing and Schlittgen: Tests if dropouts are missed at random

Listing and Schlittgen proposed a test to see if dropouts are missed at random based on the difference between means (Listing, Schlittgen 1998). This test requires monotone missingness and some additional notation to that provided in section 4.1.1 is needed. A dropout at assessment j is a patient that provides assessments up until time j but no further assessments from $j + 1$ up to time J. Let w_j indicate the number of dropouts at assessment j. The observation vectors y_i are arranged in a row such that the first n_J are observed at all time points. The next w_{J-1} vectors y_i are observed at all time points except the last one. They correspond to the individuals who are observed at time $j-1$ for the last time. Then the following w_{J-2} are observed at $j=1,\ldots,J-2$ and so on.

For each time point, a test can be based on the difference in the mean of the values of the individuals who continue to stay in the study and the mean of the values for the dropouts. Since an overall test statistic is being constructed, only the first n_J observations are used to determine the means of the non dropouts for every time point. This leads to the difference $\bar{y}_{1j} - \bar{y}_{2j}$, where

$$\bar{y}_{1j} = \frac{1}{n_J}\sum_{i=1}^{n_J} y_{ij}\,, \qquad \bar{y}_{2j} = \frac{1}{w_j}\sum_{i=1}^{w_J} y_{n_{j+i},j}$$

where $n_j = n_J + w_{J-1} + \ldots + w_{J+1}$ for $j<J-1$ and $n_j = n_J$ for $j=J-1$. These differences are combined with the number of observations belonging to the means of the dropouts to produce the statistic D, where

$$D = \frac{1}{w}\sum_{j=1}^{J-1} w_j\left(\bar{y}_{1j} - \bar{y}_{2j}\right) \qquad \text{where } w = w_1 + \ldots + w_{J-1}.$$

The idea of the test statistic is to combine the weighted differences of the means of dropouts and non-dropouts at the different time points. For the non-dropouts, only the patients staying in the study throughout are used. This ensures that a possible continuing slow change in the means of later dropouts does not mask the differences of mean values by moving the mean of the non-dropouts into the direction of the mean of the dropouts. The statistic D takes on large positive (negative) values when all means for the dropouts are smaller (greater) than the ones corresponding to the individuals remaining in the study. For use as a test statistic, this normally distributed variable must be standardised. The variance is given by

$$Var(D) = \frac{1}{w}\sigma^2 + \frac{\sigma^2}{w n_J^2}\sum_{k,j=1}^{J-1} w_k w_t \rho_{kj}\,.$$

The correlations ρ_{kj} are estimated from the data belonging to the non-dropouts only. The estimation of σ^2 can be based on the non-dropouts, since it is assumed that all y_i have the same distribution if the null hypothesis holds. The test statistic is calculated as

$$S = \frac{D}{\sqrt{\hat{V}ar(D)}}.$$

This test statistic is asymptotically normal (Listing, Schlittgen 1998). The calculation of this test statistic is not available in standard statistical software. A program was developed to calculate the test statistic for j = 3 to 6 assessments in the statistical package STATA (StataCorp 2007). An example program for j=4 assessments is provided in Appendix 4.2.

4.2.3 Listing and Schlittgen: A non-parametric test for random dropouts

Following on from the parametric test outlined in section 4.2.2, Listing and Schlittgen proposed a non-parametric test for random dropouts, which did not require the assumption of normally distributed data (Listing, Schlittgen 2003). In this procedure, a Wilcoxon rank sum test is calculated at each time point and the sum of the standardized rank statistics is used for the overall test. The only assumption is that the direction of the change in the values of the variable Y is the same. Since the single rank statistics are uncorrelated the sum is asymptotically normally distributed with a variance that can be easily obtained (Listing, Schlittgen 2003). This test requires a monotone missing data pattern. The aim is to test the hypothesis that dropouts do not show any tendency for larger (smaller) values than the people remaining in the study, at least for the next observation. The technical detail of this test can be found in the original publication (Listing, Schlittgen 2003).

4.2.4 Schmitz and Franz: A bootstrap method to test if study dropouts are missing randomly

In 2002, Schmitz and Franz proposed a nonparametric method, based on a bootstrap approach (Efron, Tibshirani 1993), for assessing whether dropouts are missed at random (Schmitz, Franz 2002). This paper built on the ideas from the Listing and Schlittgen parametric method (Listing, Schlittgen 1998), to compare scores of dropouts and non-dropouts at different assessments using a weighted

nonparametric statistic. The test proposed by Schmitz and Franz requires a monotone missing data pattern (Schmitz, Franz 2002).

The standard Wilcoxon rank-sum test can be computed at each time point comparing the dropouts with the non-dropouts. The Wilcoxon test requires all the observations to be ranked as if they were from a single sample. Then, the sum of the ranks in one group (e.g. dropouts) is calculated. A nonparametric test statistic for the problem is a weighted sum of the mean rank of scores of the dropouts at the different time points. Bootstrap techniques can be used to approximate the distribution of the test statistic under the null hypothesis (Efron, Tibshirani 1993). Further details of this method can be found in Schmitz and Franz (Schmitz, Franz 2002).

4.2.5 Ridout's logistic regression method: test for random dropouts

Ridout proposed a test for random dropout in repeated measures data that utilised logistic regression (Ridout 1991). This method is comparable to that of Diggle (Diggle 1989), but is a more flexible technique for studying the occurrence of dropouts as more complex models are allowed. Ridout's method assumes a monotone pattern of missing data and that the baseline assessment is available for all patients.

Ridout's logistic regression method is implemented as follows: at the first assessment those patients who provide a valid assessment are identified. The subset of patients for which this is their last assessment before they drop out the study is obtained. The test for complete random dropout involves testing the assumption that scores from the subset of dropouts are a random sample from the patients who provided assessment. This process is repeated for each of the assessments in turn, in order to investigate the mechanism of missingness for each assessment separately.

The response variable is dropout or not at time j in the standard logistic regression model (Hosmer, Lemeshow 1989). This logistic regression procedure can be

performed in standard statistical software. Covariates contained within the dataset along with QoL scores can be considered for inclusion in the logistic model to predict dropout. In the presence of MCAR, the dropout mechanism may depend on the values of fixed covariates – referred to as covariate-dependent drop out. If the dropout mechanism at assessment *j* depends on values of observed QoL, then there is evidence of MAR data.

4.2.6 Fairclough's method: logistic regression approach

Fairclough describes an approach using logistic regression to determine the missingness mechanism (Fairclough 2002). It is very similar to the logistic regression approach suggested by Ridout (Ridout 1991), but with one subtle difference. The response variable under Fairclough's approach is an indicator of missingness at time *j*. If the value of the outcome (QoL score) at a particular assessment is known, then the indicator equals zero; if the outcome is missing, then the indicator variable equals one. Ridout models the indicator of subsequent dropout or not for those who have been observed at the current time point (Ridout 1991).

The first step in Fairclough's logistic regression approach is to identify which covariates are associated with missingness. This is done using standard statistical tests such as t-tests or chi-squared tests. Once these have been identified, the baseline covariates are forced into a logistic model. The inclusion of the baseline or previous QoL can be tested using likelihood ratio tests (Hosmer, Lemeshow 1989). If no QoL scores are found to be significant in the model, there is a conclusion of covariate-dependent missingness. In the situation that observed QoL is significant in the model for missingness, then missingness can be said to be MAR.

4.3 Overview

A number of methods have been identified from the literature to investigate the mechanism of missingness. A monotone missingness pattern is required by Diggle (Diggle 1989), Listing and Schlittgen's parametric (Listing, Schlittgen 1998) and non-parametric (Listing, Schlittgen 2003) procedures and for the Schmitz and Franz (Schmitz, Franz 2002) method. In the situation of intermittent missingness, these procedures are not valid. To use them involves restricting the data to that which displays a monotone missing data pattern. Chapter five will show that none of the datasets used in this thesis display monotone missingness for all patients. There is always a proportion (even if small) of patients displaying an intermittent missing data pattern.

For this reason Little's test is the most appropriate hypothesis test to investigate the missing data mechanism in the example datasets. In addition, Ridout's logistic regression and Fairclough's logistic regression will be employed in chapter five. Despite the limitation of only being applicable to monotone missingness patterns, the parametric Listing and Schlittgen test (Listing, Schlittgen 1998) was chosen as a comparison to Little's hypothesis test of MCAR. This will, however, require a restriction of the dataset to those patients with a monotone missing data pattern. The implications of this will be discussed in the next chapter.

The next chapter applies the four methods indicated to the datasets, in order to investigate the missingness mechanism. There is an investigation into the mechanism of non-response and the mechanism behind reminder-response. For the latter, a restricted set of data is obtained which includes the responders only. The data obtained by reminder is then removed and the mechanism of missingness (reminder-response) is investigated.

Chapter 5 Investigating the mechanism of missing data

5.1 Introduction

Previous authors who have looked at the missing data mechanism have done so using simulated data or removed it in such a way that the mechanism is pre-determined (Musil et al. 2002; Myers 2000). This procedure is potentially misleading, as the performance of the various tests can often be anticipated through the known mechanism that was used to generate the samples. This chapter presents an alternative approach that utilises the data collected through reminders. Since the data is ultimately known, the mechanism behind reminder-response can be identified and used to inform the actual mechanism of missing data. A number of methods to investigate the mechanism of missing continuous outcome data were discussed in chapter four. Little's test (Little 1988), Listing and Schlittgen's (LS) parametric test (Listing, Schlittgen 1998), Ridout logistic regression (Ridout 1991) and Fairclough logistic regression (Fairclough 2002) were chosen to investigate the missingness mechanism for each of the seven trial datasets. In the sections that follow, for each trial dataset, the four methods will be applied.

Within each trial dataset, there were three types of responders at each assessment: immediate-responders (I), reminder-responders (R) and non-responders (N). Three data scenarios were under consideration with respect to the missing data. Scenario one contains only the immediate responders. The missing data comprises the reminder-responders and non-responders (I versus {R and N}). Scenario two uses all observed responses collected with or without reminders ({I and R} versus N). Scenario three restricts the dataset to the responders only and regards the reminder responses as missing. This enables an investigation into the mechanism behind reminder-response (I versus R).

For each of the three scenarios, within each of the seven trial datasets, the four identified methods for investigating the missingness mechanism were carried out.

In scenario three, there is the ability to investigate the mechanism behind reminder-response. Comparing between scenario one and two can assess the impact the reminder-data may have on the conclusion of the missing data mechanism.

The investigation for each trial begins with a description of the missing data pattern. Tables with mean (SD) scores for each QoL outcome are found in the Appendix. Graphical representations of these mean values are provided within this chapter. A solid line between two data points indicates that these observations were made with no missing observations in between. A dotted line between two observations indicates there were one or more missing observations between the two observed data points.

5.2 REFLUX

5.2.1 Pattern of missing data

Appendix 5.1 shows the mean QoL scores (EQ5D, SF12 and RQLS) for the 357 participants at each of the three assessments for each missing data pattern. The number of participants with each missing pattern is also shown in Appendix 5.1. This information is displayed graphically in Figure 5.1 (EQ5D), Figure 5.2 (SF12 PCS), Figure 5.3 (SF12 MCS) and Figure 5.4 (RQLS) to give an overall picture of the QoL of participants with each missing data pattern. Taking Figure 5.1 as an example, the circles represent the mean QoL score at each assessment within a pattern. The dark blue pattern represents the mean QoL for participants who provided an observation at each of the three assessments. The green pattern provides mean QoL for those participants who provided QoL scores at baseline and at 12 months but not at six months (represented by the dotted line between baseline and 12 months). The patterns which consist of a single point represent the mean scores for those participants where only one assessment of QoL was available (represented by the dot).

Generally those patients who dropped out after baseline assessment displayed the worst QoL at baseline. Those who provided later assessments indicated better QoL. The worst QoL displayed at three and 12 months occurred with those patients who did not provide a baseline assessment. Those patients who missed the baseline assessment showed better mental summary scores and reflux-specific QoL compared to those patients who provided all three assessments. However, these same patients displayed lower EQ5D and physical functioning scores. Patients providing only one assessment tended to show lower QoL, particularly at baseline. Overall QoL appeared to differ between those patients who did and did not provide missing data.

Figure 5.1: REFLUX - EQ5D mean score at each assessment by missing data pattern

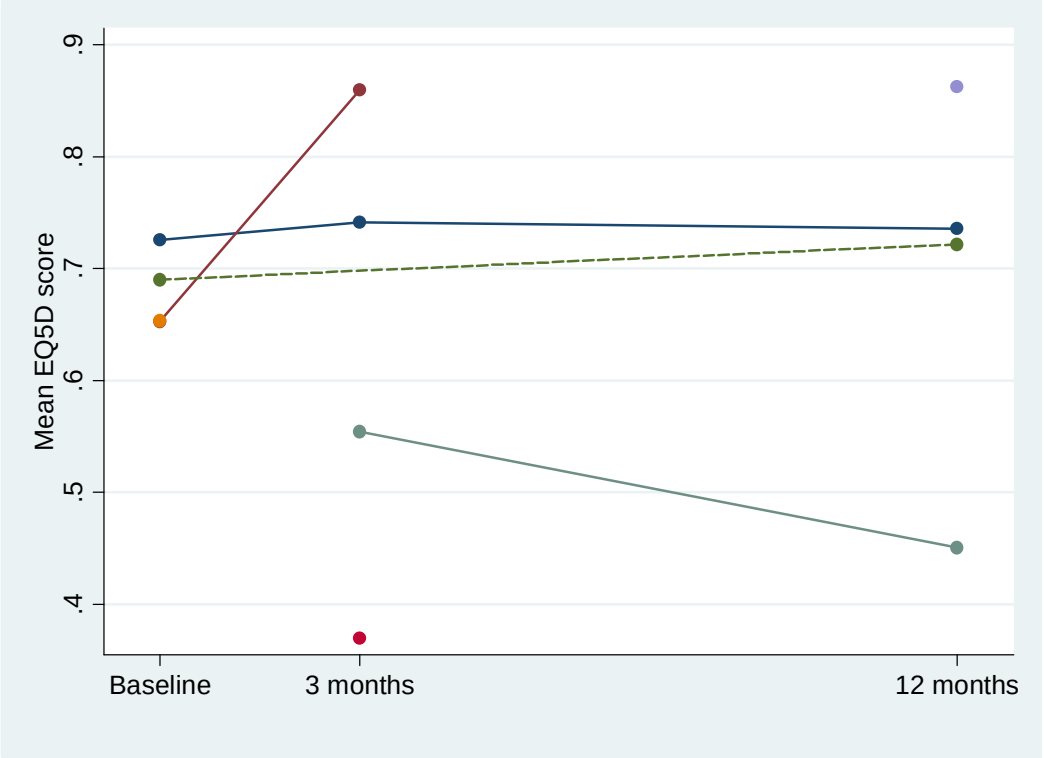


Figure 5.2: REFLUX - SF12 physical component mean score at each assessment by missing data pattern

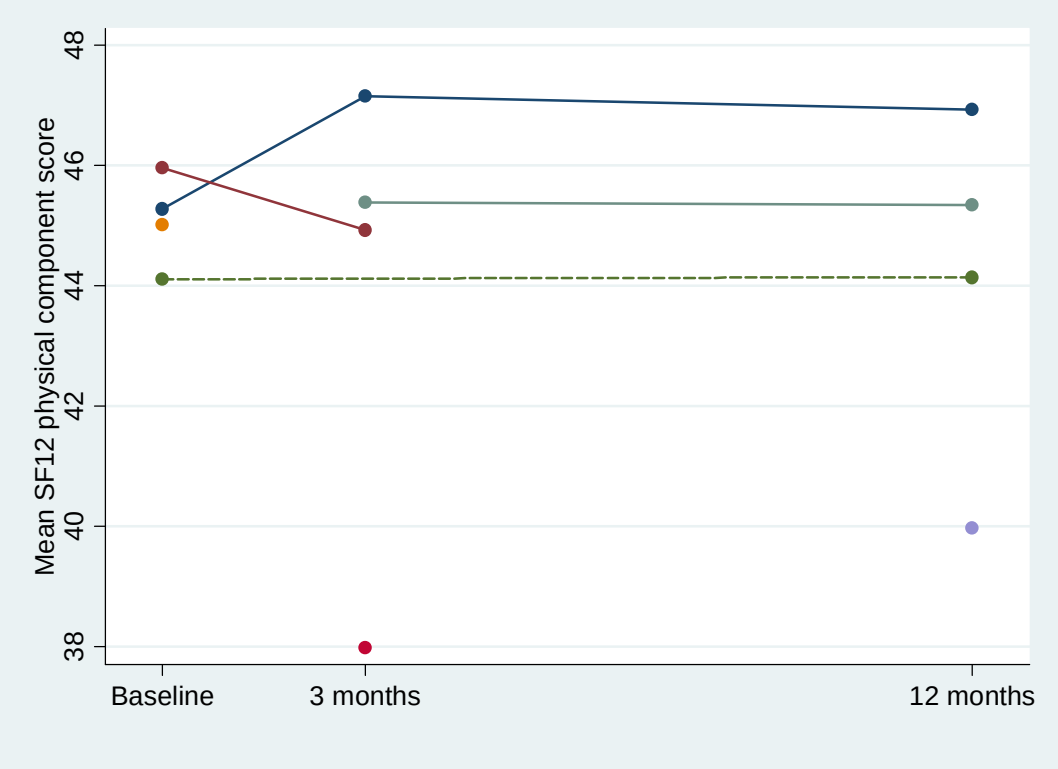


Figure 5.3: REFLUX - SF12 mental component mean score at each assessment by missing data pattern

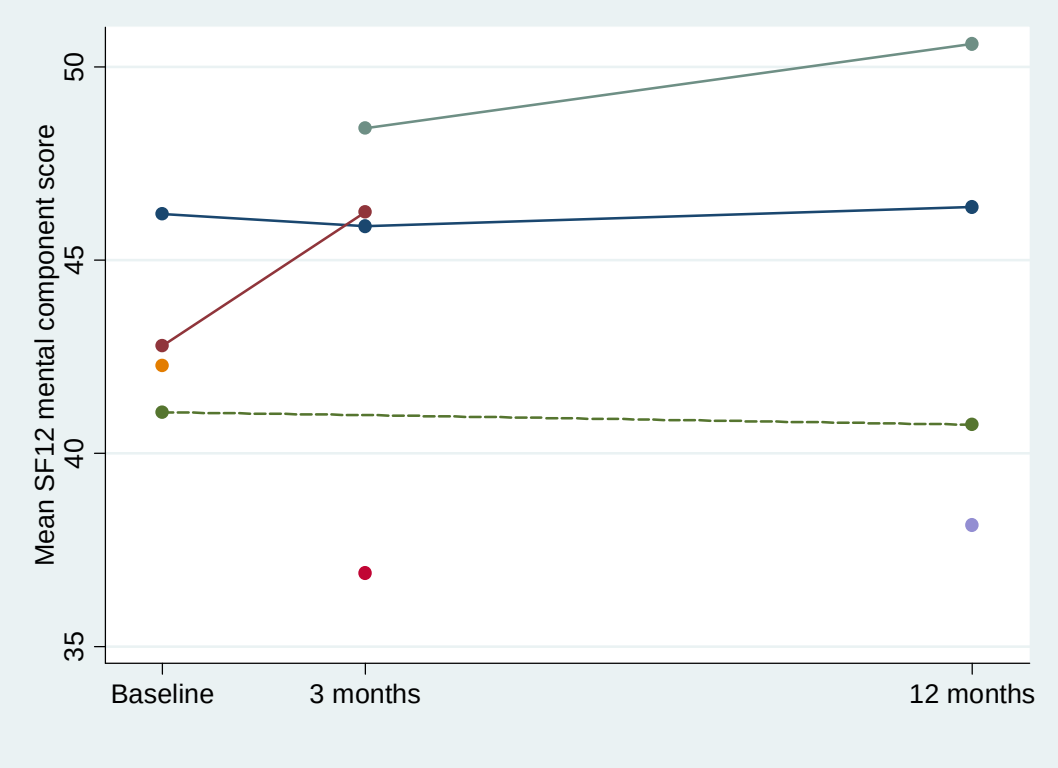
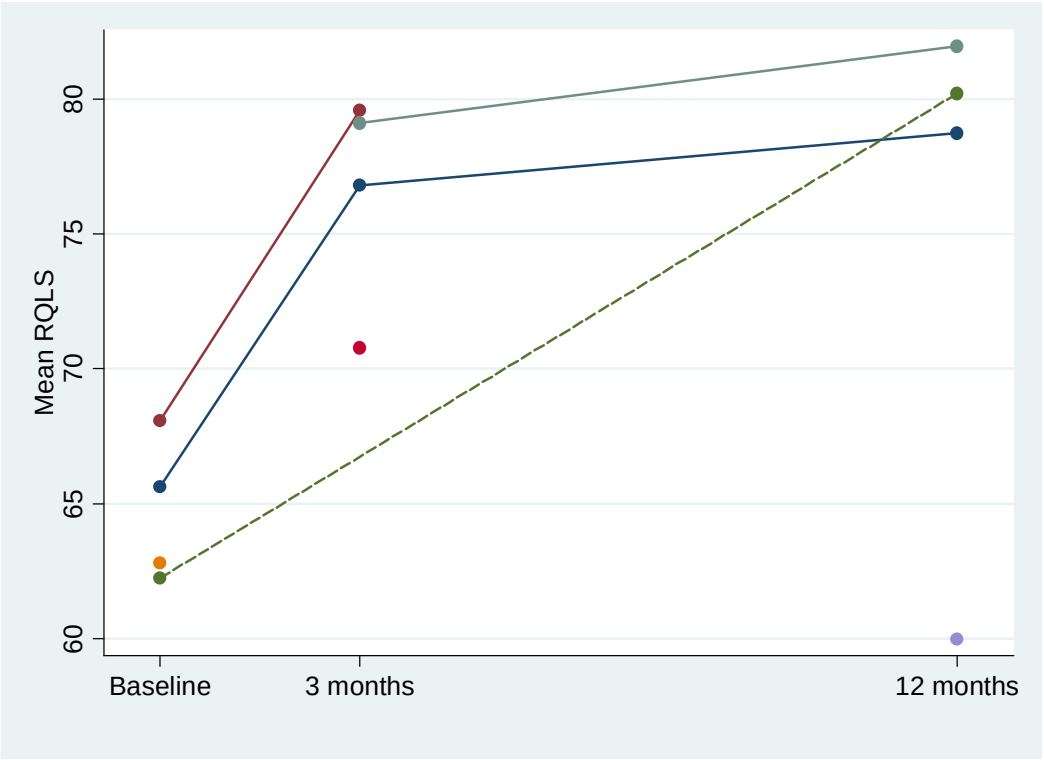


Figure 5.4: REFLUX - RQLS mean at each assessment by missing data pattern



5.2.2 Little's test

Table 5.1 shows the p-values from Little's test. There was evidence against the null hypothesis of missing completely at random (MCAR) EQ5D scores for each of the data scenarios. Therefore, the EQ5D scores cannot be regarded as MCAR, furthermore, QoL was impacting on whether or not questionnaires were returned and whether they were returned with or without the need for a reminder. For both the SF12 physical component score and the RQLS, Little's test was not significant for any of the three scenarios ($p>0.05$). Therefore, there was no evidence that the physical functioning or reflux specific QoL experience differed between those responding and not responding to the questionnaires.

Table 5.1: REFLUX - Little's test for MCAR p-values

Scenario	EQ5D	PCS	MCS	RQLS
One	0.01	0.90	0.005	0.33
Two	0.01	0.57	0.13	0.70
Three	0.02	0.91	0.01	0.43

In the case of the mental component score, there was evidence against the MCAR assumption for scenario one. Scenario two included the reminder responses and

consequently, this changed the conclusion; there was no evidence against the MCAR assumption (p=0.13). There was also evidence that the mental score differed between immediate and reminder responders (p=0.01), with the reminder responders showing poorer mental functioning.

5.2.3 Ridout Logistic regression

Scenario one

Table 5.2 shows the results of Ridout logistic regression in data scenario one. Baseline EQ5D scores were found to be a predictor of subsequent dropout having adjusted for age group (p=0.003). The odds ratio (OR) and 95% confidence interval (CI) for dropout was 0.27 (0.11, 0.65), implying that those with higher QoL were less likely to drop out, providing evidence of MAR. At three months, there was covariate dependent dropout (employment status). There was no significant difference in the three month EQ5D scores between those continuing and those dropping out (p=0.29).

Table 5.2: REFLUX - Ridout logistic regression results (scenario one)

	Total	Dropouts	Unadjusted		Adjusted	
	N (%)	N (%)	OR (95% CI)	p-value	OR (95% CI)	p-value
EQ5D						
Baseline	344 (96)	150 (44)	0.26 (0.11,0.62)	0.003	0.27 (0.11,0.65)[1]	0.003
3 months	139 (39)	63 (45)	0.52 (0.15,0.76)	0.29	-	-
SF12 physical component score (PCS)						
Baseline	336 (94)	153 (46)	0.99 (0.96,1.01)	0.25	-	-
3 months	133 (37)	58 (44)	1.00 (0.97,1.04)	0.89	-	-
SF12 mental component score (MCS)						
Baseline	336 (94)	153 (46)	0.97 (0.95,0.99)	0.002	0.97 (0.95,0.99)[2]	0.006
3 months	133 (37)	58 (44)	0.97 (0.95,1.00)	0.093	-	-
RQLS						
Baseline	327 (92)	154 (47)	0.99 (0.98,0.99)	0.019	0.99 (0.98,1.00)[2]	0.071
3 months	131 (37)	58 (44)	0.99 (0.97,1.01)	0.21	-	-

Adjusted for: [1] age group; [2] age group and BMI group;

No significant difference was found in the SF12 PCS between those continuing and those dropping out after baseline (p=0.25) or after three months (p=0.89). Dropout was found to be covariate-dependent at baseline (on age group and

whether or not the patient had erosive oesophagitis) and at three months (employment status) for the PCS.

The MCS did differ between those continuing and those dropping out after baseline (p=0.002). Adjusting for age group and body mass index (BMI) group, this difference was still evident (p=0.006) with adjusted OR (95% CI) equal to 0.97 (0.95, 0.99). Therefore, those with higher MCS were less likely to drop out and dropout was potentially MAR. At three months, baseline MCS were not predictive of dropout (p=0.093) and dropout was found to be covariate dependent (employment status).

At baseline there was a significant association between the RQLS and dropout (0.019). Adjusting for BMI group and age group, this difference was still significant at the 5% level. The adjusted OR = 0.99 provided borderline evidence that an increase in baseline QoL reduced the odds of subsequent drop out. There was no evidence that three month RQLS differed between the continuers and those dropping out (p=0.21).

Scenario two

Baseline EQ5D scores were provided by 344 participants in the REFLUX trial, with 23 (6.4%) subsequently dropping out, providing no further assessments. There was no significant difference (p=0.35) in the mean (SD) of the continuing participants (0.72 (0.25)) compared to those who dropped out after baseline (0.65 (0.33)). Comparing baseline information between the two groups showed a significant difference in source of recruitment (p=0.011). The proportion of continuers who were recruited retrospectively (47%) was significantly less than those of the drop outs (74%). Covariate dependent missingness was identified at baseline. At three months an EQ5D score was provided by 302 participants with 16 (5%) subsequently dropping out. The mean (SD) of the continuing group was 0.74 (0.27) compared with 0.74 (0.32) for the dropout group (p=0.99). Employment status was found to significantly differ between the groups (p=0.036) with a higher proportion of continuers in full time employment (64%) compared with the drop

outs (50%). Hence, there was evidence of covariate dependent drop out at three months.

Baseline SF12 component scores were available for 336 participants of which 22 (6.5%) dropped out without any further assessments. No covariates were found to be associated with dropout after baseline. The PCS was not significantly different (p=0.93) with the continuers having a mean (SD) of 45.2 (9.5) compared with 45.0 (9.1) for those dropping out. This suggested no evidence against the MCAR assumption for the missing data at baseline. At 3 months, there were 292 respondents and for 17 (6%) patients, it was their final assessment. The mean PCS of the continuing group was 47.1 (9.8) compared to 44.1 (11.1) for the dropouts. This difference was not statistically significantly (p=0.23). Similarly, the mean MCS did not differ with mean (SD) of continuers 46.0 (12.7), compared with 45.2 (14.8) in the dropout group (p=0.80). No covariates were found to be associated with dropout, providing no evidence against the MCAR assumption.

An RQLS was obtained for 327 respondents at baseline. There were 302 (92%) who continued in the study to provide a further assessment. The mean (SD) RQLS of the continuers was 65.4 (24.4) compared with 62.8 (24.3) for the drop outs. This difference was not statistically significant (p=0.61). There was evidence of covariate dependent missingness for the RQLS (source of recruitment, p=0.001), with 87% of the retrospective group continuing compared with 97% of those recruited prospectively. This provided evidence of covariate-dependent missingness at baseline for the RQLS. At three months, there were 286 subjects with a RQLS, of which 257 (90%) continued onto the final assessment. The continuers displayed lower RQLS (77.0 (23.6)) compared to the dropouts (78.7 (18.6)), but this difference was not significant (p=0.71). No covariates were found to differ providing no evidence against the assumption of MCAR data.

Scenario three

The dataset was restricted to those providing QoL scores at all three assessments and the reminder-responses were deleted. A continuer was defined as a patient who continued to respond immediately, while a drop-out was defined as someone

responding after reminder at one or more future assessments. There were 281 participants who provided all three EQ5D assessments. After baseline, 181 (64%) of these continued to respond immediately (continuer), while 100 (36%) responded by reminder (dropout). There was a significant difference (p=0.009) in their baseline QoL scores, with the continuers on average 0.09 (0.02, 0.15) units higher than the dropouts. In a logistic model adjusting for the covariates age group and BMI group, baseline EQ5D was still a significant predictor (p=0.013). At three months, there was no significant difference in the EQ5D scores of the continuers and the dropouts (p=0.38). In conclusion, reminder response was found to be MAR after baseline, but MCAR after three months.

There were 268 available SF12 scores at baseline with 172 (64%) patients continuing and 96 (36%) dropping out. There was no significant difference in the baseline PCS (p=0.27). BMI group and age group were significant predictors of dropout with those in the younger age group (18-50 years) and those in the lower BMI group (≤28) more likely to continue in the study and respond immediately. At three months, there was no significant difference (p=0.68) in PCS between those continuing on to respond immediately at the final assessment and those responding after reminder. No covariates were found to be associated with drop out.

The mean MCS, however, did show a significant difference at baseline (p=0.03). This difference was 3.10 (0.31, 5.89) with those continuing in the study displaying the better (higher) MCS. After adjusting for age group and BMI group, the baseline score was still significant in the model (p=0.046). At three months, there was no significant difference in the covariates or MCS between continuers and dropouts (p>0.05).

A similar phenomenon was found with the RQLS. A significant difference in the baseline RQLS was still found to be important having adjusted for BMI group and age group (p=0.048). The 150 (63%) responders at baseline who continued in the study had mean (SD) RQLS of 68.3 (23.5), compared to 61.1 (25.3) for the 89 (37%)

patients who dropped out after baseline. At three months, there was no significant difference in the mean MCS or covariates (p>0.05).

5.2.4 Listing and Schlittgen's test (LS test)

Scenario one

The LS test requires a monotone missing data pattern and 287 (80%) patients showed this for the EQ5D outcome. The LS test statistic was S=2.24 (p=0.033) providing evidence in favour of MAR. There were 281 patients displaying a monotone missingness pattern for the SF12 component scores. The LS test statistic was S=0.23 (p=0.39) for the PCS and S=3.12 (p=0.003) for the MCS. Hence, there was evidence in favour of MAR for the MCS but not for the PCS. There was no evidence in favour of MAR for the RQLS with S =1.38 (p=0.15).

Scenario two

There were 316 (89%) patients showing a monotone pattern of missingness for the EQ5D score (Appendix 5.1). The LS test statistic was S = 0.16 (p=0.39), providing insufficient evidence that the data was missing at random. For the SF12 component scores, 305 patients showed a monotone missingness pattern and the LS test statistic for the physical score was S=0.66 (p=0.32), providing no evidence in favour of MAR. Appendix 5.1 shows that at baseline the MCS was lowest in those patients who subsequently dropped out. However, the LS test statistic was S=1.10 (p=0.22) indicating insufficient evidence that data were missing at random. Finally, there were 290 patients with monotone missingness in the RQLS. There was no evidence of MAR data for the RQLS with S=-0.008 (p=0.40).

Scenario three

There was evidence in favour of MAR for the EQ5D score (S=2.25, p=0.031) and the SF12 MCS (S = 2.97, p=0.005), but not for the PCS (S=0.215, p=0.39) or the RQLS (S=1.59, p=0.11). Therefore, observed EQ5D QoL and mental component scores were impacting on whether or not a patient responded after reminder or not.

5.2.5 Fairclough logistic regression

Scenario one

The first stage was to identify any covariates that were associated with missingness. Age group was found to be associated with missingness at both three (p=0.025) and twelve months (p=0.004). The proportion of immediate responders was significantly greater at three months for the 51-65 year olds (46%) compared with the 18-50 year olds (35%). Similarly at 12 months, the older age group contained a higher proportion of immediate responders (47%) compared with the younger age group (32%). The previous EQ5D scores differed between responder groups at both three (p=0.043) and twelve months (p=0.003). Adjusted for age group, this difference was still evident at 12 months with adjusted OR = 0.28 (0.11, 0.69), p = 0.006. At three months, the difference was borderline significant (p=0.054) having adjusted for age group, with adjusted OR = 0.41 (0.16, 1.02). Therefore, Fairclough logistic regression suggested possible MAR data at three months and evidence of MAR at 12 months for the missing EQ5D scores.

The SF12 PCS was not significantly different between responder groups at either three months (p=0.52) or at 12 months (p=0.19), suggesting missingness was covariate dependent (on age group). The MCS was found to be MAR; the adjusted OR (for age group) at three months was 0.98 (0.96, 1.00), p=0.002 and at 12 months 0.97 (0.95, 0.99), p=0.006.

There was no significant difference in RQLS between the responder groups at three months (p=0.16). At twelve months the adjusted (for age group) OR was 0.98 (0.97, 0.99), p = 0.002 for the previous QoL term. This suggested that missingness was MAR at twelve months and that those displaying better previous RQLS were 2% less likely to provide a missing assessment at 12 months.

Scenario two

At three months, there was a significant difference in the smoking status of responders and non-responders (p=0.012). Questionnaires were returned by 88% of the non-smokers compared with 81% of the smokers. The baseline QoL (EQ5D, SF12 physical and mental component scores) did not differ between the responder

groups. This suggested at three months, missingness was covariate dependent (on smoking status).

At 12 months, no covariates were found to be predictors of non-response. Investigating the differences in previous QoL showed that there was a difference in the baseline SF12 MCS (p=0.03) between the responders (mean (SD) = 45.7 (11.7)) and non-responders (mean (SD) = 41.4 (12.7)) at 12 months. Therefore at 12 months, observed QoL was found to be a significant predictor of missingness and there was evidence of MAR data.

Scenario three

At three months there was a significant difference in age group (p=0.03) between the immediate-responders and reminder-responders. Comparing the QoL experience, showed the baseline (p=0.013) and 3-month (p=0.012) MCS differed between the two types of responder. Adjusting for age group and baseline MCS in a logistic model showed the three month score was still a potentially important predictor (p=0.048), with those displaying lower mental scores more likely to respond by reminder.

At 12 months, the significant covariate predictors of reminder-response were age group (p=0.009) and BMI group (p=0.022). For the EQ5D score, baseline and 12-month QoL assessments were significant predictors of missingness. The 12-month (current QoL) score was not included in the model after adjusting for the covariates and baseline EQ5D score (p=0.22). No difference was found in the SF12 PCS at any of the three assessments. The MCS did differ at baseline, 3-months and 12 months between the immediate-responders and reminder-responders at 12 months. In the logistic model, baseline and current MCS were important (p=0.012) suggesting possible MNAR data. RQLS at baseline differed between immediate and reminder responders at 12 months (p=0.011), but the 12 month score did not (p=0.07). The baseline score (p=0.035) was important in the logistic model but the 12 month score was not (p=0.40). This suggested the data were MAR.

5.2.6 Summary

Scenario one

Little's test found no evidence against the MCAR assumption for the PCS and RQLS. There was evidence against MCAR for the EQ5D score and the MCS. Despite being based on only those with a monotone pattern, the LS test gave a similar conclusion to Little's test. There was evidence in favour of MAR for the EQ5D and MCS, but not for the PCS and RQLS. Ridout logistic regression found evidence of MAR for the EQ5D and MCS at baseline. For the remaining situations there was evidence of covariate (age group) dependent missingness. Fairclough logistic regression showed missingness at three and twelve months to be MAR for the EQ5D and MCS. The PCS was found to be MCAR (covariate-dependent). Finally, at three months RQLS was MCAR, but at 12 months there was evidence of MAR.

Scenario two

Little's test found that generic QoL measured by the EQ5D appeared to affect whether or not the patient responded, but disease specific QoL (RQLS) did not. Ridout's logistic regression found evidence of covariate dependent dropout for the EQ5D score, and insufficient evidence against the MCAR assumption for the two SF12 component scores and the RQLS. The LS test found insufficient evidence of MAR data within the four QoL measures. Fairclough's logistic regression procedure came to the conclusion of covariate dependent missingness between responders and non-responders at each assessment.

Scenario three

This data scenario investigated the mechanism behind reminder response compared to immediate response. Little's test showed evidence against the MCAR assumption for the EQ5D score and SF12 MCS, implying QoL as measured by these two constructs was a factor in reminder response. In Ridout's regression, covariates and baseline QoL were predictors of subsequent dropout (reminder response). Those who were younger, or had poorer QoL or lower BMI were more likely to be reminder responders. At three months, no difference in QoL or covariates was found. Fairclough's logistic regression showed that previous QoL

affected response in all cases and current mental QoL affected the time of response, with those displaying better QoL less likely to need a reminder to respond. Finally, the LS test found evidence in favour of MAR for the EQ5D and SF12 MCS, but not the PCS or RQLS score.

5.3 MAVIS

5.3.1 Pattern of missing data

Appendix 5.2 shows the number of participants and mean (SD) QoL scores by missing data pattern in MAVIS. The majority (90%) of participants provided all three assessments. Figure 5.5 displays the information graphically for the EQ5D scores, Figure 5.6 for the physical component scores of the SF12 instrument and Figure 5.7 for the mental component scores.

Figure 5.5: MAVIS – EQ5D mean score at each assessment by missing data pattern

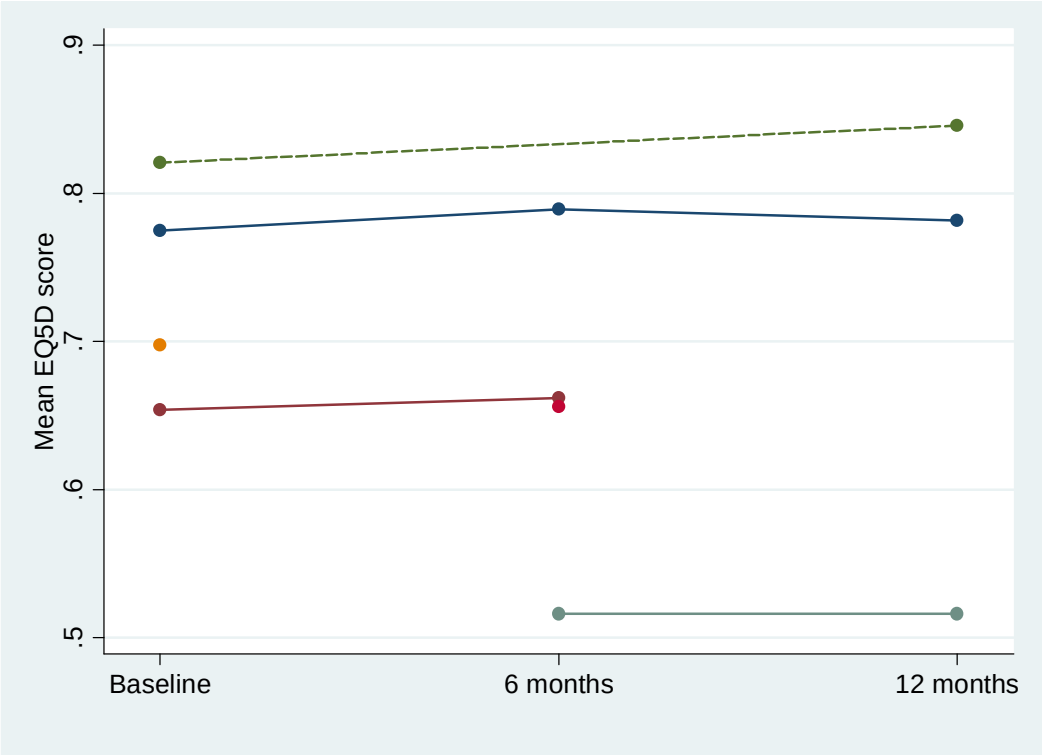


Figure 5.6: MAVIS – SF12 physical component mean score at each assessment by missing data pattern

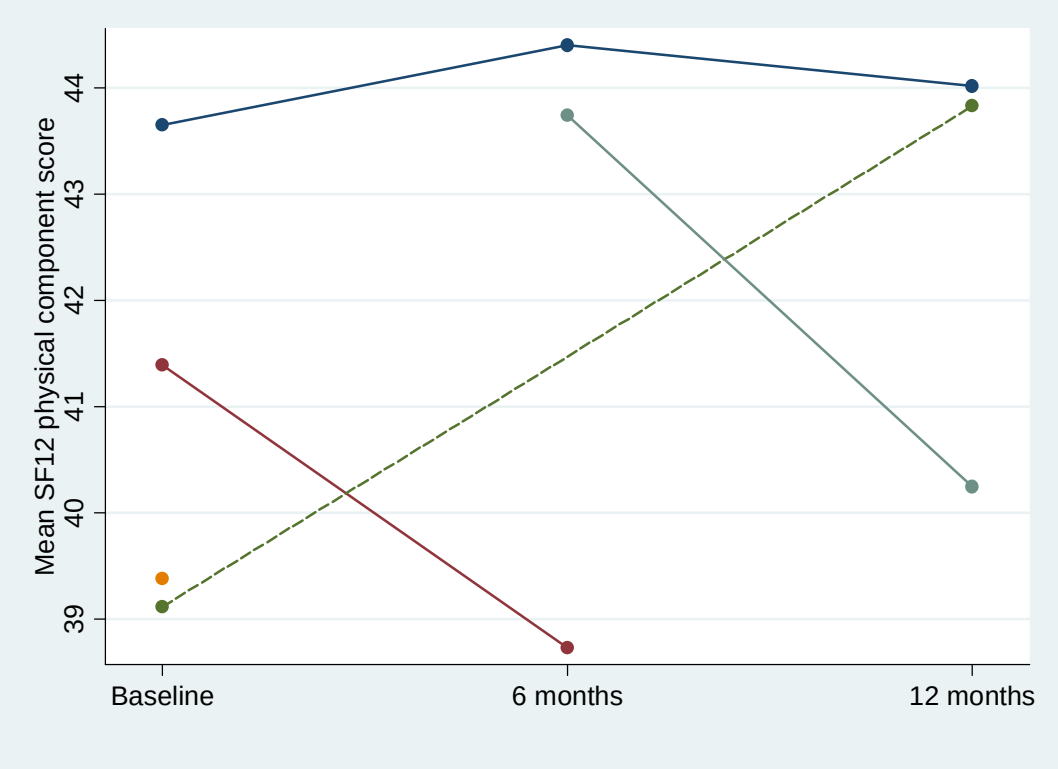
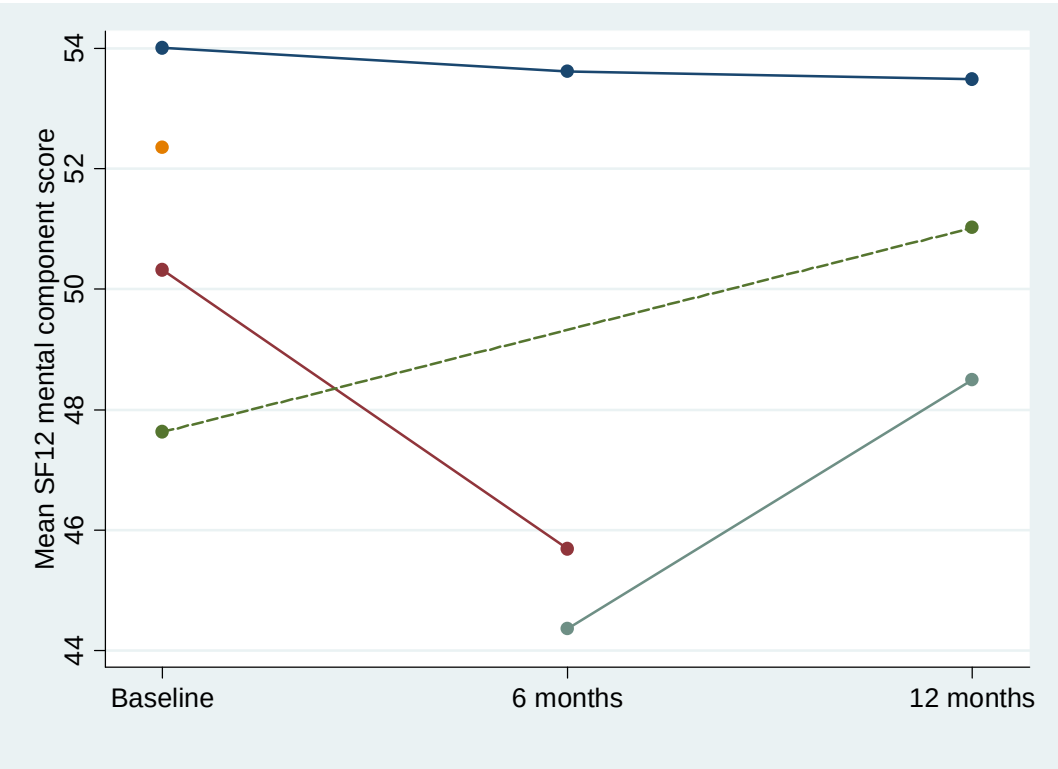


Figure 5.7: MAVIS – SF12 mental component mean score at each assessment by missing data pattern



It is easily seen from these that the mean QoL scores vary among the different patterns of missing data, suggesting that perhaps QoL has an impact on whether a patient responds to the questionnaire. Those patients without a baseline score display the worst QoL at the follow up assessments. Generally, the best QoL is observed in those who provide all three assessments. Averaging across all patterns, the mean QoL is reasonably stable from baseline to 12 months with the EQ5D score differing by 0.01 units and the SF12 scores by 0.7 units (PCS) and 0.3 units (MCS).

5.3.2 Little's test

Table 5.3 shows the p-values that resulted from Little's test for MCAR for each of the QoL scores and data scenarios. With respect to the EQ5D scores, there was evidence against the MCAR assumption in scenario two, but no evidence against MCAR for scenario one or three. For both the physical and mental component scores a significant result was found for either scenario one and two, providing evidence against the MCAR assumption in favour of either MAR or MNAR. The QoL experience between immediate and reminder responders (scenario three), was not found to be significantly different for any of the three QoL scores. This suggested that reminder response was perhaps undertaken completely at random.

Table 5.3: MAVIS - Little's test for MCAR p-values

Scenario	EQ5D	PCS	MCS
One	0.20	0.015	0.031
Two	0.015	0.002	<0.001
Three	0.91	0.37	0.17

5.3.3 Ridout Logistic regression

Scenario one

Table 5.4 displays the results of the Ridout logistic regression process for scenario one. Baseline EQ5D scores were found to be associated with dropout (p=0.012). This difference was less significant (p=0.07) once the covariates age group and sex had been adjusted for. Although not statistically significant at the 5% level, there was some evidence that those patients showing better QoL at baseline were less

likely to dropout afterwards. The six month EQ5D score was not found to be associated with dropout (p=0.18) and there was no evidence in favour of MAR.

Table 5.4: MAVIS - Ridout logistic regression results (scenario one)

	Total N (%)	Dropouts N (%)	Unadjusted OR (95% CI)	p-value	Adjusted OR(95% CI)	p-value
EQ5D						
Baseline	908 (99)	60 (7)	0.27 (0.10,0.75)	0.012	0.37 (0.13,1.09)[1]	0.07
6 months	823 (90)	120 (15)	0.56 (0.24,1.30)	0.18	-	-
SF12 physical component score (PCS)						
Baseline	906 (99)	61 (7)	0.97 (0.95,0.99)	0.007	0.98 (0.96,1.00)[1]	0.064
6 months	818 (90)	122 (15)	0.98 (0.97,1.00)	0.053	0.99 (0.97,1.01)[2]	0.29
SF12 mental component score (MCS)						
Baseline	906 (99)	61 (7)	0.98 (0.95,1.01)	0.13	-	-
6 months	818 (90)	122 (15)	0.98 (0.96,0.99)	0.029	0.98 (0.96,1.00)[2]	0.088

Adjusted for: [1] age group and sex; [2] residence type, heart problems and chronic infection.

As with the EQ5D scores, the physical component score was associated with dropout after baseline (p=0.007), but having adjusted for sex and age group, this was no longer the case (p=0.064) at the 5% level. At six months, the odds of dropout for the PCS was borderline significant. Once the covariates had been adjusted for, there was no evidence of MAR suggesting that missing PCS were covariate-dependent. Baseline MCS were not found to differ between those continuing and those dropping out (p=0.13). The covariates associated with dropout after baseline assessments were residence type, sex and age group. At six months, the unadjusted odds ratio for dropout was significant (p=0.029). Adjusting for residence type, heart problems and presence of chronic infection, the inclusion of the six month MCS term in the model was not significant (p=0.088). Hence, within the MAVIS trial, according to Ridout's logistic regression procedure, missing QoL scores in scenario one were found to be covariate dependent.

Scenario two

There were 908 participants who provided a baseline EQ5D score and for 6%, it was their final assessment. Tests of association between dropout after baseline and the covariates showed a significant relationship with sex (p = 0.022) and age group (p=0.004). In a logistic regression model with these covariates, the baseline

EQ5D score was not an important addition to the model (p=0.13), suggesting covariate-dependent drop out.

At six months, the logistic regression model included the covariates sex, age group and residence type. The addition of the six month score to the model was significant (p=0.021). Therefore, after adjusting for covariates, the six month QoL score was an important factor in determining whether the patient was likely to dropout after this point. The odds ratio for dropout of the six month EQ5D score was 0.16, 95% CI (0.04, 0.68) suggesting a person in better health was less likely to dropout.

Adjusting for sex and age group there was no difference in the PCS at baseline between those who continued and those who dropped out (p=0.08). Hence, there was no evidence against the MCAR assumption. In the case of the MCS, there was no difference in the baseline score between continuers and those who dropped out (p=0.25).

At six months, there was a significant difference in both the PCS (p=0.003) and MCS (p<0.001) of those who continued and those who dropped out, with drop outs displaying lower QoL. Covariates found to be associated with dropout were age group, diabetic status, residence type and number of current medications. Adjusting for these covariates in a logistic model, both the PCS (p=0.01) and MCS (p<0.001) were significant predictors of dropout. This suggested that QoL experience was affecting whether or not a patient dropped out. Those showing higher QoL at six months were less likely to drop out.

Scenario three

At baseline, 825 patients provided an EQ5D score, of which 7 went on to be reminder-responders at both the follow-up assessments. Given the low number in the dropout group, the results should be interpreted with caution. The baseline EQ5D was not found to differ between the two groups (p=0.77). Of those responding at baseline 123 (15%) were reminder-responders at one of the two follow up assessments. Their baseline scores did not differ to that of the

immediate responders (p=0.71). At six months follow up, the EQ5D scores did not significantly differ (p=0.99) between those who were immediate responders at the final follow up and those who dropped out (reminder-response). None of the covariates were found to be predictors of reminder-response. Following the same process for the SF12 data showed that physical and mental component scores were not significantly different between those who continued to respond immediately and those who subsequently responded by reminder.

5.3.4 The LS test

Scenario one

There were 881 (97%) patients who displayed a monotone pattern of missingness for the EQ5D score. The LS test statistic was S=2.79 (p=0.008). For the physical and mental component scores (N=875), the LS test was S=3.41 (p=0.001) and S=3.08 (p=0.003) respectively. In all three cases there was evidence in favour of MAR, suggesting that QoL experience was affecting whether or not a patient responded immediately or not.

Scenario two

There were 904 (99%) patients who showed a monotone pattern of missingness for the EQ5D score. The LS test statistic was S = 4.02 (p<0.001), providing evidence in favour of MAR data. Those patients who completed all assessments displayed higher (better) EQ5D scores. There were 897 (98.5%) patients who showed a monotone pattern for the SF12 component scores. The LS test statistic was significant for both the PCS (S=4.10, p<0.001) and MCS (S=4.27, p<0.001), providing evidence in favour of MAR data in both cases.

Scenario three

The calculation of the LS test statistic showed no significant results for the three QoL scores: EQ5D - S=0.002 (p=0.40); SF12 PCS – S=0.729 (p=0.31); SF12 MCS - S=0.330 (p=0.38). Hence, there was no evidence in favour of MAR for any of the three QoL scores.

5.3.5 Fairclough logistic regression

Scenario one

Age group was found to be a predictor of missing EQ5D scores at six months, with the older age group less likely to respond immediately. At twelve months current medication, residence type and age group were predictors of missingness. Those living in the community were more likely to be immediate responders (81%) compared to those in institutional care (55%). Those patients receiving fewer medications were more likely to be immediate responders. The adjusted OR for missingness at six months was 0.35 (0.14, 0.87), p=0.024, and at twelve months, 0.56 (0.26, 1.21), p=0.14 for previous EQ5D scores. Therefore, at six months there was evidence of MAR and at 12 months no evidence against MCAR (covariate-dependent) for the missing EQ5D scores.

Males (92%) were more likely to provide SF12 scores by immediate response compared to females (87%) at six months (p=0.012). As with the EQ5D scores at 12 months, current medication, residence type and age group were associated with missingness. The adjusted OR for missingness at six months for the previous PCS was 0.97 (0.95, 0.99), p=0.002 and at 12 months was 0.98 (0.97, 1.00), p=0.05. For the MCS at six months the adjusted OR = 0.97 (0.95, 0.99), p=0.04 and at twelve months OR=0.99 (0.97, 1.01), p=0.23. Hence, there was evidence against MCAR in favour of MAR at six months, but insufficient evidence against MCAR at 12 months.

Scenario two

At six months, 862 (95%) of questionnaires were returned, while at 12 months 840 (92%) were returned. At six months, only age group (p=0.001) was significantly associated with missingness, with those in the youngest age group (65-74 years) more likely to return the questionnaire. In addition to age group (p<0.001) at 12 months, residence type (p<0.001) was also a significant predictor of missingness. Baseline EQ5D (p=0.04) and SF12 PCS (p=0.04) were significantly associated with missingness. At 12 months, all three baseline QoL scores were significant predictors of missingness (EQ5D: p<0.001, SF12 PCS: p=0.004 and SF12 MCS: p=0.017).

A logistic model for the missingness indicator at six months found the inclusion of the baseline QoL score not to be significant (p>0.05). At 12-months, the inclusion of the baseline EQ5D score was significant (p=0.003), but the SF12 scores were not. A participant was more likely to provide missing response at 12 months if they were older, resident in care or had a lower baseline EQ5D score.

Scenario three

There were 853 (94%) patients with both baseline and six month scores EQ5D scores, of which 32 (4%) responded after reminder. None of the covariates were found to be significant predictors of reminder response (p>0.05). The baseline EQ5D was not found to differ between the immediate and reminder responders (p=0.85). The same results was found for the six month EQ5D score (p=0.97). At 12 months none of the covariates were predictors of reminder response. The QoL scores did not significantly differ between the types of responder (immediate or reminder). Similar results were found with the two SF12 component scores. There was no difference in the baseline characteristics or baseline QoL (physical and mental) of those who responded immediately or after reminder at six or twelve months. This investigation showed there was no evidence against the MCAR assumption for response after reminder.

5.3.6 Summary

Scenario one

Little's test of MCAR found no evidence against the MCAR assumption for missing EQ5D scores, but did find that missing SF12 component scores were not MCAR. The LS test found in favour of MAR for each of the three QoL scores. Ridout logistic regression found that despite some differences in QoL, after adjusting for the covariates, QoL was no longer a predictor of dropout and dropout was covariate-dependent. Finally, the Fairclough logistic regression process suggested QoL scores were MAR at six months, but MCAR at 12 months.

Scenario two

In summary, Little's test provided evidence against the MCAR assumption in the comparison of responders versus non-responders. In Ridout's regression there

was covariate dependent missingness after baseline for each QoL score. At six months, patients displaying lower EQ5D or SF12 component scores were more likely to drop out. There was evidence against the null hypothesis for the LS test, providing evidence in favour of MAR for each of the three QoL scores. Fairclough's logistic regression showed at six months that missingness was covariate-dependent. At 12 months, in addition to covariates, baseline EQ5D was a significant predictor of missingness indicating MAR. However, the SF12 component scores were not suggesting MCAR for this QoL measure.

Scenario three

None of the four tests of missingness found QoL to be a predictor of reminder response. Little's test found no evidence against the MCAR assumption, while the LS test found no evidence of MAR data. Ridout's method did not find any covariates or QoL scores to be predictors of subsequent reminder response. Fairclough's logistic regression found that reminder response seemed to be made completely at random.

5.4 RECORD

5.4.1 Pattern of missing data

Appendix 5.3 shows the number of participants and mean (SD) QoL scores for each missing data pattern in RECORD. The data are represented graphically in Figure 5.8 for the EQ5D score, Figure 5.9 for the physical summary score and Figure 5.10 for the mental component score. It is clear from this that those patients who provided all three assessments showed the best QoL at each assessment. Baseline QoL (four month) was lowest for those patients who did not provide any further assessments. Those patients who missed the second assessment (12 months) showed poorer QoL than those who provided all three. Therefore, it appears the QoL did differ between those who did and did not have missing data.

Figure 5.8: RECORD – EQ5D mean scores by missing data pattern

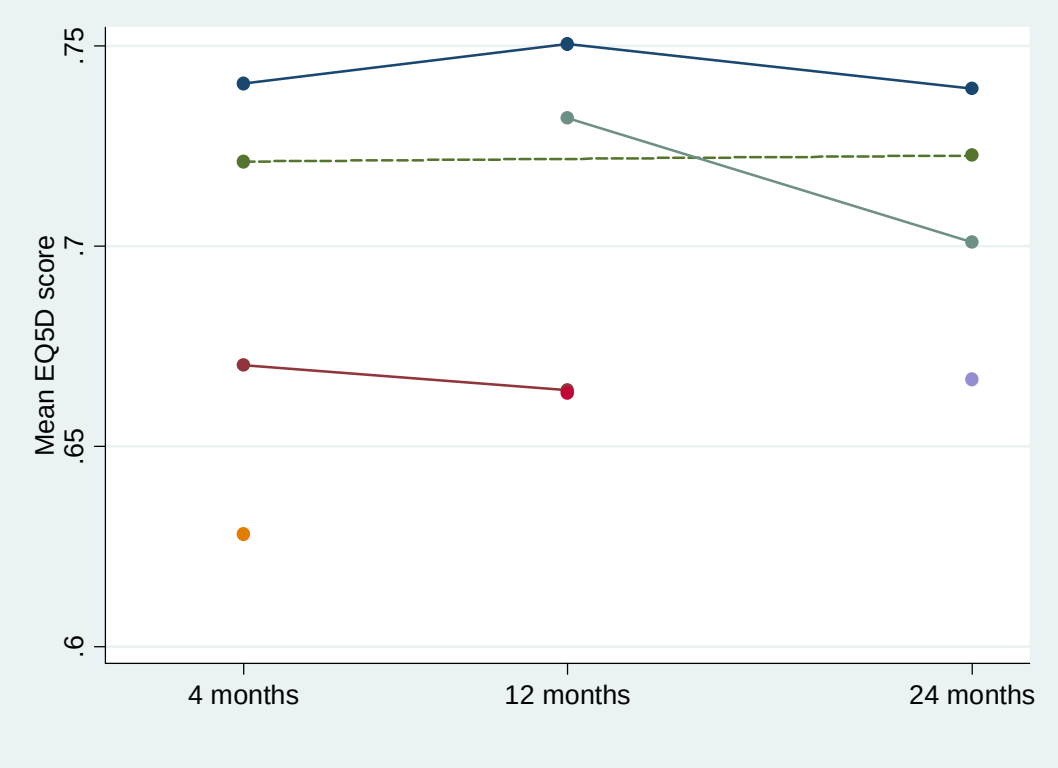


Figure 5.9: RECORD – SF12 physical component mean scores by missing data pattern

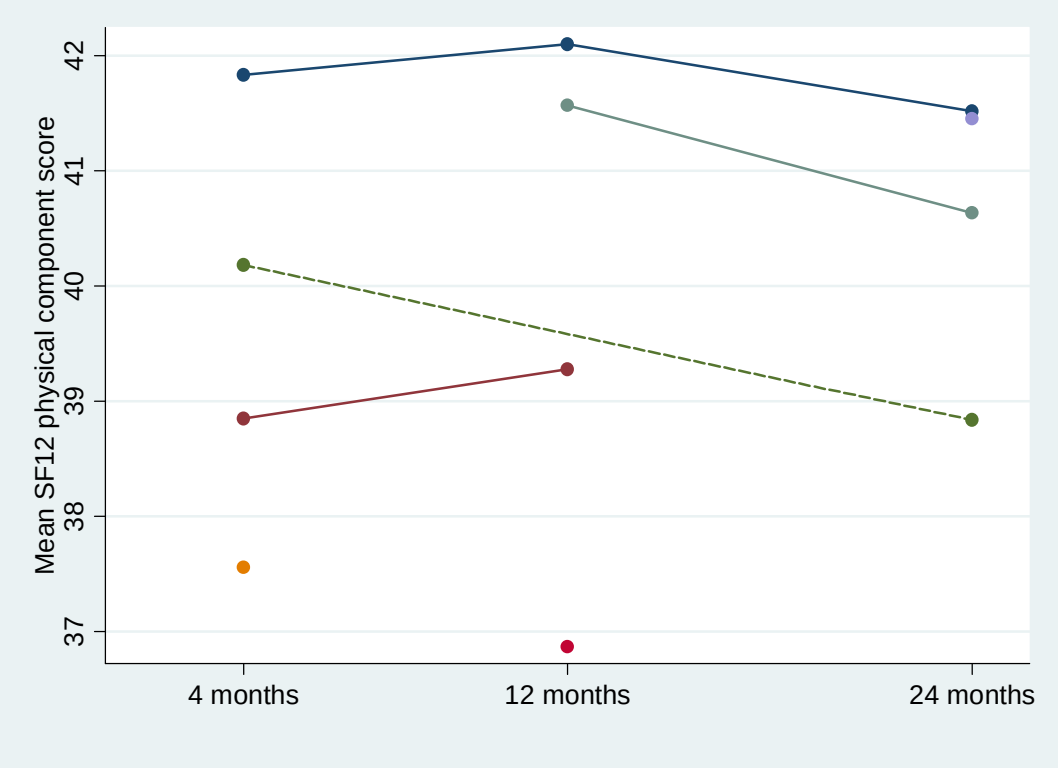
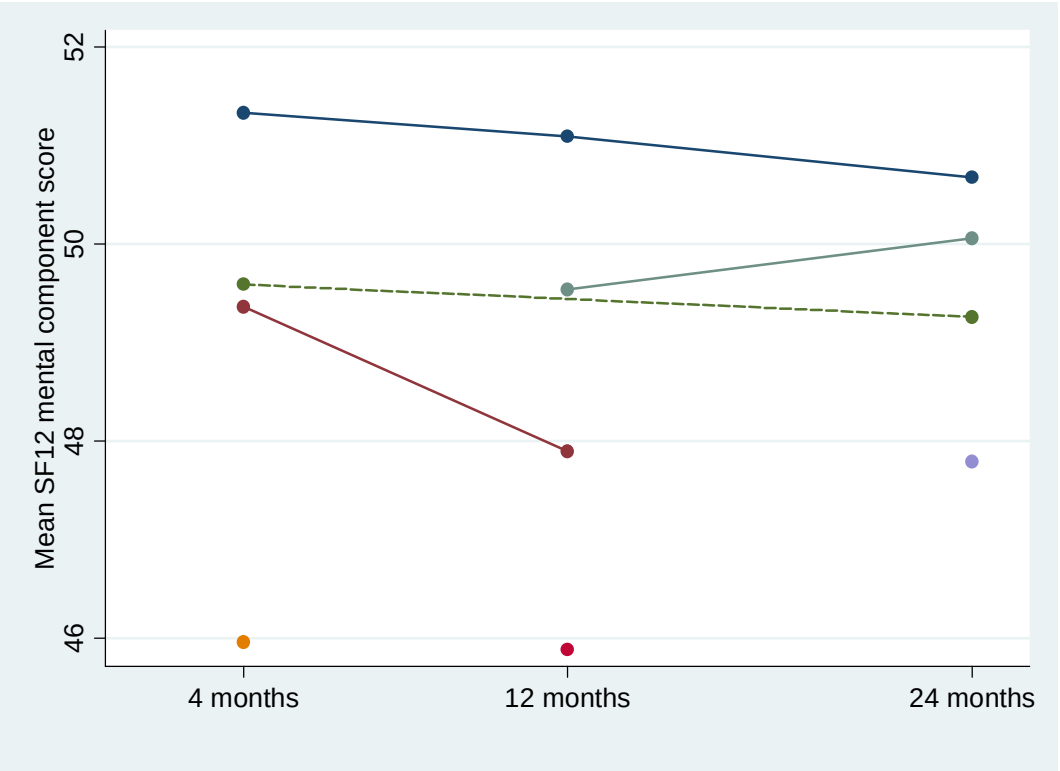


Figure 5.10: RECORD – SF12 mental component mean scores by missing data pattern



5.4.2 Little's test

Little's test provided significant evidence against the MCAR assumption in favour of MAR/MNAR for each of the three QoL scores and three data scenarios ($p<0.001$ in each case). Hence, QoL measured by the EQ5D, physical and mental components of the SF12 affected whether or not a patient responded immediately, responded at all or, having responded, whether they did so by reminder.

5.4.3 Ridout Logistic regression

Scenario one

Table 5.5 shows the results of Ridout logistic regression under data scenario one for RECORD. At four months 55% of participants provided an EQ5D score, of which 17% subsequently dropped out and provided no further assessments. Adjusting for age group and locomotor ability, there was a significant difference in the four month EQ5D scores ($p<0.001$). At 12 months, a similar phenomenon was seen with adjusted OR = 0.49 (0.35, 0.70). Similarly, having adjusted for some baseline covariates, the physical and mental component scores were significantly

different between continuers and drop outs at both four and 12 months (Table 5.5). In each case, there was evidence of MAR data.

Table 5.5: RECORD - Ridout logistic regression results (scenario one)

	Total	Dropouts	Unadjusted		Adjusted	
	N (%)	N (%)	OR (95% CI)	p-value	OR(95% CI)	p-value
EQ5D						
4 months	2908 (55)	507 (17)	0.31 (0.21,0.45)	<0.001	0.44 (0.30,0.65)[1]	<0.001
12 months	2648 (50)	704 (27)	0.41 (0.30,0.58)	<0.001	0.49 (0.35,0.70)[2]	<0.001
SF12 physical component score						
4 months	2732 (52)	483 (18)	0.98 (0.97,0.98)	<0.001	0.98 (0.97,0.99)[3]	<0.001
12 months	2559 (48)	683 (27)	0.98 (0.97,0.99)	<0.001	0.987 (0.978,0.995)[4]	<0.001
SF12 mental component score						
4 months	2732 (52)	483 (18)	0.96 (0.95,0.97)	<0.001	0.97 (0.96,0.98)[3]	<0.001
12 months	2559 (48)	683 (27)	0.97 (0.96,0.98)	<0.001	0.975 (0.97,0.98)[4]	<0.001

Adjusted for: [1] age group and locomotor ability; [2] age group, sex and residence type prior to fracture; [3] age group, residence type prior to fracture, locomotor ability; [4] type of recruiting fracture, age group, residence type after fracture

Scenario two

EQ5D scores were obtained for 3907 patients at four months, of which 88% continued in the study and 12% dropped out after this time. There was a significant difference in the EQ5D score between the two groups ($p<0.001$), with those dropping out displaying on average 0.1 units worse QoL. Locomotor ability ($p<0.001$), type of recruiting fracture ($p=0.001$), age group ($p<0.001$) and residence type before and after fracture (both $p<0.001$) were significantly associated with dropout. In a logistic model adjusting for the other covariates (age group, locomotor ability, treatment and type of fracture), the four month EQ5D score was significant. The adjusted odds ratio for the EQ5D score was 0.34 (0.23, 0.49), implying those with a higher 4-month QoL score were less likely to drop out.

Appendix 5.4 shows the number of continuers and dropouts along with mean (SD) QoL scores at each of four and 12 months. In each case, a significant difference was found between the QoL scores ($p<0.001$). At 12 months, 3488 (66%) patients provided an assessment with 18% of these dropping out. The unadjusted odds ratio was 0.30 (0.22, 0.41), $p<0.001$ implying those with higher 12 month EQ5D scores were less likely to dropout. The covariates that were identified as being significant predictors of dropout were sex ($p=0.03$), locomotor ability

(p<0.001), type of fracture (p=0.001), age group (p<0.001) and residence prior to and after fracture (both p<0.001). In a logistic model, all but residence type prior to fracture remained significant. The adjusted OR for the 12 month EQ5D was 0.43 (0.30, 0.59), p<0.001, indicating those with higher QoL were significantly less likely to drop out.

The number of patients providing four-month SF12 component scores was 3644 (69%), with 87% continuing and 13% dropping out. Covariates associated with dropout after four months were marital status, type of fracture, age group and residence type prior to and after fracture. In a stepwise logistic model, age group and residence type after fracture were significant. There was a significant difference in the mean score of the PCS with mean difference 3.7 (2.5, 4.8), p<0.001. The addition of the four month PCS to the logistic model was significant (p<0.001). Similarly, the MCS was significantly different between continuers and dropouts, with mean difference 4.92 (3.8, 6.0), p<0.001. Adding this term into the logistic model was significant (p<0.001). In both cases, QoL was significant in the model and thus, data were likely MAR.

At 12 months, 3368 (64%) patients provided a score, with 82% continuing. The covariate model to predict dropout included locomotor ability, age group and residence type after fracture. There was a significant difference in the mean PCS (3.3 (2.3, 4.3), p<0.001). In the logistic model, the addition of the PCS was significant (p<0.001). For the MCS at 12m, there was also a difference between continuers and dropouts, with mean difference 3.45 (2.5, 4.5), p<0.001. This difference existed after adjusting for the identified covariates in a logistic model (p<0.001).

Scenario three

At four months, locomotor ability was a significant predictor of subsequent reminder response. No covariates were predictors of reminder response after the 12 month assessment. Appendix 5.5 shows the mean (SD) QoL scores and the unadjusted odds ratios for dropout. QoL was significant in the logistic model, indicating those with higher QoL scores were less likely to dropout. QoL was still

significant in the four month model, after adjusting for the covariate locomotor ability. Hence, there was evidence of MAR data for reminder response.

5.4.4 The LS Test

Scenario one

There were 2401 (45%) patients who displayed a monotone missing data pattern for the EQ5D score and 2249 (42%) for the SF12 component scores. The LS test statistic was statistically significant for the EQ5D score ($S=3.74$, $p<0.001$), providing evidence in favour of MAR. Similarly for the SF12 component scores, there was statistical evidence in favour of MAR for the PCS ($S=4.27$, $p<0.001$) and MCS ($S=5.16$, $p<0.001$).

Scenario two

There were 3634 (69%) patients which showed a monotone pattern of missingness for the EQ5D score, with 2606 (72%) responding to all three assessments. The LS test statistic was $S = 12.61$ ($p<0.001$), which was highly significant. The LS test statistic was $S=9.56$ ($p<0.001$) for the SF12 PCS and $S = 12.54$ ($p<0.001$) for the MCS. In all three cases, there was evidence in favour of the MAR assumption. Those dropping out earlier in the study displayed lower QoL.

Scenario three

In data scenario three, there was evidence in favour of MAR for each of the three QoL scores. The LS test statistic was $S=3.23$ ($p=0.002$) for the EQ5D score; $S=4.45$ ($p<0.001$) for the PCS; and $S=6.07$ ($p<0.001$) for the MCS. Those responding after reminder tended to display lower (worse) QoL.

5.4.5 Fairclough logistic regression

Scenario one

Covariates associated with missingness at 12 months were residence type after fracture, age group, locomotor ability, marital status and treatment. At 24 months age group, locomotor ability, treatment, type of recruiting fracture and residence type prior to fracture were associated with missingness.

Table 5.6: RECORD - Fairclough logistic regression results (scenario one)

	Complete N (%)	Missing N (%)	Adjusted OR for previous QoL OR (95% CI)	p-value
EQ5D				
12 months	2648 (50)	2644 (50)	0.66 (0.46, 0.93)[1]	0.018
24 months	2511 (47%)	2781 (53%)	0.46 (0.34, 0.62)[2]	<0.001
SF12 physical component score				
12 months	2559 (48%)	2733 (52%)	0.988 (0.980, 0.996)[1]	0.002
24 months	2480 (47%)	2812 (53%)	0.982 (0.974, 0.990)[2]	<0.001
SF12 mental component score				
12 months	2559 (48%)	2733 (52%)	0.977 (0.969, 0.985)[1]	<0.001
24 months	2480 (47%)	2812 (53%)	0.976 (0.969, 0.983)[2]	<0.001

Adjusted for: [1] residence type after fracture, age group, locomotor ability, marital status and treatment; [2] age group, locomotor ability, treatment, type of recruiting fracture and residence type prior to fracture.

Table 5.6 shows the results of Fairclough logistic regression for data scenario one. At 12 months there were 2648 (50%) patients who provided EQ5D scores, with 2644 (50%) missing. In a logistic model adjusting for the identified covariates, the OR for the previous QoL term was significant (p=0.018). At 24 months, the previous QoL term was also significant in the logistic model. In this instance, 2511 (47%) had provided response, with 2781 (53%) missing. Therefore, there was evidence of MAR for the EQ5D scores at both 12 and 24 months and those patients showing higher (better) QoL at previous assessments were less likely to provide missing response at the current assessment. With respect to the SF12 scores, there were 52% missing at 12 months and 53% missing at 24 months (Table 5.6). As with the EQ5D scores, missingness was found to be MAR for both the PCS and MCS at 12 and 24 months.

Scenario two

Age group, locomotor ability and residence type after fracture were related to missingness (scenario two) at both 12 and 24 months. The missing data mechanism was found to be MAR in all cases (Table 5.7). Those patients showing better previous QoL were less likely to provide missing response at the current assessment.

Table 5.7: RECORD - Fairclough logistic regression results (scenario two)

	Complete N (%)	Missing N (%)	Adjusted OR for previous QoL OR(95% CI)[1]	p-value
EQ5D				
12 months	3488 (66%)	1804 (34%)	0.48 (0.35, 0.66)	<0.001
24 months	3204 (61%)	2088 (39%)	0.37 (0.28, 0.48)	<0.001
SF12 physical component score				
12 months	3368 (64%)	1924 (32%)	0.98 (0.97, 0.99)	<0.001
24 months	3149 (60%)	2143 (40%)	0.98 (0.97, 0.99)	<0.001
SF12 mental component score				
12 months	3368 (64%)	1924 (32%)	0.97 (0.96, 0.98)	<0.001
24 months	3149 (60%)	2143 (40%)	0.97 (0.96, 0.98)	<0.001

Adjusted for: [1] residence type after fracture, age group, locomotor ability

Scenario three

At 12 months, residence type after fracture, age group, locomotor ability, marital status and sex were associated with reminder response. At 24 months, age group, locomotor ability and sex were associated with reminder-response. The current score was known (from reminder-response), therefore, the inclusion of the term in the logistic model could be assessed. The current QoL score was significant in the model having adjusted for covariates and previous QoL in each case, except for the PCS at 24 months (Table 5.8). Therefore, there was evidence that reminder responders were reporting lower QoL and that reminder response was not at random.

Table 5.8: RECORD - Fairclough logistic regression results (scenario three)

	Immediate N (%)	Reminder N (%)	Adjusted OR for current QoL OR (95% CI)	p-value
EQ5D				
12 months	2648 (76)	840 (24)	0.51 (0.34, 0.77)[1]	0.01
24 months	2511 (78)	693 (22)	0.49 (0.32, 0.76)[2]	0.001
SF12 physical component score				
12 months	2559 (76)	809 (24)	0.98 (0.97, 0.99)[1]	0.001
24 months	2480 (79)	669 (21)	0.997 (0.98, 1.01)[2]	0.64
SF12 mental component score				
12 months	2559 (76)	809 (24)	0.987 (0.977, 0.997)[1]	0.008
24 months	2480 (79)	669 (21)	0.985 (0.975, 0.991)[2]	0.005

Adjusted for: [1] residence type after fracture, marital status, locomotor ability; [2] age group, locomotor ability and sex

5.4.6 Summary

Scenario one
Each of the four methods agreed that data were likely to be MAR. This was the case for each of the three QoL scores. Those patients who displayed lower QoL were more likely to be non-responders.

Scenario two
The test proposed by Little, provided evidence against the MCAR assumption for each of the three QoL scores. This conclusion was also reached by the LS test, which provided evidence in favour of MAR data. In Ridout's logistic regression, at each of four and 12 months, the baseline QoL was an important predictor of dropout in addition to the covariates. This suggested missingness was at least MAR. Fairclough's logistic regression method agreed. Patients displaying lower QoL in their observed scores were more likely to be non-responders.

Scenario three
The investigation into the mechanism behind reminder response using the four different methods gave a similar conclusion. Little's test provided evidence against the MCAR assumption while the LS test provided evidence in favour of MAR. Ridout logistic regression found the data to be MAR. Fairclough logistic regression suggested reminder response was MNAR and that those with lower QoL were more likely to be reminder responders.

5.5 KAT

5.5.1 Pattern of missing data

Appendix 5.6 shows the number of participants and mean (SD) QoL scores for KAT at each assessment, split by missing data pattern. There were 16 potential missing data patterns. The mean QoL scores for each of these patterns are shown in Figure 5.11 for the EQ5D score, Figure 5.12 for the SF12 physical component score, Figure 5.13 for the mental component score and Figure 5.14 for the Oxford Knee Score (OKS). The majority of patients (93%) provided a baseline assessment.

There was an improvement in QoL scores post surgery compared to baseline for all patterns of missing data.

Those patients who missed one or more assessments tended to display lower QoL than those who responded at each of the four assessments. Those patients who provided a later assessment, but perhaps missed an early one, tended to show better QoL than those who missed later assessments. There did appear to be a relationship between QoL and the missing data pattern, with those patients who provided some missing assessments tending to show lower scores at the assessments that they did provide.

Figure 5.11: KAT – EQ5D mean score at each assessment by missing data pattern

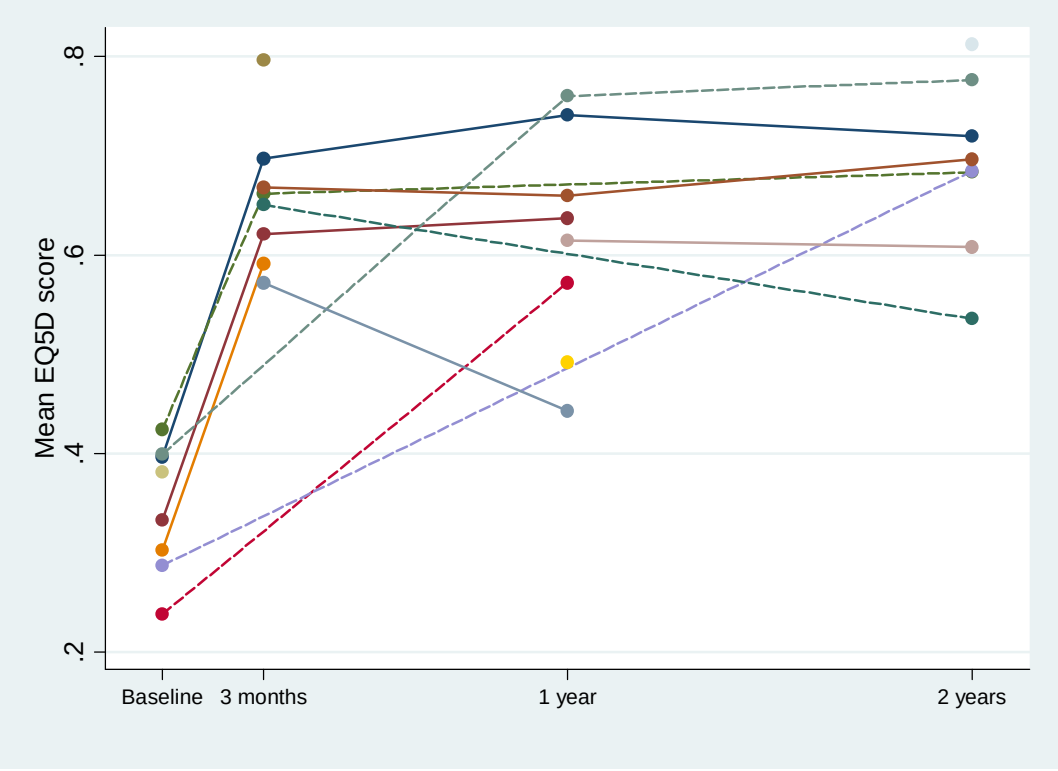


Figure 5.12: KAT – SF12 physical component mean score at each assessment by missing data pattern

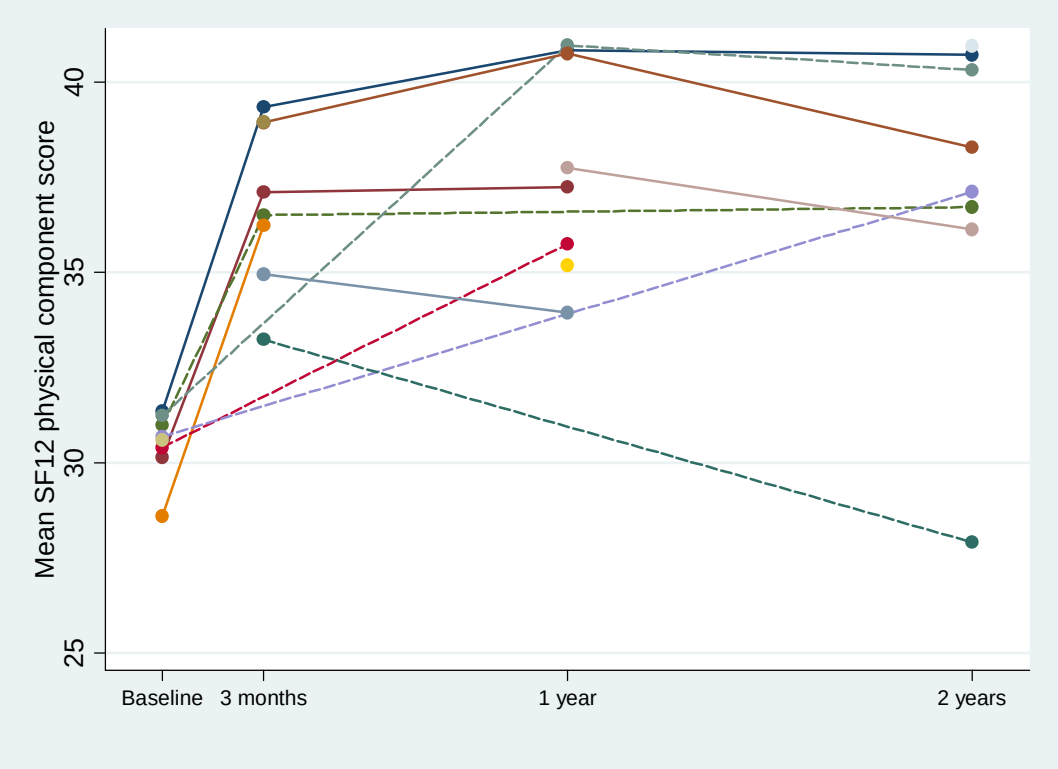


Figure 5.13: KAT – SF12 mental component mean score at each assessment by missing data pattern

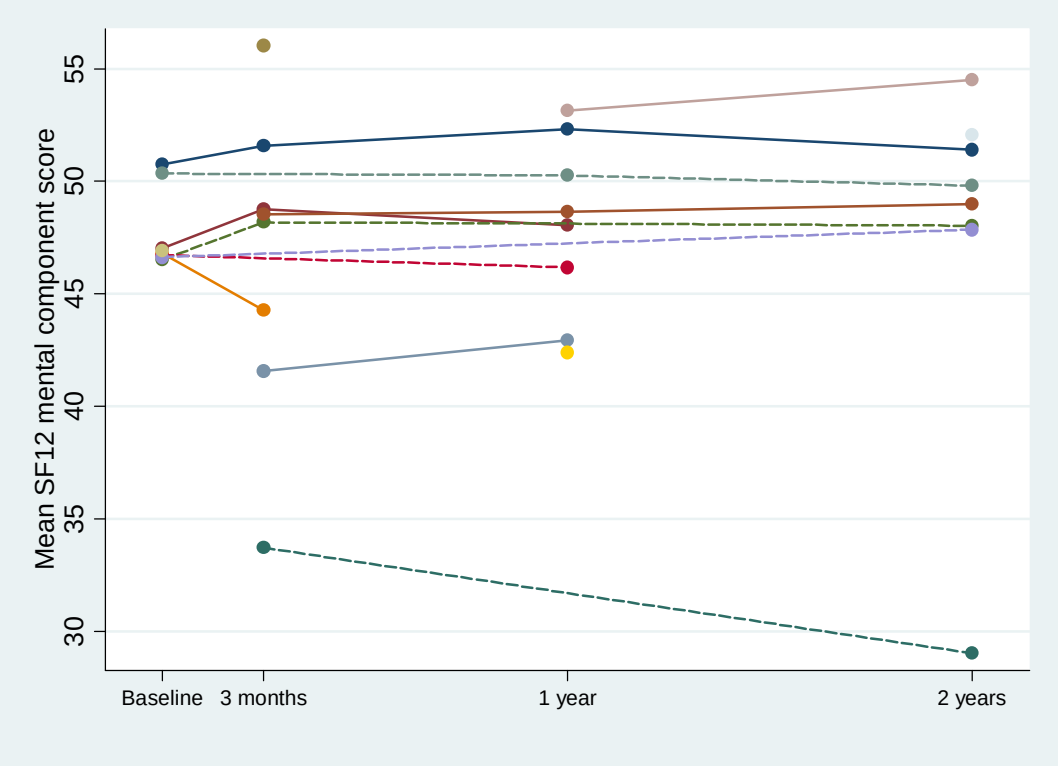
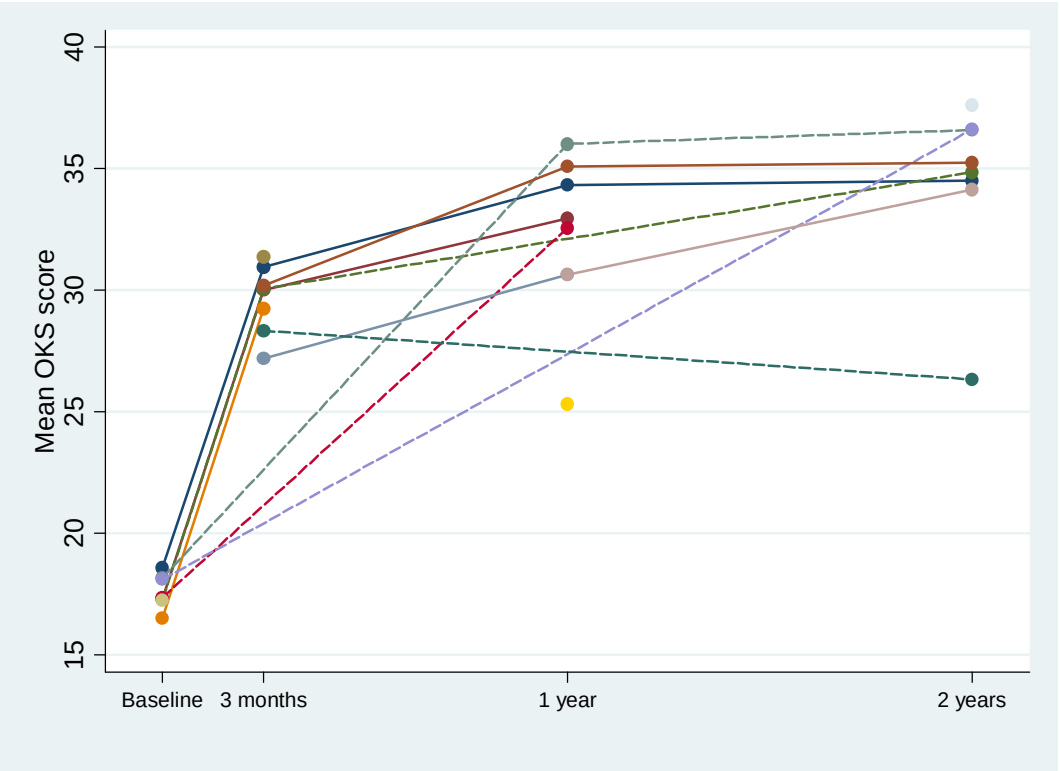


Figure 5.14: KAT – OKS mean score at each assessment by missing data pattern



5.5.2 Little's test

Table 5.9 shows the results of Little's test of MCAR for each of the data scenarios and QoL scores. Each hypothesis test found evidence against the null of MCAR and suggested that missingness was not MCAR. Therefore, QoL was a significant factor in response and if a patient did respond, it was also a significant factor in whether they did so immediately or after reminder (scenario three).

Table 5.9: KAT p-values from Little's test for MCAR

Scenario	EQ5D	PCS	MCS	OKS
One	<0.001	<0.001	<0.001	<0.001
Two	<0.001	<0.001	0.001	0.001
Three	<0.001	0.032	0.02	0.015

5.5.3 Ridout Logistic Regression

Scenario one

At baseline, whether or not the patient had further re-admissions was related to subsequent drop out. At three months and one year, ASA grade was associated with dropout. Apart from the physical score at baseline and the OKS at three

months, the QoL scores were predictive of dropout having adjusted for the identified covariates (Table 5.10). Those displaying better QoL were less likely to dropout at a subsequent assessment, implying MAR data. For the physical score at baseline and OKS at three months, there was no evidence against MCAR (covariate-dependent).

Table 5.10: KAT - Adjusted OR's for result of Ridout regression (scenario one)

	Number (%)		Adjusted OR (95% CI)	
Assessment	Total	Drop outs	for dropout	p-value
EQ5D				
Baseline	2195 (93)	175 (8)	0.52 (0.31, 0.85)[1]	0.01
3 months	1798 (76)	174 (10)	0.36 (0.20, 0.67)[2]	0.001
1 year	1701 (72)	363 (21)	0.31 (0.20, 0.50)[3]	<0.001
SF12 Physical component score				
Baseline	2162 (92)	188 (9)	-	-
3 months	1744 (33)	182 (10)	0.97 (0.95, 0.99)[2]	0.001
1 year	1663 (31)	361 (22)	0.97 (0.96, 0.98)[2]	<0.001
SF12 Mental component score				
Baseline	2162 (92)	188 (9)	0.97 (0.96, 0.99)[1]	<0.001
3 months	1744 (33)	182 (10)	0.98 (0.96, 0.99)[2]	<0.001
1 year	1663 (31)	361 (22)	0.97 (0.96, 0.98)[2]	<0.001
OKS				
Baseline	2192 (93)	264 (12)	0.97 (0.95, 0.99)[1]	0.001
3 months	1608 (68)	254 (16)	-	-
1 year	1477 (63)	446 (30)	0.98 (0.96, 0.99)[2]	<0.001

Adjusted for: [1] whether or not patient had re-admissions; [2] ASA grade.

Scenario two

Table 5.11 shows the results of Ridout logistic regression for scenario two. In a logistic model for dropout after baseline and secondly after three months, ASA grade ($p<0.001$) was significant. At one year, both age ($p=0.009$) and sex ($p=0.037$) were significant predictors of dropout. No difference was found between the mean baseline EQ5D scores of those continuing and those who dropped out after baseline ($p=0.90$). The adjusted odds ratio for dropout after three months was OR = 0.29 (0.13, 0.60) ($p<0.001$). Patients displaying higher QoL scores at six months were less likely to dropout with OR = 0.20 (0.12, 0.33) ($p<0.001$). This showed that

for the EQ5D score (except at baseline) current observed QoL was a significant predictor of drop out thereafter.

In the case of the SF12 scores, the covariates associated with dropout after baseline were ASA grade and any post-operative complications. At three months, ASA grade and age were significantly associated with dropout and finally, at one year age was associated with dropout. There was no significant difference in the PCS at baseline between the continuers and dropouts (p=0.35). The addition of the baseline MCS to the logistic model was significant (p<0.001). The adjusted OR (95% CI) was 0.97 (0.95, 0.99) (Table 5.11).

Table 5.11: KAT - Adjusted OR's for result of Ridout regression (scenario two)

	Number (%)		Adjusted	
Assessment	Total	Drop outs	OR (95% CI)	p-value
EQ5D				
Baseline	2195 (93)	94 (6)	0.95 (0.44, 2.08)[1]	0.90
3 months	2008 (85)	89 (4)	0.29 (0.13, 0.60)[1]	<0.001
1 year	1777 (89)	213 (11)	0.20 (0.12, 0.33)[2]	<0.001
SF12 Physical component score				
Baseline	2063 (95)	99 (5)	0.99 (0.96, 1.02)[3]	0.35
3 months	1850 (95)	94 (5)	0.97 (0.95, 0.99)[4]	0.008
1 year	1722 (88)	232 (12)	0.97 (0.95, 0.99)[5]	0.005
SF12 Mental component score				
Baseline	2063 (95)	99 (5)	0.97 (0.95, 0.99)[3]	<0.001
3 months	1850 (95)	94 (5)	0.95 (0.93, 0.97)[4]	<0.001
1 year	1722 (88)	232 (12)	0.97 (0.96, 0.99)[5]	0.001

Adjusted for: [1] ASA grade; [2] age and sex; [3] ASA grade and any complications; [4] ASA grade and age; [5] age.

At three months both the PCS and MCS of continuers were significantly different to the drop outs (p=0.008 and p<0.001 respectively). This difference was still evident having adjusted for the appropriate covariates (Table 5.11). For the one-year SF12 component scores the current observed QoL was found to be a significant (p<0.001) predictor of dropout in the logistic model. The adjusted (for age) OR= 0.97 (0.95, 0.99) for the PCS and 0.97 (0.96, 0.99) for the MCS. At baseline and at three months, there was no difference in the OKS scores of those continuing and those dropping out (p=0.19 and p=0.08 respectively). At 12 months,

subsequent dropout was associated with the baseline OKS score but not the 12 month scores.

Scenario three

No covariates were found to be associated with dropout (subsequent reminder response) at baseline or three months. The current EQ5D was not a predictor of dropout at baseline (p=0.28) and borderline significant at three months (OR = 0.36 (0.13, 1.01), p=0.053). At one year however, ASA grade was a significant predictor of reminder response. The adjusted odds ratio for dropout was 0.40 (0.22, 0.75), p=0.004, implying current QoL measured by the EQ5D was a significant predictor of dropout. This suggested the mechanism behind reminder response in this trial was potentially MAR.

In the case of the SF12 PCS, no significant association was found with dropout at baseline (p=0.61). At three months, some association with dropout was found (OR=0.97 (0.94, 0.99), p=0.04). Finally at one year, after adjusting for ASA grade and three month PCS, the twelve month PCS was not significant in the model (p=0.84). Results for the MCS and Oxford Knee score were similar. Neither mental QoL, nor functional status (OKS), was found to be predictive of subsequent reminder response at baseline or at three months. While at one year, after adjusting for ASA grade and three month MCS, the one year score was not an important factor in the model, with adjusted OR = 0.99 (0.97,1.01), p=0.29. However, in the case of the OKS the one year score was borderline significant in the model (p=0.05) with adjusted OR = 0.97 (0.94, 1.00). This suggested that there was a possibility of MAR for reminder response.

5.5.4 The LS test

Scenario one

There were 1771 (75%) patients who displayed a monotone missingness pattern for the EQ5D data. The LS test statistic S=7.21 (p<0.001) was significant and provided evidence in favour of the MAR assumption. A similar finding was seen for the SF12 component scores and OKS. There were 1705 (72%) patients who

provided a monotone missingness pattern for the SF12 scores and 1652 (70%) for the OKS. The LS test statistic for the PCS was S = 5.90 ($p<0.001$); MCS S=7.52 ($p<0.001$); OKS S=5.45 ($p<0.001$). In each case, there was evidence in favour of MAR suggesting QoL experience was affecting whether or not patients responded to the follow up QoL questionnaires.

Scenario two

As in scenario one, there was evidence in favour of MAR in each case. There were 1983 (84%) patients with monotone missingness for the EQ5D score and the LS test statistic S=5.86 was significant ($p<0.001$). With respect to SF12 scores 1881 (80%) patients showed a monotone missingness pattern. The test statistic was S=5.40 ($p<0.001$) and S=8.50 ($p<0.001$) for the PCS and MCS respectively. Finally, for the OKS, there were 1763 (75%) with a monotone pattern and S=3.69 ($p<0.001$). In each case the observed QoL scores were shown to be impacting on response and there was evidence in favour of MAR.

Scenario three

There was evidence in favour of MAR for each of the four QoL scores. The results of the LS test were: EQ5D (S=3.99, $p<0.001$); PCS (S=3.33, $p=0.002$); MCS (S=2.12, $p=0.042$); OKS (S=4.64, $p<0.001$). Hence, there was evidence in favour of MAR for reminder response and it was found that observed QoL was related to whether or not a patient responded after reminder or not.

5.5.5 Fairclough logistic regression

Scenario one

At three months whether or not the patient had re-admissions was associated with missingness for the EQ5D and OKS score. At one year, type of knee arthritis and any post-operative complications were associated with missing EQ5D and SF12 score. At two years, missing EQ5D and SF12 scores were associated with ASA grade. Appendix 5.7 displays the adjusted OR for the previous QoL term from Fairclough's logistic regression procedure. The previous EQ5D scores were significant at each of three months, one year and two years, suggesting missingness was MAR. The previous PCS was not significant at three months

(suggesting MCAR), but was significant at one and two years (suggesting MAR). The same phenomenon was seen for the OKS. Previous MCS were significant in the model at each of three months, one year and two years, suggesting missing MCS was MAR.

Scenario two

At three months, the extent of knee arthritis (one knee) and whether they had re-admissions was found to differ between responders and non-responders (26% vs. 22%, p=0.03 and 8% vs. 2%, p<0.001 respectively). At one year, extent of knee arthritis (p=0.011), post-operative complications (p=0.013), readmissions and ASA grade (<0.001) were related to non-response. At 2 years, ASA grade (p<0.001) and type of knee arthritis (p=0.011) were significantly associated with non-response. Previous QoL between responders and non-responders was compared. A difference existed for the MCS at three months and all the QoL scores at both one and two years.

A logistic model was created to assess the inclusion of previous QoL in the model having adjusted for the covariates. The adjusted odds ratios for the previous QoL term are shown Appendix 5.8. It can be seen that previous QoL experience was still a predictor of non-response for both the EQ5D and SF12 at one and two years, but only for the MCS at three months. However, in general it appeared that previous QoL was a predictor of non-response after adjusting for appropriate covariates.

Scenario three

Age was found to differ between immediate and reminder responders at both three months (p=0.002) and one year (p=0.005). In addition, at one year, the type of knee arthritis (p=0.04) differed between the two types of responders. At two years, no covariates were found to be associated with responder type. Both the previous EQ5D and previous OKS scores differed between immediate and reminder-responders at all three assessments. The SF12 PCS differed at one and two years, with the MCS differing only at 2 years.

It was possible to investigate whether current QoL was affecting at what point the patients responded (immediately or after reminder). After adjusting for covariates, and previous QoL, the current QoL was an important addition to the model at 3 months for EQ5D (p=0.009) and PCS (p=0.001), and at one year for all three QoL measures (EQ5D (p=0.017), PCS (p=0.003) and MCS (p=0.007)). At 3 months and one year, previous OKS was significant in the model, but current QoL was not. While at 2 years, the previous OKS was not significant, but current scores were (p=0.023). Therefore, there was evidence of possible MNAR data at three months for EQ5D and PCS; at one year for EQ5D and both SF12 component scores; at two years for the OKS. Reminder response was found to be MAR in the remaining cases.

5.5.6 Summary

Scenario one

Little's test suggested the data were not MCAR, while the LS test found evidence in favour of MAR for the missing QoL scores in scenario one. Similarly, both Ridout logistic regression and Fairclough logistic regression suggested missingness was MAR.

Scenario two

In summary, Little's test was found to be significant for each of the QoL scores. This provided evidence against the MCAR assumption, implying the QoL experience differed between those responding and not responding to the questionnaires. Those not responding tended to provide lower (poorer) QoL scores. The results of the LS test agreed with this conclusion, finding evidence in favour of MAR for all QoL scores. Using Ridout logistic regression found evidence of covariate dependent missingness at baseline for the EQ5D score with MAR at both three months and one year. There was evidence of covariate-dependent missingness for the PCS at baseline and three months and evidence of at least MAR for the MCS and PCS at 12 months. Functional status measured by the OKS impacted on dropout after the 12 months assessment, with those with lower baseline QoL more likely to drop out. Fairclough's logistic regression found

that in addition to identified covariates, observed QoL was an important factor in predicting missingness, providing evidence of MAR data.

Scenario three
Little's test provided evidence that the QoL experience differed between the immediate and reminder responders. The LS test also found in favour of MAR for reminder response. In Fairclough's logistic regression, current QoL (suggesting MNAR) was significant at one year, but only significant in the model at three months for the EQ5D and PCS. Ridout logistic regression found observed QoL to be predictive of subsequent reminder response only after the one year assessment.

5.6 PRISM

5.6.1 Pattern of missing data

Appendix 5.9 displays the number of participants and mean (SD) QoL scores observed during PRISM split by missing data pattern. This information is displayed graphically in Figure 5.15 (EQ5D), Figure 5.16 (SF36 PCS), Figure 5.17 (SF12 MCS) and Figure 5.18 (Arthritis Index). Only 4% of participants completed each of the five assessments. The EQ5D score increased with follow up from 0.58 at baseline to 0.64 at four years. Those patients with the lowest baseline scores tended to provide few or no subsequent assessments. The PCS tended to decrease from baseline to follow up within a particular pattern, but overall the average across patterns was 36.3 at baseline compared with 37.1 at four years. The average MCS across response pattern was 48.7 at baseline compared to 47.4 at four years. The Arthritis index score rose from 35.8 units at baseline to 36.9 at four years. A feature of the data for each of the QoL scores was that patients with a greater number of observed assessments tended to report the higher (better) QoL.

Figure 5.15: PRISM – EQ5D mean score by missing data pattern

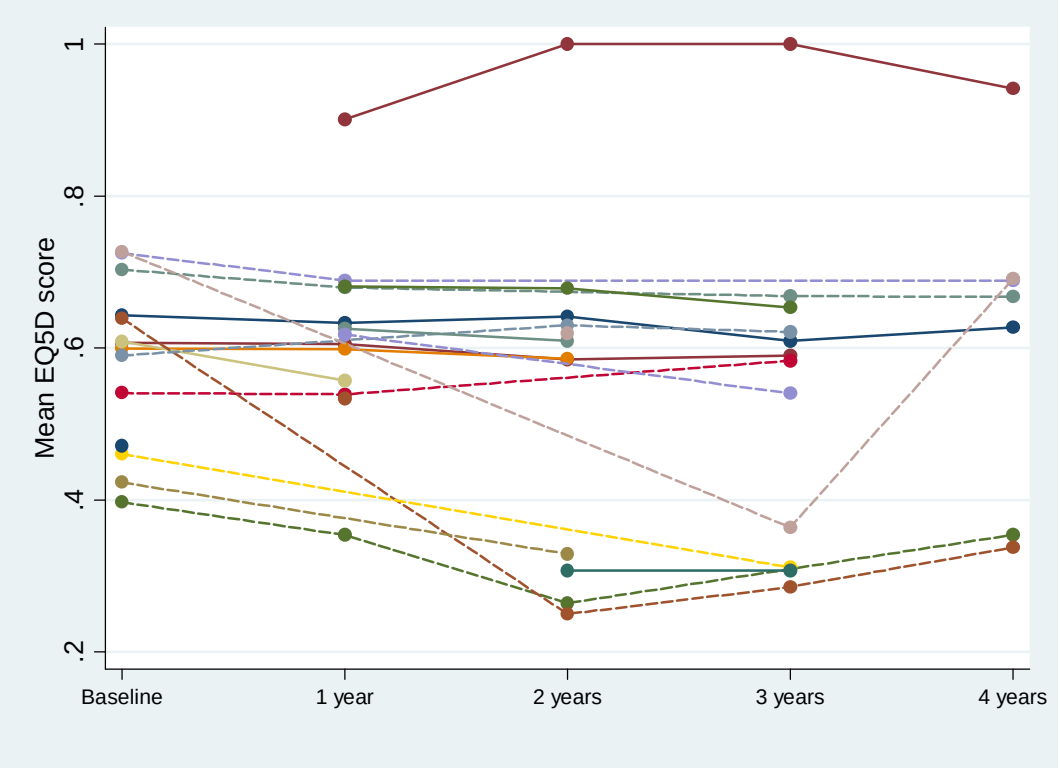


Figure 5.16: PRISM – SF36 physical component mean score by missing data pattern

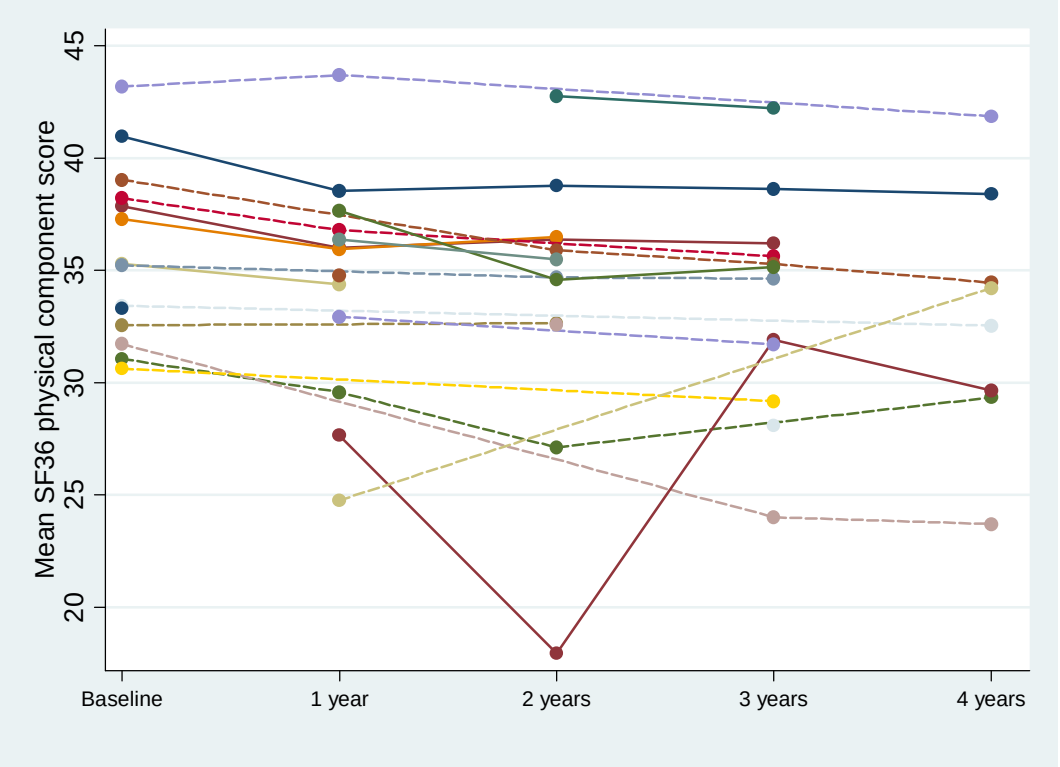


Figure 5.17: PRISM – SF36 mental component mean score by missing data pattern

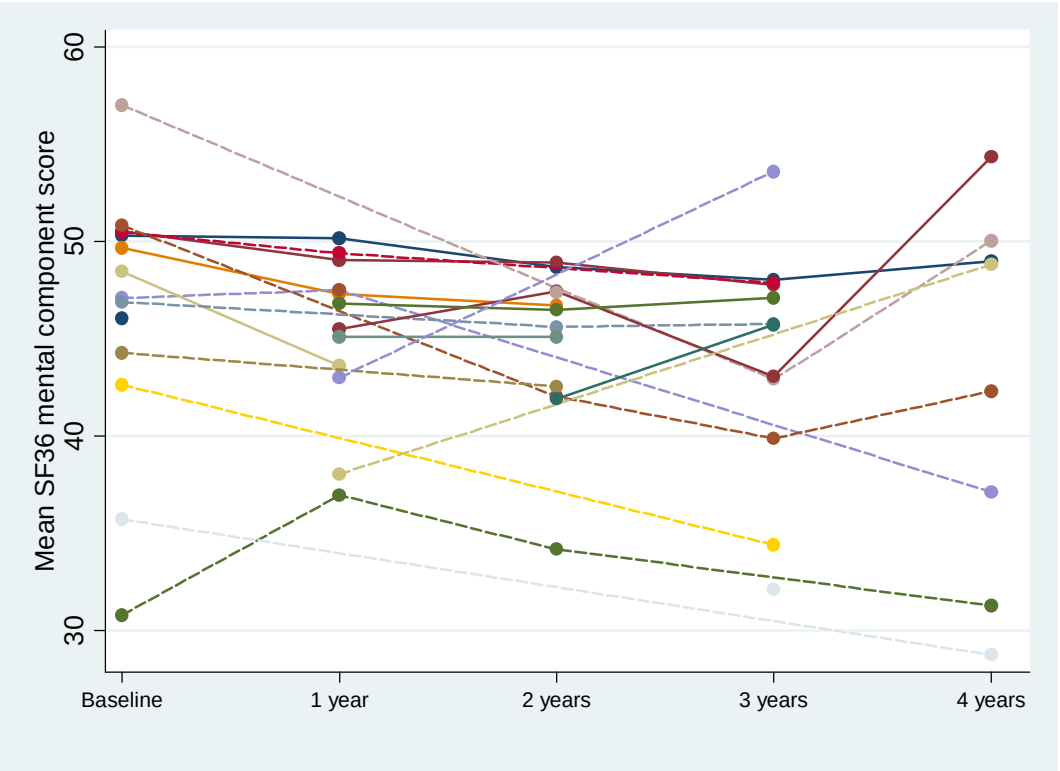
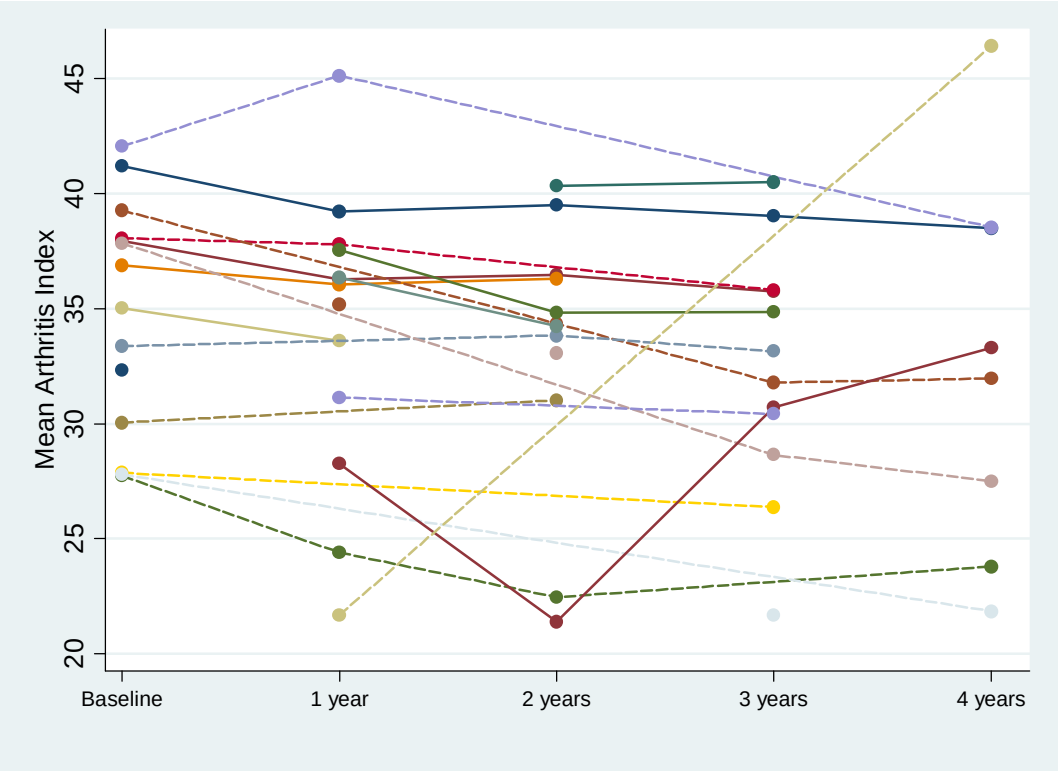


Figure 5.18: PRISM – Arthritis Index mean score by missing data pattern



5.6.2 Little's test

Table 5.12 shows the p-values arising from Little's test of the null hypothesis of MCAR. In scenario one, there was evidence against MCAR for the mental summary score and Arthritis Index. Evidence against the null hypothesis was borderline for the PCS (p=0.05).

Table 5.12: PRISM - Little's test for MCAR p-value

Scenario	EQ5D	PCS	MCS	Arthritis Index
One	0.44	0.05	<0.001	0.002
Two	0.014	0.013	<0.001	<0.001
Three	0.82	0.46	0.43	0.45

In scenario two, the null hypothesis was rejected for each of the QoL scores ($p<0.05$) concluding that missingness was not MCAR. Those patients displaying lower QoL scores early on were more likely to provide missing responses. In scenario three there was no evidence against the MCAR assumption ($p>0.05$) for any of the four QoL scores. This implied that of those patients responding, QoL did not affect whether they did so immediately or after reminder.

5.6.3 Ridout Logistic regression

Scenario one

Table 5.13 displays the results of the Ridout logistic regression process for data scenario one. Prior to adjusting for covariates, baseline ($p<0.001$) and one year ($p=0.03$) EQ5D scores were predictors of subsequent dropout. Adjusting for the relevant covariates, baseline scores were still significant in the model but were borderline significant for the one year scores. There was no evidence of MAR after the two and three year assessments.

In the case of the PCS and the Arthritis index, there was a difference in baseline scores between those who continued and those who dropped out ($p<0.001$). Patients displaying better scores were less likely to dropout. At one, two and three years follow up, there was no evidence against MCAR as QoL scores did not differ between continuers and dropouts ($p>0.05$). For the MCS, it was at one year

where observed QoL was a predictor of subsequent dropout (p<0.001). On the whole, there was evidence that dropout after baseline was dependent on baseline scores, but drop out thereafter was not, and was found to be MCAR.

Table 5.13: PRISM - Adjusted OR's for result of Ridout regression (scenario one)

	Number (%)		OR (95% CI)	
Assessment	Total	Drop outs	for dropout	p-value
EQ5D				
Baseline	1250 (94)	212 (74)	0.43 (0.27, 0.68)[1]	<0.001
1 year	1018 (77)	278 (27)	0.65 (0.41, 1.04)[2]	0.07
2 years	740 (56)	343 (53)	0.85 (0.52, 1.40)[1]	0.52
3 years	408 (31)	369 (90)	0.85 (0.20, 3.52)[3]	0.82
SF36 Physical component score				
Baseline	1198 (90)	247 (21)	0.98 (0.96, 0.99)[4]	<0.001
1 year	888 (67)	242 (27)	0.99 (0.98, 1.00)*	0.20
2 years	718 (54)	410 (57)	1.00 (0.98, 1.01)*	0.58
3 years	367 (28)	330 (90)	0.98 (0.95, 1.01)*	0.23
SF36 Mental component score				
Baseline	1198 (90)	247 (21)	0.98 (0.96, 1.01)[4]	0.28
1 year	888 (67)	242 (27)	0.97 (0.96, 0.99)[5]	<0.001
2 years	718 (54)	410 (57)	0.99 (0.98, 1.00)[6]	0.11
3 years	367 (28)	330 (90)	0.98 (0.96, 1.01)*	0.28
Arthritis Index (ASHI)				
Baseline	1198 (90)	247 (21)	0.97 (0.96, 0.99)[4]	<0.001
1 year	888 (67)	242 (27)	0.99 (0.98, 1.00)[5]	0.11
2 years	718 (54)	410 (57)	0.99 (0.98, 1.01)*	0.40
3 years	367 (28)	330 (90)	0.97 (0.95, 1.00)[7]	0.055

* Unadjusted; Adjusted for: [1] if patient has siblings; [2] marital status and if parents were alive; [3] femur bone involved in Paget's and if there was clinician reported bone pain; [4] marital status; [5] marital status and if patient had siblings; [6] past fractures and if parents were alive; [7] if the patient had received a bone scan.

Scenario two

Appendix 5.10 shows the results for Ridout's logistic regression for data scenario two. At baseline, marital status (p=0.03) and having siblings (p=0.013) were associated with subsequent missing EQ5D scores. At one year, marital status (p=0.016) and whether or not the pelvic bone was involved in Paget's disease (p=0.006) were associated with dropout. At two years only existence of siblings was significant (p=0.038). At three years, having had a bone scan (p=0.002), femur bone involved with Paget's (p=0.023) and clinician reported bone pain (p=0.011)

had significant associations with the indicator of dropout. In a logistic model for post-baseline dropout, the addition of the baseline EQ5D score was significant and the adjusted odds ratio was OR= 0.31 (0.18, 0.52). Similarly, at one year, after adjusting for identified covariates, the EQ5D score was significant with adjusted OR = 0.58 (0.34, 0.98). At two and three years however, once the covariates had been adjusted for, the EQ5D was not a significant predictor of dropout.

Covariates associated with missing SF12 scores after baseline were marital status and at one year, whether the patient had previous vertebral fractures. Drop out after two years was associated with whether or not a patient had received a bone scan, while at year three in addition to the bone scan, whether or not a patient had a femoral fracture was associated with dropout. In the case of the PCS, at baseline the component score was an important predictor in addition to the covariates (p=0.001). At the remaining assessments, the PCS was not required in the logistic model. In contrast, the MCS was an important predictor of dropout after baseline, year one and year two, but not at year three. Hence, in general there was evidence of MAR in respect of the mental component of QoL but not the physical component.

The arthritis index differed at baseline and year one between those who continued and those who dropped out, with those remaining in the study displaying a better arthritis index. The covariates associated with missingness were the same as for the SF12 scores. At baseline, the Arthritis index was important in the logistic model ($p<0.001$). At one year (p=0.08), at year two (p=0.24) and at year three (p=0.16), having adjusted for the appropriate covariates, the current QoL was not a significant predictor of subsequent dropout.

Scenario three

There was insufficient data for responders at years three and four to be included. This section was restricted to those responding at baseline, year one and year two. No covariates were identified to be significantly associated with reminder response. Table 5.14 shows the results of Ridout logistic regression, where the dropout indicator being modelled was that of reminder response. At none of the

assessments were the four QoL scores significantly different between continuers and dropouts. Therefore, there was no evidence to show that the observed QoL was impacting on subsequent reminder response and the mechanism could be assumed to be MCAR.

Table 5.14: PRISM - Ridout logistic regression results for reminder response

	Continuers		Drop outs		Unadjusted	
Assessment	N	Mean (SD)	N	Mean (SD)	OR (95% CI)	p-value
EQ5D						
Baseline	775	0.61 (0.29)	16	0.59 (0.25)	0.79 (0.15, 4.12)	0.78
1 year	637	0.61 (0.28)	115	0.57 (0.31)	0.65 (0.33, 1.27)	0.21
SF36 Physical component score						
Baseline	678	37.8 (11.5)	14	37.1 (9.5)	0.99(0.95, 1.04)	0.83
1 year	562	36.0 (10.9)	96	36.4 (11.4)	1.00 (0.98, 1.02)	0.73
SF12 Mental component score						
Baseline	678	50.0 (11.1)	14	48.8 (13.2)	0.99 (0.95, 1.04)	0.68
1 year	562	48.5 (11.4)	96	47.1 (12.3)	0.99 (0.97, 1.01)	0.27
Arthritis Index (ASHI)						
Baseline	678	37.7 (12.7)	14	36.2 (12.3)	0.99 (0.95, 1.03)	0.67
1 year	562	36.3 (11.9)	96	36.3 (12.4)	1.00 (0.98, 1.02)	0.94

5.6.4 The LS test

Scenario one

In data scenario one, there were 1103 (83%) patients who displayed a monotone pattern for the EQ5D score and 1023 (77%) for the PCS, MCS and Arthritis Index. There was no evidence in favour of MAR for any of the four QoL measures with the results of the LS test as follows: EQ5D score S=0.45 (p=0.36); PCS S=0.40 (p=0.37); MCS S=0.48 (p=0.36); Arthritis Index S=0.53 (p=0.35).

Scenario two

There were 1118 (84%) patients who showed a monotone pattern of missingness for the EQ5D. The LS test statistic was S=0.47 (p=0.36), therefore providing insufficient evidence against the assumption of MCAR. For both SF12 component scores and the Arthritis Index 1023 (77%) patients showed a monotone missingness pattern. The LS test statistic was S=0.53 (p=0.35) for the PCS, S=0.46 (p=0.36) for the MCS and S=0.62 (p=0.33) for the Arthritis Index. In all three cases, there was no evidence in favour of MAR.

Scenario three

Only 55 patients responded at all five time points, resulting in an insufficient number of response patterns to allow the calculation of the LS test statistic. The data was therefore restricted to the responders of the first three assessments. Appendix 5.11 shows the results of the LS test for the mechanism behind reminder response of the first three assessments. The result was non-significant for all four QoL scores, providing insufficient evidence in favour of MAR. Therefore, observed QoL was not found to impact on whether a patient, who responded immediately, subsequently went onto to being a reminder responder.

5.6.5 Fairclough logistic regression

Scenario one

Table 5.15 shows the results of Fairclough logistic regression.

Table 5.15: PRISM - Fairclough logistic regression results (scenario one)

Assessment	Complete N (%)	Missing N (%)	Previous QoL term OR (95% CI)	p-value
EQ5D				
1 year	1018 (77)	306 (23)	0.45 (0.29, 0.68)	<0.001
2 years	740 (56)	584 (44)	0.52 (0.36, 0.75)[1]	<0.001
3 years	408 (31)	916 (69)	0.71 (0.48, 1.05)[2]	0.08
4 years	53 (4)	1271 (96)	0.76 (0.30, 1.93)[3]	0.56
SF36 Physical component score				
1 year	888 (67)	436 (33)	0.98 (0.96, 0.99)[2]	<0.001
2 years	718 (54)	606 (46)	0.99 (0.98, 1.00)[1]	0.07
3 years	367 (28)	957 (72)	1.00 (0.99, 1.01)[4]	0.58
4 years	51 (4)	1273 (96)	0.99 (0.96, 1.02)[3]	0.42
SF36 Mental component score				
1 year	888 (67)	436 (33)	0.97 (0.96, 0.98)[2]	<0.001
2 years	718 (54)	606 (46)	0.98 (0.97, 0.99)[1]	0.001
3 years	367 (28)	957 (72)	0.98 (0.97, 0.99)[4]	0.004
4 years	51 (4)	1273 (96)	0.99 (0.97, 1.02)[3]	0.66
Arthritis Index (ASHI)				
1 year	888 (67)	436 (33)	0.97 (0.96, 0.98)[2]	<0.001
2 years	718 (54)	606 (46)	0.987 (0.977, 0.997)[1]	0.007
3 years	367 (28)	957 (72)	0.99 (0.98, 1.00)[4]	0.18
4 years	51 (4)	1273 (96)	0.98 (0.96, 1.01)[3]	0.18

Adjusted for: [1] age and marital status; [2] age; [3] age and whether or not had bone scan; [4] age and whether or not had previous fracture.

There was evidence that previous QoL was a predictor of missing EQ5D and Arthritis Index scores at one and two years; missing PCS at one year; missing MCS

at one, two and three years follow up. Those displaying better previous QoL were less likely to provide missing response. There was evidence that missingness was MAR at a number of assessments.

Scenario two

The proportion of patients not providing the EQ5D rose from 3% at baseline to 95% at four years (section 3.6). At each assessment, the association with a number of baseline variables was assessed. After a stepwise procedure, the covariate logistic model included age and type of Paget's disease at year one; age, marital status and pelvic bone involved with Paget's at year two; age and pelvic bone involved with Paget's at year three; age, had a bone scan, and femur involved in Paget's disease at year four.

The inclusion of the EQ5D term in the logistic model was significant at year one ($p<0.001$), year two ($p<0.001$) and year three ($p=0.036$), but at four years EQ5D was not a significant predictor of missingness in addition to the covariates ($p=0.65$). At year one, the adjusted odds ratio for previous QoL was 0.33 (0.21, 0.53), suggesting those with better previous QoL were more likely to complete the current EQ5D assessment. This trend was repeated at year two and three with OR = 0.32 (0.22, 0.50) and OR = 0.68 (0.47, 0.98) respectively.

A significant association with non-response to the SF12 questionnaire was found for age at year one; age, marital status, pelvic bone involved in Paget's at year two; age, pelvic bone involved in Paget's, previous clinical vertebrae fractures at year three; age, femur bone involved in Paget's, had a previous bone scan at year four. In a logistic model, the addition of previous QoL was significant at one and two years for each of the PCS ($p<0.001$ and $p=0.041$), MCS ($p<0.001$ and $p=0.001$) and Arthritis Index ($p<0.001$ and $p=0.004$). This suggests that the missingness was at least MAR, as the missingness was dependent on observed values. At year three and year four, the addition of the previous QoL was not a significant predictor of missingness having adjusted for the identified covariates.

Scenario three

The proportion completing after reminder ranged from 6% at year one to the highest of 27% at year three. The covariates that showed significant associations with reminder response to the EQ5D questions, at year one were ($p<0.001$), marital status ($p<0.001$), pelvic bone involved in Paget's ($p<0.001$), and one or more parents still alive ($p=0.005$). At year two, pelvic bone involved in Paget's ($p=0.032$) and history of previous fractures ($p=0.013$) were important. At year three and year four, none of the baseline covariates were found to have an association with the type of responder (immediate or reminder). There was no significant difference in the previous EQ5D score between immediate and reminder responders at any of the four assessments ($p>0.05$). Previous QoL did not differ between the immediate and reminder responders. It was concluded that there was evidence of covariate dependent response by reminder.

There was no significant difference in the previous PCS, MCS or Arthritis Index between immediate and reminder responders ($p>0.05$). At year one, marital status and whether or not the patients parents were still alive were predictors of reminder response. For the two year follow up, significant covariates were whether the patients had brothers or sisters and whether the patient had had any previous vertebral fractures.

5.6.6 Summary

Scenario one

Little's test found evidence against the MCAR assumption for both the MCS and the Arthritis Index. The LS test found no evidence in favour of MAR for the four QoL scores for those patients displaying a monotone missing data pattern. Ridout logistic regression suggested MAR data after baseline, but MCAR thereafter. Fairclough logistic regression suggested missingness was MAR at one and two years, but more likely covariate dependent at three and four years follow up.

Scenario two

Little's test of MCAR found evidence against MCAR for each of the QoL scores. Ridout logistic regression found that at baseline the current QoL was a significant

predictor of subsequent dropout for each QoL scores. The same conclusion was found at year one, except for the PCS and arthritis score which were found to be covariate-dependent drop out. At year two, QoL impacted on subsequent missing MCS, with the other three found to be covariate dependent. At year three, subsequent drop out was covariate dependent for all four QoL measures. The LS test found insufficient evidence against the MAR assumption. Fairclough logistic regression found evidence that previous QoL was impacting on non-response at year one to three suggesting MAR, with those displaying lower scores more likely to drop out.

Scenario three

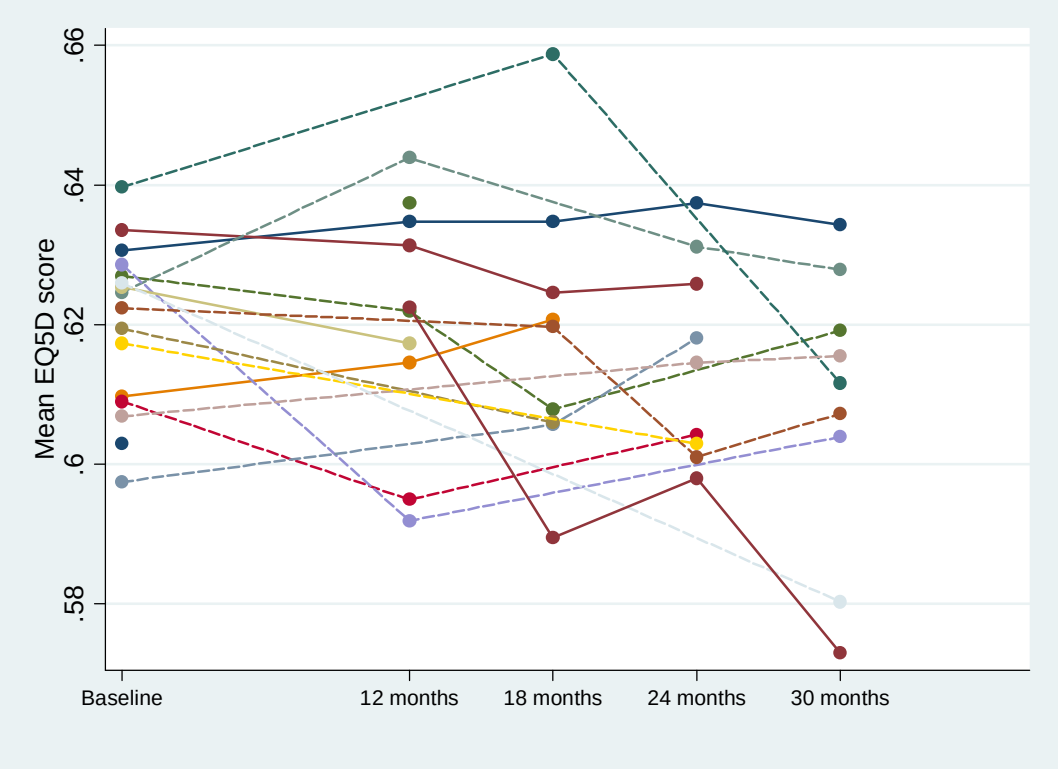
All four methods showed there was no apparent difference in QoL between the immediate and reminder responders. The mechanism of reminder response was found to be MCAR.

5.7 TOMBOLA

5.7.1 Pattern of missing data

Appendix 5.12 displays the number of participants and mean (SD) EQ5D score for the responders in TOMBOLA. Some additional data were collected at six weeks, but only a small subset of participants was issued with this questionnaire. Therefore, for the purposes of this chapter, only the remaining five QoL assessments (not six) were used, generating 32 possible missing data patterns. Figure 5.19 displays the mean EQ5D scores for the 18 patterns which were displayed by 10 or more participants. Despite the differing QoL at a particular time point between the patterns, the overall QoL is fairly stable among the participants who provide QoL scores. Those patients providing all assessments generally showed better QoL.

Figure 5.19: TOMBOLA – mean EQ5D score for each missing data pattern (≥ 10 patients per pattern)



5.7.2 Little's test

There were five assessments of EQ5D in TOMBOLA. In data scenario one, Little's test statistic was $X^2 = 86.2$ (p=0.11). For data scenario two, $X^2 = 73.8$ (p=0.30) and in data scenario three, $X^2 = 75.1$ (p=0.29). Hence, there was insufficient evidence to reject the null hypothesis of MCAR data in each case. Therefore, the observed QoL did not appear to be impacting on whether or not a patient responded and if they did respond, whether they needed a reminder.

5.7.3 Ridout Logistic regression

Scenario one

Table 5.16 shows the results of Ridout logistic regression for scenario one. At baseline, subsequent dropout was associated with the covariates age group, marital status, ethnicity, training, smoking and activity level. Baseline EQ5D scores were significantly different (p<0.001). This was still evident in the adjusted logistic model (p<0.001). The adjusted OR = 0.35 (0.20,0.63) suggested that those

who provided better baseline scores were less likely to drop out of the study and that drop out was MAR.

Table 5.16: TOMBOLA - Ridout logistic regression results (scenario one)

	Total	Dropouts	Unadjusted		Adjusted	
Assessment	N (%)	N (%)	OR (95% CI)	p-value	OR (95% CI)	p-value
Baseline	3300 (97)	809 (25)	0.37 (0.22, 0.64)	<0.001	0.35 (0.20, 0.63)[1]	<0.001
12 months	1725 (51)	241 (14)	0.39 (0.17, 0.92)	0.03	0.47 (0.18, 1.25)[2]	0.13
18 months	1543 (45)	281 (18)	0.69 (0.30, 1.60)	0.39	-	-
24 months	1492 (44)	461 (31)	0.49 (0.23, 1.07)	0.07	-	-

Adjusted for: [1] age group, smoking status, training, activity level, marital status and ethnicity; [2] age group, smoking status, marital status and employment status

At twelve months, age group, smoking status, marital status and employment status were related to dropout. Adjusting for these in the logistic model, the 12-month score was not significant (p=0.13). The 18 and 24 month QoL scores were not associated with dropout. Covariates associated with dropout were age group and ethnicity after 18 months and age group and smoking status after 24 months. Dropout was found to be covariate dependent at 12, 18 and 24 months.

Scenario two

The covariates associated with dropout were as follows: at baseline age group, smoking, marital status, employment and ethnic group; at 12 months, - age group, smoking status, employment status, marital status; at 18 months age group, smoking status, marital status and ethnic group; at 24 months - age group and smoking status. Table 5.17 describes the mean (SD) and unadjusted odds ratio between continuers and those that drop out. At baseline and at two years, there was evidence that those about to drop out showed lower QoL at that time point (p<0.001 and p=0.043 respectively).

Table 5.17: TOMBOLA - Ridout logistic regression results (scenario two)

	Total	Drop outs	Unadjusted		Adjusted	
Assessment	N (%)	N (%)	OR (95% CI)	p-value	OR (95% CI)	p-value
Baseline	3300 (97)	683 (21)	0.32 (0.18, 0.56)	<0.001	0.28 (0.15, 0.53)[1]	<0.001
12 months	2337 (69)	225 (10)	0.59 (0.24, 1.44)	0.25	-	-
18 months	2105 (62)	225 (11)	0.56 (0.23, 1.33)	0.19	-	-
24 months	2005 (59)	358 (18)	0.46 (0.21, 0.98)	0.043	0.26 (0.12, 0.59)[2]	0.001

Adjusted for: [1] age group, smoking status, employment status and ethnicity; [2] age group and smoking.

Having adjusted for the covariates at baseline and at 24 months, the current QoL score was a significant predictor of dropout after this time (Table 5.17). Those patients with lower QoL at baseline were more likely to drop out. At 24 months, those with lower QoL were less likely to provide the final assessment.

Scenario three

There were 1422 (42%) patients who provided all five assessments. The number continuously responding by reminder was limited. Dropout after a particular assessment was considered to be a patient who responded after reminder on at least one occasion on the remaining assessments. There were no covariates found to be associated with the indicator of subsequent reminder response at any of the assessments ($p>0.05$). The unadjusted odds ratios for dropout are found in Appendix 5.13. There was no evidence that current QoL was a predictor of dropout (or reminder response for at least one of the subsequent assessments). Therefore, there was no evidence against MCAR for reminder response.

5.7.4 The LS test

The LS test requires monotone missingness and 2254 (66%) showed this in scenario one with $S=2.63$ ($p=0.012$). Scenario two contained 2670 (79%) with a monotone pattern and $S=4.36$ ($p<0.001$). Both of these tests found evidence against the null hypothesis. In scenario three, there were 1422 patients providing all five assessments with 901 (63%) patients showing a monotone pattern (of reminder response). There was no evidence against the null hypothesis ($S=-0.81$, $p=0.29$) in this case and observed QoL was not found to be a predictor of the missingness mechanism (reminder response).

5.7.5 Fairclough logistic regression

Scenario one

In a stepwise logistic model the following baseline covariates were associated with the missing indicator: baseline – employment status and ethnicity; 12 months – age group, training, smoking status, marital status, employment status and ethnicity;

18-30 months – age group, smoking, marital status, employment status and ethnicity.

Inclusion of the previous QoL term (last known value) was significant at each of the follow up assessments. The adjusted OR (95% CI) were as follows: 12 months – OR = 0.31 (0.18, 0.54), p<0.001; 18 months – OR = 0.55 (0.32, 0.93), p=0.025; 24 months - OR = 0.48 (0.29, 0.82), p=0.006; 30 months – OR = 0.30 (0.17, 0.51), p<0.001. Therefore, at each of the follow up assessments there was evidence of MAR and that previously observed QoL was a predictor of missingness. Those patients displaying better previous QoL were less likely to provide missing assessments.

Scenario two

In a stepwise logistic model the following covariates were found to be significant: baseline – ethnic group and employment status; 12 months - age group, ethnic group, smoking status, marital status, employment status, training; 18-30 months - age group, ethnic group, smoking status, marital status, employment status. Having identified the relevant covariates, the significance of the previous QoL score was determined. The adjusted ORs (95% CI) were as follows: 12 months – OR = 0.27 (0.15, 0.49), p<0.001; 18 months - OR = 0.54 (0.32, 0.93), p=0.026; 24 months - OR = 0.39 (0.23, 0.65), p<0.001; 24 months – OR = 0.35 (0.21, 0.60), p<0.001. Those patients with higher QoL scores were less likely to drop out.

Scenario three

Age group, ethnicity and smoking status were found to be associated with responder type (immediate or reminder). In addition to these covariates, previous QoL was not significant in the model (p>0.05). There was some evidence at 18 months that current QoL was significant in the model (p=0.046) and that reminder response was found to be covariate dependent.

5.7.6 Summary

Scenario one

Little's hypothesis test found no evidence against the MCAR assumption. The LS test on the monotone patterned data suggested MAR. Ridout logistic regression

suggested drop out after baseline was MAR, but dropout after any follow up assessment was more likely MCAR (covariate-dependent). Finally, Fairclough logistic regression suggested missingness at each follow up assessment was MAR with those displaying poorer observed QoL more likely to provide missing responses.

Scenario two

Little's test found no evidence against the MCAR assumption. The LS test however, found some evidence in favour of MAR. Ridout logistic regression found evidence of MAR at baseline and at two years, but MCAR (covariate-dependent) at the remaining time points. Fairclough logistic regression found that previous QoL was a significant predictor in the missingness model and therefore, MAR was likely.

Scenario three

The mechanism of reminder response was found to be MCAR by Little's test. Fairclough logistic regression found evidence of MAR, but a possibility of MNAR at 18 months. Ridout regression found no evidence against the MCAR assumption, as neither covariates nor current QoL were found to be predictors of response by reminder. The LS test found no evidence in favour of MAR, suggesting observed QoL was not a predictor of reminder response.

5.8 The NPC Trial

5.8.1 Pattern of missing data

Appendix 5.14 displays the number of participants and mean (SD) scores for the pain, physical functioning and emotional functioning dimensions of the QLQ-C30 instrument. This information is presented pictorially in Figure 5.20 (pain), Figure 5.21 (physical functioning) and Figure 5.22 (emotional functioning). Across the majority of patterns the pain score decreases (less pain) or remained stable with time. This probably reflected that those who remained in the study for longer had less pain than those who dropped out. Those patients who dropped out after baseline (n=155) had the worst pain at baseline. Physical functioning (PF) was

reasonably stable within a particular pattern but tended to decrease in those patterns which contained patients dropping out early. Those patients who dropped out after baseline displayed the worst baseline physical functioning scores. The same phenomenon was seen with the emotional functioning (EF) scores.

Figure 5.20: NPC trial - pain mean score at each assessment by missing data pattern

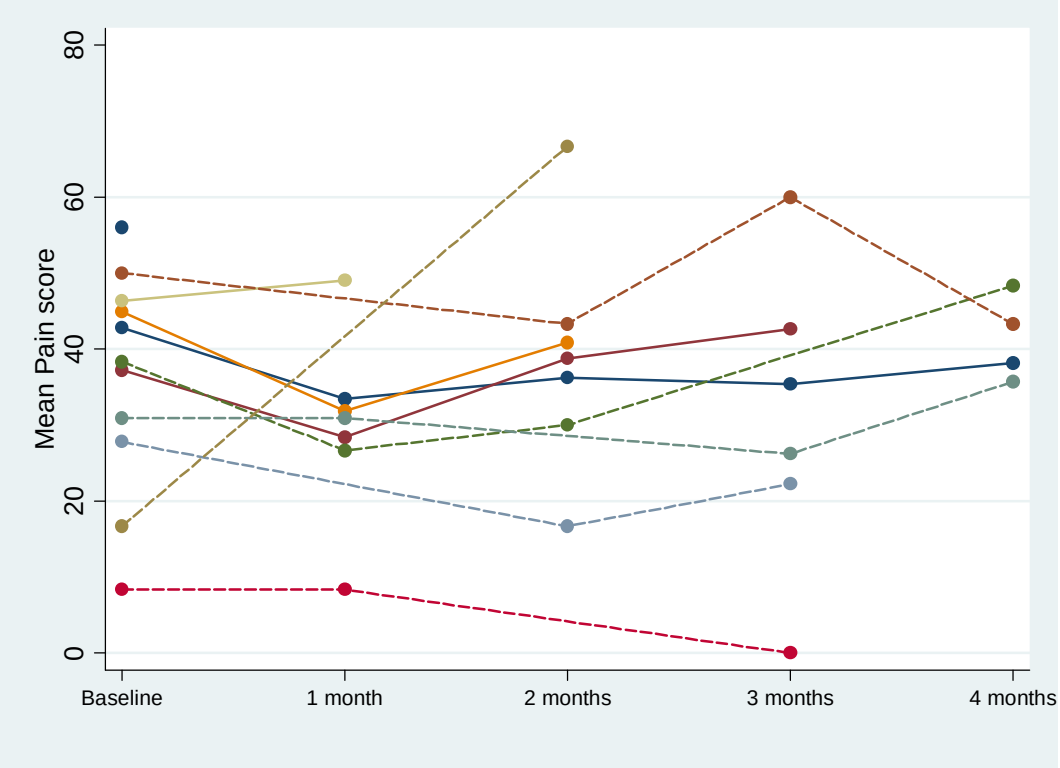


Figure 5.21: NPC trial - physical functioning mean score at each assessment by missing data pattern

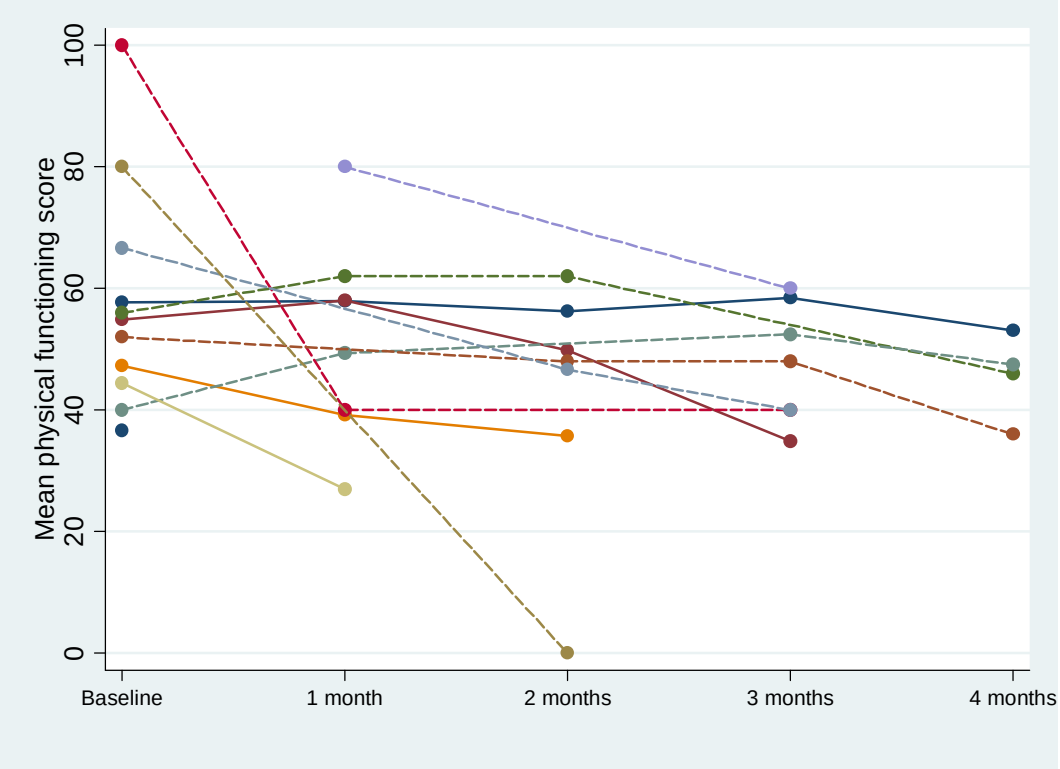
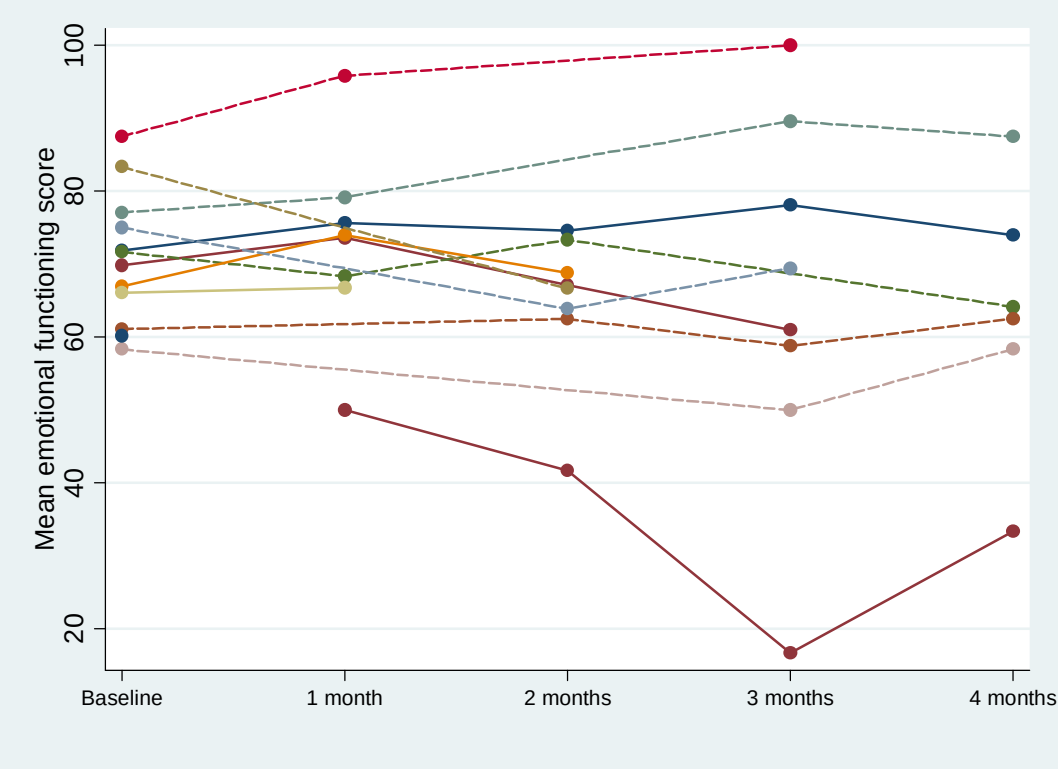


Figure 5.22: NPC trial - emotional functioning mean score at each assessment by missing data pattern



5.8.2 Little's test

There were 434 patients in the dataset, with 114 (26%) responding at each of the five assessments (baseline, 1m, 2m, ,3m and 4m). Just over a third (36%) of patients responded at baseline but provided no further assessments. Table 5.18 shows the results for Little's test for each of the three identified dimensions of the QLQ-C30 and for each data scenario.

Table 5.18: NPC Trial - Little's test p-values

QLQ-C30 Dimension	Scenario One	Two	Three
Pain	0.20	0.005	0.68
Physical functioning	<0.001	<0.001	0.67
Emotional functioning	0.045	<0.001	0.20

It can be seen that there was evidence against the null hypothesis of MCAR for the physical and emotional functioning scores in scenario one. In data scenario two, there was evidence against MCAR for each of the three QoL dimensions. A restriction to responders in scenario three saw a change in the conclusion (Table 5.18). In this instance, there was no evidence against the null hypothesis of MCAR and response after reminder could be regarded as completely at random.

5.8.3 Ridout logistic regression

Scenario one

Appendix 5.15 shows the results of Ridout logistic regression. At baseline, the covariates associated with dropout were the Karnofsky performance index and the cluster pair. At one month, sex was associated with dropout and at three months the Karnofsky index was associated with dropout. At two months no covariates were associated with dropout. Dropout was found to be MCAR at two and three months for the pain score and at two months for the emotional functioning score. QoL dimension scores were associated with dropout at all other times. This was still the case after adjusting for the identified covariates. Therefore, in this instance dropout was found to be MAR.

Scenario two

At baseline, Karnofsky performance status was found to be associated with dropout. Those patients dropping out displayed lower scores (worse performance status). No covariates were associated with dropout after the other assessments. Appendix 5.16 shows the mean (SD) QoL dimension scores of continuers and dropouts after each of the assessments and the odds ratio for dropout.

There was a significant difference in the pain scores of dropouts and continuers at both baseline and one month. Those dropping out were displaying worse pain (higher scores). Although pain scores were higher for dropouts at two and three years, the difference was not significant. The physical functioning scores were significantly different at each assessment, with those dropping out showing significantly worse physical functioning. The same was found with the emotional functioning scores, where apart from two months, patients with poorer emotional functioning were more likely to drop out. Therefore on the whole, Ridout regression provides evidence that observed QoL (pain, physical and emotional functioning) was an important predictor or drop out suggesting the data are MAR.

Scenario three

Data was restricted to include only those who responded at each assessment. There were 114 (26%) patients with pain scores, 112 (26%) with physical functioning and 107 (25%) with emotional functioning scores. Data collected by reminder was set to missing. In this scenario, there was insufficient data to employ Ridout's logistic regression method at baseline, one month and two months follow up. The process was possible at three months to assess the mechanism behind dropout thereafter (reminder response at four months). There were 97 patients with three month pain scores of which 19 (20%) subsequently dropped out. The three month score was not a significant predictor of dropout (OR = 0.985 (0.97, 1.00), p=0.10).

For the physical functioning scores at three months, 95 (22%) patients responded and 19 (20%) dropped out thereafter. As with the pain scores, the physical functioning scores were not a predictor of dropout (OR=0.99 (0.98, 1.01), p=0.47).

There were 92 patients who provided three month emotional functioning scores of which 18 (20%) dropped out thereafter. Dropout (reminder response) after three months was found to be MCAR for each of the three QoL dimensions. There was insufficient data available to investigate dropout earlier on.

5.8.4 The LS test

Scenario one

The LS test was not statistically significant for the pain score (p=0.26) and for the emotional functioning score (p=0.11). This suggested there was no evidence in favour of MAR. The LS test was significant for the physical functioning score (p=0.013) providing evidence in favour of MAR and suggesting that observed QoL was impacting on missingness.

Scenario two

The LS test required monotone missingness. In all three cases, the test was statistically significant with S=-2.53 (p=0.016) for the pain dimension, S = 5.31 (p<0.001) for physical functioning and S=3.23 (p=0.002) for emotional functioning. Therefore, there was evidence in favour of the MAR assumption that observed QoL was a predictor of missingness.

Scenario three

There was insufficient data at month three and four to carry out the LS test. Therefore, the data was restricted to consider the first three assessments only (baseline, one month and two months). There were 134 patients who displayed a monotone pattern for the pain and physical functioning scores at each of the first three assessments. This reduced to 131 patients for the emotional functioning score. In each case the LS test was not significant and there was no evidence in favour of MAR. The LS test statistic was S=0.09 (p=0.40) for the pain scores; S=1.15 (p=0.21) for the physical functioning scores and S=1.59 (p=0.11) for the emotional functioning scores.

5.8.5 Fairclough Logistic Regression

Scenario one

Table 5.19 shows the results of the Fairclough logistic regression process on data scenario one.

Table 5.19: NPC Trial - Fairclough logistic regression results (scenario one)

Assessment	Complete N (%)	Missing N (%)	Adjusted[1] OR(95% CI)	p-value
Pain				
1 month	203 (47)	231 (53)	1.00 (0.99, 1.01)	0.35
2 months	149 (34)	285 (66)	1.01 (1.00, 1.02)	0.001
3 months	130 (30)	304 (70)	1.009 (1.003, 1.016)	0.004
4 months	98 (23)	336 (77)	1.007 (1.00, 1.014)	0.04
Physical Functioning (PF)				
1 month	203 (47)	231 (53)	0.99 (0.98, 1.00)	0.022
2 months	148 (34)	286 (66)	0.98 (0.97, 0.99)	<0.001
3 months	130 (30)	304 (70)	0.986 (0.98, 0.99)	0.001
4 months	97 (22)	337 (78)	0.98 (0.97, 0.99)	<0.001
Emotional Functioning (EF)				
1 month	201 (46)	233 (64)	0.99 (0.98, 1.00)	0.39
2 months	148 (34)	286 (66)	0.98 (0.97, 0.99)	<0.001
3 months	130 (30)	304 (70)	0.986 (0.978, 0.995)	0.002
4 months	96 (22)	338 (78)	0.977 (0.966, 0.987)	<0.001

[1] Adjusted for Karnofsky performance status {<70, >=70}; PF – physical functioning; EF – emotional functioning

The Karnofsky performance status was associated with missingness in each case. Adjusting for the Karnofsky index, the previous QoL term was a significant predictor of missingness in the model in all but two cases (pain and emotional functioning at one month). On the whole, previous QoL was found to be associated with missingness indicating MAR. Patients with less pain, better physical and emotional functioning were less likely to provide missing response.

Scenario two

The Karnofsky performance status was associated with non-response at each month of follow-up ($p<0.001$ at month 1 and 2, $p=0.001$ at month 3 and 4). Those patients showing lower scores (poorer performance) were less likely to be responders. Appendix 5.17 shows the results of the Fairclough logistic regression process. In summary, the baseline pain scores differed between responder groups at each follow up. At two and three months, the one month pain score differed between responders and non-responders ($p<0.001$ and $p=0.01$ respectively). At

four months, other than baseline score, no other pain scores differed between responders and non-responders.

In a logistic model for response at one month (in addition to the Karnofsky score) the baseline pain score was significant (p=0.008). At two months, in addition to Karnofsky score, the one month pain score was significant in modelling non-response (p=0.002), while the baseline score was not (p=0.10). Similarly at three months, in addition to the Karnofsky score the one month pain score was required in the model (p=0.018), but baseline pain score was not (p=0.70). Finally, at four months, in addition to the Karnofsky status, the baseline pain score was significant in the model (p=0.025). Therefore, at each follow up assessment, missingness was related to both Karnofsky status and a previous pain score, implying missingness was MAR.

The baseline PF scores differed between responder groups at each follow up (Appendix 5.17). In addition to the Karnofsky score, the baseline PF score was significant in a logistic model for response at one month (p=0.002). Similarly at two and three months, the one month PF score was a significant predictor of non-response, in addition to the Karnofsky score (both p<0.001). Finally at four months, in addition to the Karnofsky status, the baseline PF score was important in the model (p=0.003). Therefore, at each of the follow up assessments, missingness was related to both Karnofsky status and a previous PF score, implying missingness was at least MAR.

The baseline EF scores differed between responder groups at follow up (Appendix 5.17). The one month EF scores did not differ between responder types at two, three or four months (all p>0.05). The three month EF score differed between responders and non-responders at four months (p=0.005). In a logistic model, the baseline score was important in addition to the Karnofsky performance status at one month (p<0.001), but not at two months (p=0.42) or three months (p=0.2). At four months, both the three month score (p=0.001) and baseline score (p=0.04) proved significant in the model.

Scenario three

For pain, the baseline score (p=0.012), and two month (p=0.04) scores differed between the types of responders (immediate or reminder) at four months. No other previous scores differed. Using the feature of the reminder data, the current scores could be tested as to whether they predicted non-response. None were found to be significant. Therefore, in relation to the pain scores, on the whole there was no difference in the pain levels between the immediate and reminder responders.

There was no difference in the previous or current PF scores between the two types of responders at the four follow up times (p>0.05 in all cases). Baseline EF scores differed between the type of responder at one month (p=0.026) and at two months (p=0.036). At one month, the current score also differed between immediate and reminder responders. In a logistic model for reminder response at one month, in addition to the one month score, the baseline score was not required (p=0.40), suggesting that the current EF was better at predicting the reminder response.

5.8.6 Summary

Scenario one

Little's test of MCAR found evidence against the null hypothesis for the emotional and physical functioning scores. There was no evidence against MCAR for the pain scores. The LS test found in favour of MAR for the physical functioning scores but not for the missing emotional functioning and pain scores. Both Ridout and Fairclough logistic regression found evidence of MAR data in most cases.

Scenario two

There was evidence against the MCAR assumption for Little's test of all three QLQ-C30 dimensions. Ridout logistic regression provided evidence that observed QoL (pain, physical and emotional functioning) was an important predictor of drop out, suggesting the data were MAR. The LS test also provided evidence of MAR, as did Fairclough's logistic regression approach.

Scenario three

The investigation into the mechanism behind reminder response found no evidence against MCAR from Little's test. Fairclough logistic regression found evidence of MAR. For reminder response at one month, the current emotional functioning score was found to be significant, suggesting the possibility of MNAR. Ridout logistic regression found dropout after three months to be MCAR for each of the tree QoL dimensions. The LS test found no evidence in favour of MAR.

5.9 Discussion

The initial stage of investigating the missing data mechanism is to look at the missing data pattern. The figures provided here displayed the mean QoL score at each assessment split by missing data pattern. In general, this showed that the QoL was better among the participants who provided all assessments or those that only missed the baseline assessment. Those participants who provided only one assessment tended to show poor QoL at this assessment. This pattern suggests that QoL is impacting on whether or not participants respond and that poor observed responses are indicative of missingness at a later assessment.

Chapter four identified four methods to investigate the missing data mechanism for the missing QoL data in the trial datasets. The two hypothesis tests (Listing, Schlittgen 1998, Little 1988) gave an idea of the overall mechanism of missing responses while the two logistic regression procedures (Fairclough 2002, Ridout 1991) looked specifically at the mechanism of missingness at a particular assessment. The advantage of Little's test over the LS test is that Little's test allows for intermittent missingness pattern. The LS test requires monotone missingness which will often not be the case. Using the LS test in this situation will discard some of the observed data.

Table 5.20: Summary of the missingness mechanism from the two hypothesis tests

	LITTLE'S TEST			LS Test		
Scenario	**ONE**	**TWO**	**THREE**	**ONE**	**TWO**	**THREE**
REFLUX						
EQ5D	MAR			MAR	not MAR	MAR
PCS	MCAR			not MAR		
MCS	not MCAR	MCAR	not MCAR	MAR	not MAR	MAR
RQLS	MCAR			not MAR		
MAVIS						
EQ5D	MCAR	not MCAR	MCAR	MAR		not MAR
PCS	not MCAR					
MCS						
RECORD						
EQ5D	not MCAR			MAR		
PCS						
MCS						
KAT						
EQ5D	not MCAR			MAR		
PCS						
MCS						
OKS						
PRISM						
EQ5D	MCAR	not MCAR	MCAR	not MAR		
PCS	MCAR/MAR					
MCS	not MCAR					
ARTH	not MCAR					
TOMBOLA						
EQ5D	MCAR	MCAR	MCAR	MAR	MAR	not MAR
NPC Trial						
PAIN	MCAR	not MCAR	MCAR	not MAR	MAR	not MAR
PF	not MCAR			MAR		
EF	not MCAR			not MAR		

PF – physical functioning; EF – emotional functioning; PCS – physical component score; MCS – mental component score

Ridout logistic regression models dropout after an observed assessment, while Fairclough logistic regression models missing response at a particular assessment. The approaches are similar in terms of identifying suitable covariates and then assessing the inclusion of QoL terms in the model, but the dependent variable is subtly different. If the mechanism of missingness at a particular assessment is of interest, then one of the two logistic regression procedures is useful. Despite the differences between methods, for a particular scenario/QoL dimension combination within a trial, all four methods gave consistent conclusions. A summary of these conclusions is found in Table 5.20 for the two hypothesis tests and in Table 5.21 for the two logistic regression procedures.

Table 5.21: Summary of the missingness mechanism from the logistic regression methods

	Ridout logistic regression			Fairclough logistic regression		
QoL score	**ONE**	**TWO**	**THREE**	**ONE**	**TWO**	**THREE**
REFLUX						
EQ5D	MAR at baseline MCAR at 3m	MCAR	MAR at baseline MCAR at 3m	MAR	MCAR at 3m	MNAR
PCS	MCAR	MCAR	MCAR	MCAR		MNAR
MCS	MAR at baseline MCAR at 3m	MCAR	MCAR	MAR		MNAR
RQLS	MCAR	MCAR	MCAR	MAR		MAR
MAVIS						
EQ5D	MCAR	MAR at baseline MCAR at 6m	MCAR	MAR at 6m MCAR at 12m	MCAR at 6m MAR at 12m	MCAR
PCS	MCAR		MCAR	MAR at 6m MCAR at 12m	MCAR	MCAR
MCS	MCAR		MCAR		MCAR	MCAR
RECORD						
EQ5D	MAR			MAR		MNAR
PCS						
MCS						
KAT						
EQ5D	MAR	MCAR at baseline MAR at 3m/1y	MAR	MAR	MCAR at baseline MAR at 3m/1y	MAR at 2y MNAR at 3m/1y
PCS	MCAR at baseline MAR at 3m/1y	MCAR at baseline MAR at 3m/1y	MCAR			MAR at 2y MNAR at 3m/1y
MCS	MAR	MAR	MCAR			MAR at 3m/2y MNAR at 1y
OKS	MCAR at 3m MAR at baseline/1y	MCAR at baseline/3m MAR at 1y	MCAR			MAR at 3m/1y MNAR at 2y
PRISM						
EQ5D	MAR at baseline MCAR at 1y/2y/3y	MAR at baseline/1y MCAR at 2y/3y	MCAR	MAR at 1y/2y MCAR at 3y	MAR	MCAR
PCS		MAR at baseline MCAR at 1y/2y/3y		MAR at 1y MCAR at 2y/3y	MAR at 1y/2y MCAR at 3y	
MCS		MAR at baseline/1y/2y MCAR at 3y		MAR at 1y/2y/3y		
ARTH		MAR at baseline MCAR at 1y/2y/3y		MAR at 1y/2y MCAR at 3y		

TOMBOLA					
EQ5D	MAR at baseline MCAR thereafter	MAR at baseline/2y MCAR elsewhere	MCAR	MAR	
NPC Trial					
Pain	MAR at baseline/1m MCAR at 2m/3m/4m		MCAR at 3m	MAR	MCAR
PF	MAR		MCAR at 3m		
EF	MAR		MCAR at 3m		

PF – physical functioning; EF – emotional functioning; PCS – physical component score; MCS – mental component score

Scenario two represents the situation of most clinical trials and the platform on which researchers would base their investigation. On the whole, missingness was found to be MAR (associated with covariates and observed QoL) in the trial datasets presented here. Participants who had shown lower observed QoL were more likely to provide missing responses. Ignoring this finding in analysis could potentially bias the results. Patients displaying lower QoL at other assessments are liable to display lower QoL at the missing assessment. If this missing data are ignored, then it is possible that calculated means on the observed sample are inflated and not reflective of the true mean.

Scenario one looked at the mechanism of non-response where response data included only that collected without the need for reminder. In scenario two, the data collected through reminders was included as part of the response data. The mechanism of missingness identified was not always the same in scenario one and two. This suggested that the reminder data had an important role to play. In a trial which does not employ a reminder system, only the data collected immediately would be available. If the investigation into the missingness mechanism was based on only this data, then one could potentially get a distorted view of the missing data mechanism. Obtaining as much data as possible in any setting is always going to give a more informed decision and ultimately reduce any potential bias in analysis results.

There were some situations where the mechanism of missing data differed between QoL measures within a particular trial, but since the outcomes would be

analysed separately this is not of too much concern. The problem lies when the mechanism of missing data differed between two QoL dimensions within a single instrument. For example, in REFLUX, missing physical summary scores from the SF12 instrument were found to be MCAR while mental summary scores were more likely MAR. This actually is an interesting finding and suggests that the mental functioning of the participants was more a factor in non-response rather than the physical functioning. For instruments that contain more than one dimension, the investigation into the missing data mechanism may indicate which particular QoL construct is potentially informing missingness rather than QoL as a whole.

The advantage of scenario three is that the current (observed) QoL was known. This allowed an investigation as to whether current scores were different between immediate and reminder responders having adjusted for the covariates that were predictors of reminder-response. In several situations, this was shown to be MNAR suggesting that the missingness was informative. The use of reminder data allowed this conclusion. In the more usual setting this is not possible, as the data required are missing.

In any study that contains missing data, the missingness mechanism should be identified in advance of any analysis so that the most appropriate methods can be identified (Curran et al. 1998a). Chapter two showed that in the majority of reported clinical trials, there was no formal discussion about reasons for missingness and no investigation into the mechanism of missingness. In the unlikely situation that data can be confirmed as being MCAR, complete case analysis or simple methods of imputation could be used. In the more likely situation of MAR data, multiple imputation is useful (Carpenter, Kenward 2007). An alternative could be to use available case analysis and in the longitudinal setting a repeated measures model could be appropriate. When data is thought likely to be MNAR, more sophisticated approaches such as joint modelling or pattern mixtures models should be used (Fairclough 2002).

5.9.1 Conclusion

The conclusion from the work presented in this chapter is that, depending on the dataset and scenario, the mechanism of missing data differed. In a typical trial situation, before any analysis and/or imputation is carried out, the mechanism of missing data should be identified. For an overall view, Little's test should be applied to scenario two as this test allows for all missing data patterns (both monotone and intermittent). If the mechanism behind missingness at a particular endpoint is required then one of the two logistic regression procedures can be used. These approaches can also be used to identify the mechanism behind reminder-response; which may help in identifying the mechanism behind non-response. The rationale behind this suggestion is that the reminder-responders are perhaps more representative of the true non-responders, than the immediate-responders are.

As discussed above the missingness mechanism is important in determining the most appropriate method of imputation or analysis strategy. Without the reminder-responses, you may get a distorted view of the missing data mechanism which may lead to inappropriate analyses. The results of this chapter will be used later in this thesis in the application of the different ways of dealing with missing data and why they may or may not be appropriate for each of the trials. Chapter six explains the theory behind simple and multiple imputation, while chapter seven applies these methods to the seven trial datasets. Chapter eight provides theory for alternative model-based strategies and chapter nine applies these methods to the data.

Chapter 6 Methods of imputation for missing data

6.1 Introduction

One way to deal with missing data is imputation, whereby a reasonable alternative value replaces one that is missing. The process of imputation is based on completion, rather than deletion. Imputation can be used in one of two ways: to impute missing forms (unit non-response) or to impute missing items (item non-response). The first occurs when the whole QoL assessment is missing. The second occurs when a questionnaire has only been partially completed with one or more questions unanswered. A number of QoL instruments have built-in methodology for imputing items from multi-item scales. For example, in the EORTC QLQ-C30, if at least half the items from a scale have been answered, it is assumed that the missing items have values equal to the average of those items, which are present for that respondent (Aaronson et al. 1993, Fayers et al. 2001). This half-item rule has been validated in a number of questionnaires and is implemented for the QLQ-C30 and SF12/36 used in this thesis.

There are a number of methods available to determine the "best guess", which is then substituted for each missing value. In a trial dataset, additional information about the patient is often provided. This information can be used to make an educated guess as to the most likely QoL value that would have been reported if the patient had returned the questionnaire. The process of imputation results in a complete sample, with the missing observations estimated by the imputed values. This allows standard analysis approaches to be carried out. However, it should be noted that imputation cannot completely replace lost information. Although a seemingly complete dataset is created, it is one that has been augmented in some sense. Imputation is a device to facilitate analysis and it is not a substitute for real data (Fayers, Machin 2001). Imputation can be attractive when compared to model-based methods. Imputation is reasonably straightforward to implement and the results are relatively easy to interpret. The main focus of this chapter is the imputation of the whole QoL dimension score, rather than the individual items

that comprise the score. Different approaches to imputation will be discussed. Chapter seven will apply these approaches to the example trial datasets.

6.2 Simple imputation

Simple imputation is the process when a single alternative value is imputed for a missing value. These procedures come in two broad categories: longitudinal and cross sectional methods. Longitudinal methods utilise data observed at other assessments and which are specific to the patient. Cross-sectional methods make use of observed data at the given assessment, including data from other patients. Both types can incorporate other information such as sex, age or any other covariate included in the dataset. Various methods of simple imputation are described below.

6.2.1 Last value carried forward

As the name suggests, last value carried forwards (LVCF) (often denoted LOCF – last observation carried forward) is the last known QoL assessment carried forward to one that is missing. For example, if a patient completes the first and second assessment, but does not respond to the third, then the value obtained at the second assessment is carried forwards and imputed for the third (missing) assessment. The major disadvantage of LVCF is that it assumes QoL is remaining constant over time, which is perhaps not realistic. For example, in a palliative care setting, QoL is known to be decreasing with time, so carrying forward a previously observed value is likely to give an inflated view of QoL at the current missing assessment.

6.2.2 Baseline carried forward

Baseline carried forward (BCF) is similar to LVCF, but the baseline value is carried forward rather than the last observed value. This method has the same disadvantage as LVCF. There is the assumption that QoL is unchanged from baseline. Depending on the characteristics of the study group, this may give an

inflated view of their QoL (perhaps in a palliative care setting where QoL is likely decreasing) or a conservative view (where treatment is deemed to be improving QoL). This method is equivalent to LVCF when imputing data at the second assessment, or when the only observed score, is the baseline score.

6.2.3 Next value carried backwards

Next value carried backwards (NVCB) works on the same principle as LVCF and BCF, but carries backwards the next known value. Clearly this method is only applicable when future values are known. It may be of limited use, especially when it is the final QoL assessment that is missing. As with LVCF, the method assumes that QoL is stable over time and specifically that QoL in the future is the same as the QoL at the assessment in question.

6.2.4 Horizontal mean imputation

In a study where QoL is collected across several assessments, the mean of the observed assessments for each patient can be imputed for a missing assessment. This method will give the same imputed value as LVCF if there is only one other assessment, or if there is no change in the observed QoL assessments. The same reservations that exist for LVCF also apply here. Horizontal mean imputation is not recommended if there is any evidence of a systematic decline or rise in QoL scores over time (Fayers, Machin 2007).

6.2.5 Simple mean imputation

The mean score of those participants who completed a particular QoL assessment is obtained. This value is then imputed for all those participants who provided missing response at that assessment. This will cause the mean in the augmented complete dataset to be unchanged, but the standard error will be artificially reduced. This can lead to distorted significant tests and falsely narrow confidence intervals (Fayers, Machin 2001; Fairclough 2002). The correlation between different scores may be affected by the imputation (Fayers, Machin 2007). Mean

imputation can be carried out using a stratified procedure. The mean of patients with similar characteristics is used rather than the mean of the whole sample.

6.2.6 Maximum/minimum value imputation

As the name suggests, maximum (or minimum) value imputation substitutes the maximum (or minimum) value observed. This may be carried out in either a longitudinal or cross-sectional manner. In the longitudinal approach, the maximum (minimum) observed value for a particular patient is imputed for each of their missing values. In the cross-sectional approach, the maximum (or minimum) value of the sample is imputed for each of the missing observations. In the latter, the problem of reduced standard errors that was discussed for simple mean imputation also applies.

6.2.7 Hot deck imputation

The hot deck procedure selects at random from patients with observed QoL data. The QoL score of one of these patients is then imputed for one that has missing assessment. It is possible to restrict the set of observed responses to patients who match on particular characteristics (e.g. gender, age group, treatment group). This provides a stratified hot-deck procedure.

6.2.8 Regression

A regression model can be determined to provide predicted values for imputation. An advantage of this method is that, it can incorporate other information from the same subject or other subjects. This could include both QoL scores and other covariates/patient characteristics. Candidate variables for inclusion are those that are either associated with the missing indicator, or are strong predictors of the QoL outcome. Regression analyses are then carried out to determine the best model for imputation.

6.2.9 Summary

The methods described above carry out imputation in a longitudinal (horizontal) or cross-sectional (vertical) manner. An advantage is that they are easy to implement. Simple imputation does not require any specialist software and the routines can be programmed into any standard statistical software. Once imputation has occurred, the values are treated as equivalent to those which have been observed. This can be problematic. A major disadvantage of simple imputation method is that variances/standard deviations are underestimated. This has consequences for computing standard errors, subsequent test statistics and confidence intervals (Fairclough 2002). Simple imputation methods are likely to be biased unless the mechanism is known to be MCAR, in which case methods such as mean imputation may be appropriate. Therefore, it is imperative to consider the missing data mechanism when deciding on an appropriate imputation procedure (Fairclough 2002).

6.3 Multiple Imputation

6.3.1 What is multiple imputation?

Simple imputation can create more problems that it solves, distorting estimates, standard errors and hypothesis tests. Multiple imputation (MI) can overcome some of these problems. Instead of filling in a single value for each missing value, MI replaces each missing value with a set of plausible values that represent the uncertainty about the imputed value. MI does not attempt to estimate each missing value through simulated values, but instead draws a random sample of the missing values from its distribution. This process results in valid statistical inferences, which properly reflect the uncertainty surrounding the missing values.

The basic strategy is to impute multiple (M) values (say 5) for the missing data, incorporating both the variability of the QoL measure and the uncertainty about the missing observations. The initial model for imputation could be a regression model or can be based on the hotdeck procedure. The imputation process is carried out M times, generating M complete datasets. Each data set can then be

analysed in the usual way, resulting in M analyses producing, for example, M regression coefficients. These are then combined to give an overall summary estimate and significance test.

Rubin and Schenker presented this method of combining results from a data analysis performed 'M' times, once for each of 'M' imputed data sets, to obtain a single set of results (Rubin, Schenker 1991). From each analysis, the estimates and their standard errors need to be calculated and saved. Suppose that E_j is an estimate of interest (e.g. regression coefficient) obtained from data set $j=1,\ldots,m$ and S_j is the standard error associated with E_j. The overall estimate is the average of the individual estimates. For the overall standard error, one must first calculate the within-imputation variance

$$\bar{S} = \frac{1}{m}\sum_{j=1}^{m} S_j \,,$$

and the between imputation variance,

$$B = \frac{1}{m-1}\sum_{j=1}^{m}\left(E_j - \bar{E}\right)^2 .$$

The total variance, V is then

$$V = \bar{S} + \left(1+\frac{1}{m}\right)B .$$

The overall standard error is equal to the square root of V. The degrees of freedom are given by

$$df = (m-1)\left(1+\frac{m\bar{S}}{(m+1)B}\right)^2 .$$

A significance test of the null hypothesis $E=0$ is performed comparing the ratio $t = \bar{E}/\sqrt{V}$ to the appropriate t-distribution (Rubin, Schenker 1991).

If the proportion of missing data is quite small, then single imputation may be quite reasonable. Without special corrective measures, single imputation inferences tend to overstate precision because they omit the between imputation component of variability. MI does not require or assume that non-response is ignorable. Imputations may in principle be created under any kind of

assumptions or model for the missing data mechanism and the resulting inferences will be valid under that mechanism.

6.3.2 Software for multiple imputation

There have been recent developments in the availability of built-in procedures for MI in existing statistical software packages. STATA undertakes MI by chained equations using the ICE command (Royston 2004; Royston 2005). SAS undertakes MI using the procedure PROC MI to carry out the imputation, followed by PROC MIANALYZE to combine the results (SAS Institute Inc. 2004). A specific software package, SOLAS (Statistical Solutions Inc, Sargus, MA, USA) has been developed to handle missing data and perform MI. The chosen software package for this work was SAS. The procedures are described in the next section.

6.3.3 Multiple imputation in SAS

Multiple imputation can be carried out using PROC MI and the results combined using PROC MIANALYZE within the statistical software package SAS (SAS Institute Inc. 2004). Several MI methods are available and their appropriateness depends on the type of outcome variable (continuous, binary or categorical) as well as the missing data pattern (monotone or intermittent). The QoL outcomes contained within this work are all continuous. The details of the multiple imputation methods available are provided below. Further details on the methods for categorical outcomes can be found in the SAS Help and Documentation (SAS Institute Inc. 2004).

Two parametric methods are available, which assume a multivariate normal distribution for a continuous variable. They are the regression method (Rubin 1987) and the predictive mean matching (PMM) method (Heitjan, Little 1991). Both these methods are based on modelling the QoL outcome. The non-parametric procedure available is the propensity score method. This method models the missing indicator, rather than the QoL outcome (Rubin 1987). All three methods require the data to be of a monotone missing data pattern. One

option available for intermittent missing data is the use of Markov Chain Monte Carlo (MCMC) imputation (Schafer 1997). Each of these methods is now described.

Regression method for monotone missingness

The regression method is the default imputation method within SAS for continuous variables in a data set with a monotone missing data pattern. In the regression method, a standard regression model is fitted for a continuous outcome (e.g. QoL score) with the covariates constructed from those available within the dataset. Based on the fitted regression model, a new regression model is simulated from the posterior predictive distribution of the parameters. This new model is used to impute the missing values for each variable (Rubin 1987). A formal explanation is now supplied.

The standard regression model is fitted to the observed data for variable Y_j and its covariates X_1, X_2,…, X_k such that

$$Y_j = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k .$$

The fitted model produces regression parameter estimates $\hat{\beta} = \left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k\right)$ and the associated covariance matrix $\hat{\sigma}_j^2 V_j$, where V_j is the usual X′X inverse matrix derived from the intercept and covariates. Next, the imputed values for each imputation need to be generated.

The following steps are used (SAS Institute Inc. 2004):

1. New parameters $\beta_* = \left(\beta_{*0}, \beta_{*1}, \ldots, \beta_{*0k}\right)$ and σ_{*j}^2 are drawn from the posterior predictive distribution of the parameters. Thus, they are simulated from $\left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k\right)$, σ_j^2 and V_j. The variance is drawn as $\sigma_{*j}^2 = \hat{\sigma}_j^2 \left(n_j - k - 1\right)/ g$ where g is a $\chi_{n_j-k-1}^2$ random variate and n_j is the number of non-missing observations for Y_j. The regression coefficients are drawn as $\beta_* = \hat{\beta} + \sigma_{*j} V_{hj}^{'} Z$ where $V_{hj}^{'}$ is the upper triangular matrix in the Cholesky decomposition, $V_j = V_{hj}^{'} V_{hj}$ and Z is a vector of k+1 independent random normal variates.
2. The missing values are then replaced by

$$\beta_{*0} + \beta_{*1}x_1 + \beta_{*2}x_2 + \ldots + \beta_{*k}x_k + z_i\sigma_{*j}$$

where $x_1, x_2, \ldots, x_k$ are the values of the covariates and z_i is a simulated normal deviate.

It is possible to use different regression models for each variable undergoing imputation.

Predictive mean matching for monotone missingness

The predictive mean match model also fits a regression model and obtains a set of predicted values. For each piece of missing data, a set of observed values for which the predicted values are closest to the predicted value of the missing observation are obtained. One of these observed values is randomly selected and imputed for the missing value. An advantage of this method over regression is that the imputed values are always within the range of the observed data. The process is as follows:

1. As for the regression method new parameters are drawn from the posterior predictive distribution.
2. For each missing value, a predicted value $y_{i*} = \beta_{*0} + \beta_{*1}x_1 + \beta_{*2}x_2 + \ldots + \beta_{*k}x_k$ is calculated from the covariate values $x_1, x_2, \ldots, x_k$.
3. A set of k_0 observations, whose corresponding predicted values are closest to y_{i*} is generated.
4. The missing value is replaced by a value randomly selected from the k_0 (default $k_0 = 5$) observed values.

The predictive mean matching method ensures the imputed values are plausible and may be more appropriate than regression if the normality assumption is violated (Fairclough 2002).

Propensity score method for monotone missingness

The propensity score method is an alternative non-parametric imputation procedure available for continuous variables, when the data set has a monotone missing pattern. For a variable with missing values, a score is generated for each

observation to estimate the probability that the observation is missing. This is called the propensity score. The observations are then grouped (usually five) according to these propensity scores. Within each group, a set of observed scores are randomly selected with replacement to create a new set of observed scores. This is known as the Approximate Bayesian Bootstrap (ABB) (Rubin 1987). The steps involved in the propensity score method are described below (SAS Institute Inc. 2004).

Firstly, create an indicator variable R_j with value 1 for observations missing and 0 otherwise. Fit the logistic regression model

$$\log it(p_j) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

where $x_1, x_2, \ldots, x_k$ are covariates for Y_j, $p_j = \Pr(R_j = 0 \mid x_1, x_2, \ldots, x_k)$

and

$$\log it(p) = \log\left(\frac{p}{1-p}\right).$$

Create a propensity score for each observation to estimate the probability that it is missing. Divide the observations into a fixed number of groups (usually 5) based on the propensity scores. Apply an ABB imputation to each group as follows: in group k, suppose Y_{obs} denotes the n_1 observations with non-missing Y_j values and Y_{mis} denotes the n_0 observations with missing Y_j. The ABB imputation first draws n_1 observations randomly with replacement from Y_{obs} to create a new data set Y_{obs}*. The process then draws the n_0 values for Y_{mis} randomly with replacement from Y_{obs}*. This method is effective for inference about the distributions of individual imputed variables, such as univariate analysis, but not appropriate for analyses involving relationship among variables, such as a regression (SAS Institute Inc. 2004).

Methods for arbitrary (intermittent) missingness

The MI methods discussed so far are only suitable for datasets with a monotone missing data pattern. An alternative that is useful if the data shows an intermittent pattern is the Markov Chain Monte Carlo (MCMC) method (Schafer 1997). MCMC is used to generate pseudo-random draws from multidimensional

probability distributions using Markov chains. A Markov chain is a sequence of random variables, in which the distribution of each element depends only on the value of the previous one. MCMC is applied as a method for exploring posterior distributions in Bayesian inference. MCMC can be used in two ways for arbitrary missingness. The first uses MCMC on all the missing data. The second uses MCMC to replace enough missing values to make the data monotone. This is followed by one of the other more flexible monotone methods (described above) for the remainder of the data. The MCMC method assumes the data are from a multivariate normal distribution.

MCMC simulates the joint distribution and obtains simulation-based estimates of posterior parameters. Assuming the data are multivariate normal, data augmentation can be applied to Bayesian inference with missing data by repeating the 'imputation' I-step and 'posterior' P-step. They are defined as follows:

1. The imputation I-step: Given an estimated mean and covariance matrix, the I-step simulates the missing values for each observation independently, drawing from the conditional distribution for $Y_{i(mis)} \mid Y_{i(obs)}$.
2. The posterior P-step: Given a complete sample, the P-step simulates the population mean vector and covariance matrix. These new estimates are used in the next I-step. Informative or non-informative priors can be used.

The two steps are iterated long enough for the results to be reliable (see (SAS Institute Inc. 2004, Schafer 1997)).

6.3.4 Choice of imputation model

Five imputed datasets (by default) are created using PROC MI. The MIANLAYZE procedure then uses Rubins rule's (Rubin, Schenker 1991) to produce a combined estimate, the corresponding between and within imputation variance, standard errors, confidence intervals and p-values. The formulation of the imputation model is an important step in all methods (Carpenter, Kenward 2007). Failure to accommodate the model structure appropriately can cause bias in the resulting

analysis (Fay 1992). Additional variables which are related to both missingness and outcome should also be included (Fairclough 2002). For each trial dataset, the covariate set to be used in the imputation model was identified using standard statistical procedures (e.g. t-tests and chi-squared tests). Those covariates which were significantly associated with both the outcome (QoL) and the indicator of reminder response were considered.

6.4 Assessing the accuracy of imputation

The unique feature of the datasets under consideration is the extra portion of data that was collected through reminders, which would have otherwise been missing. Using this data, allows the calculation of the accuracy of the imputation as the true value is known. Ultimately what is of most interest in the analysis of a clinical trial, is producing an unbiased estimate of treatment difference. It is of less concern if a bias exists at the patient level, when the main focus is the treatment difference. Chapter seven uses the data for the responders and deletes that which was obtained by reminder. The missing values are then imputed, using the approaches outlined earlier in this chapter. The ANCOVA presented in chapter three can be carried out on each of the imputed datasets. The accuracy of each imputation method is then assessed using a measure of bias and precision as outlined below.

The absolute bias in the calculated treatment difference estimate is

$$b = |\text{observed treatment difference} - \text{treatment difference under imputation}|.$$

The precision of the estimate is important when determining accuracy of the different methods. The width of the confidence interval for the observed result and that under imputation can be obtained. The ratio of the confidence interval widths can then be used as a measure of precision, where

$$\text{Ratio} = (95\%\ \text{CI width under imputation}) / (95\%\ \text{CI width of observed}).$$

Ideally the ratio would be equal to one, such that the observed precision is also seen under imputation. The 'best' imputation method can then be identified by considering both the measure of bias and precision. The ratio of the CI width was used as the measure of precision to aid the interpretation. The treatment difference is presented with the 95% CI to give an indication to the interval in which the true population value lies. A reasonable method of imputation would hope to maintain the width of this interval, and hence precision of this estimate.

6.5 Overview

This chapter has outlined some of the available simple and multiple imputation procedures. The major advantage that MI has over simple imputation is that it produces more appropriate standard errors which can lead to more appropriate confidence intervals (Fairclough 2002). A selection of these imputation procedures will be carried out on the example datasets in chapter seven. The process involved in this is discussed in detail there. As discussed above, utilising the data collected through reminders allows the calculation of the accuracy of the imputation as the missing values are in fact known. The advantage of this approach over previous authors is that the missing data are not simulated and are not missing subject to a pre-specified pattern. The performance of different imputation strategies can be assessed and to some extent be predicted based on the knowledge of the missing data mechanism found in chapter five. The aim is to identify suitable imputation procedures that could be used in future clinical trials, to aid analysis and provide as unbiased as possible estimates of treatment difference. This will ultimately improve the available evidence base and inform clinical practice.

Chapter 7 Investigating the accuracy and impact of imputation

7.1 Application to the datasets

7.1.1 Introduction

This chapter aims to determine the accuracy of a number of imputation strategies for missing QoL outcomes. The previous chapter outlined the various methods of imputation that are available. Those methods which are implemented in this thesis are shown in section 7.1.2. In this chapter, these methods will be applied to the example trial datasets. Chapter five investigated the missing data mechanism using the data collected by reminder. In the current chapter, the reminder data are used as a tool to enable the accuracy of different imputation procedures to be investigated. The reminder responses are regarded as missing, imputed, and the accuracy of imputation assessed, since the true value of the missing value is known. Hence, in the analysis that follows the trial datasets are restricted to those who responded at both baseline and the assessment of primary interest. Observed baseline scores were required as the analysis of covariance (ANCOVA) adjusted for baseline QoL.

Once these restricted datasets were obtained, the data collected by reminder at the primary endpoint was removed. This missing data was then imputed using one of the methods previously described (chapter six). After imputation a complete augmented dataset was obtained. The trial ANCOVA analysis was performed and the resulting estimate of treatment difference compared to that which was actually reported. There is a discrepancy between the numbers of patients used for each of the QoL measures within a particular trial. Although the various QoL measures were contained on the same postal questionnaire, some missing items may have occurred within a particular QoL measure. This meant that the QoL score could not be calculated and thus, for that patient the score was missing. Since this thesis deals with missing forms, any forms for which some items were missing and scores could not be calculated were assumed to be missing for that particular QoL measure. Therefore, in some cases for a particular patient it was

possible to have an observed EQ5D score, but missing SF12 component scores (or vice-versa).

To assess the accuracy of imputation two criteria will be used: a measure of bias and a measure of precision. These were described in section 6.4. For the purposes of this work, it is assumed that the reported estimate of treatment difference is the true value. This then allows the accuracy of the imputation to be assessed. Clearly, this is not actually the case as the most unbiased estimate would be obtained if all patients were included in the analysis and there was no missing data. Making this assumption allows the accuracy of the imputation to be assessed, due to the fact that the imputation is being carried out on real, known data.

From this process, for each trial, the 'best' simple and multiple imputation method were identified. The best method was not necessarily the one which showed least bias, but the method which had low bias and a good estimate of precision (ratio of confidence interval width close to one). These two imputation methods were then used to impute the actual missing data found in the original trial dataset in order to assess what effect the use of imputation may have had on the trial conclusions. The remainder of this chapter presents the results of this process on a trial by trial basis and is followed by a summary and discussion of the findings.

7.1.2 Methods of imputation

Chapter six detailed a number of imputation strategies which are available. The simple imputation methods which will be applied to the datasets here are: simple mean; maximum value; minimum value; baseline carried forward (BCF); last value carried forward (LVCF); regression (see each trial for the model description). For the multiple imputation methods, the variables to be included in the imputation model needed to be identified. Carpenter and Kenward suggest that the imputation model should be as complex as the model used for analysis (Carpenter, Kenward 2007). Those variables which are both predictors of (QoL) outcome and missingness will be useful in the model. Identification of these

variables was undertaken using standard statistical techniques (t-tests, chi-squared tests and so on) and the inclusion in a multivariate model was assessed. Following this, two sets of imputation models were implemented in the multiple imputation process for the regression method, predictive mean match (PMM) method and the propensity score method. Firstly, a model containing identified covariates only (non- QoL) and secondly, the same covariate set plus any previous QoL scores. These methods require a monotone missing data pattern. Therefore, the MCMC method is used to make the data monotone followed by one of the other methods, regression, predictive mean match or propensity model. The MCMC method was also carried out on all the missing data (both intermittent and monotone). Therefore, for each trial dataset seven different multiple imputation procedures were carried out, in addition to the six simple imputation procedures.

7.2 REFLUX

The primary outcome in REFLUX was the reflux QoL score (RQLS). In addition, the EQ5D and SF12 (physical and mental component) scores were employed. Table 3.3 showed the mean (SD) QoL scores at each assessment split by treatment group. The results of the trial analysis were provided in Table 3.4, where an ANCOVA adjusting for the sex, age, BMI and the baseline QoL score was used. There was a significant difference in the 12 month QoL scores between the two treatment groups for the RQLS and SF12 physical component score, but no significant difference found for the EQ5D and SF12 mental component score.

7.2.1 Simple imputation of reminder response data

There were 309 (87%) patients who provided both baseline and 12 month EQ5D scores. Of which, there were 176 (57%) reminder responders at 12 months. For the SF12 scores, at 12 months 299 (84%) patients responded and also provided baseline scores (170 (57%) were reminder-responders). Finally for the RQLS, 265 (74%) patients had responded at both baseline and 12 months, with 145 (55%) reminder-responders. Each of the six simple imputation methods described in section 7.1.2, were applied to the missing data (reminder-responses) at 12 months.

The ANCOVA analysis was carried out each of these imputed datasets and the accuracy of imputation assessed.

Table 7.1: REFLUX - ANCOVA results after imputation of reminder scores

	12 month treatment comparison					
	Difference	SE	95% CI	p-value	Bias	Ratio of CI width
EQ5D (N=309)						
Observed data	**0.047**	**0.030**	**(-0.003, 0.097)**	**0.07**	**-**	**-**
Mean	0.016	0.016	(-0.015, 0.047)	0.30	0.031	0.62
Maximum	0.028	0.020	(-0.011, 0.068)	0.16	0.019	0.79
Minimum	-0.030	0.049	(-0.126, 0.066)	0.54	0.077	1.92
BCF	0.033	0.016	(0.002, 0.064)	0.04	0.014	0.62
LVCF	0.068	0.025	(0.018, 0.117)	0.001	0.021	0.99
Regression*	0.065	0.014	(0.037, 0.093)	<0.001	0.018	0.56
RQLS (N=276)						
Observed data	**14.05**	**2.29**	**(9.53, 18.6)**	**<0.001**	-	**-**
Mean	5.42	1.51	(2.46, 8.39)	<0.001	8.63	0.65
Maximum	6.15	1.87	(2.46, 9.84)	0.001	7.90	0.81
Minimum	3.12	3.70	(-4.16, 10.4)	0.399	10.93	1.61
BCF	5.92	1.91	(2.16, 9.68)	0.002	8.13	0.83
LVCF	12.50	2.41	(7.79, 17.3)	<0.001	1.55	1.05
Regression*	5.60	1.43	(2.78, 8.42)	<0.001	8.45	0.62
SF12 physical component score (N=299)						
Observed data	**3.51**	**0.88**	**(1.77, 5.25)**	**<0.001**	-	**-**
Mean	1.35	0.68	(-0.001 , 2.69)	0.050	2.16	0.77
Maximum	2.32	1.13	(0.09, 4.55)	0.041	1.19	1.33
Minimum	-0.64	1.97	(-4.53, 3.24)	0.745	4.15	2.23
BCF	1.760	0.64	(0.50, 3.03)	0.006	1.75	1.01
LVCF	3.130	0.89	(1.38, 4.88)	0.001	0.38	1.01
Regression*	1.880	0.61	(0.67, 3.09)	0.002	1.63	0.70
SF12 mental component score (N=299)						
Observed data	**1.54**	**1.18**	**(-0.78, 3.86)**	**0.19**	-	**-**
Mean	0.24	0.70	(-1.15, 1.62)	0.74	1.30	0.60
Max	1.10	1.08	(-1.04, 3.23)	0.31	0.44	0.92
Min	-1.67	1.83	(-5.27, 1.92)	0.36	3.21	1.55
Baseline CF	0.61	0.70	(-0.77, 1.99)	0.39	0.93	0.59
LVCF	2.42	1.10	(0.25, 4.59)	0.03	0.88	0.94
Regression*	0.50	0.62	(-0.72, 1.72)	0.42	1.04	0.53

* baseline QoL, sex, treatment, BMI group and age group;
SE=standard error of treatment difference estimate;
Bias = |observed treatment difference – treatment difference under imputation|

The results are shown in Table 7.1 for each of the four QoL scores. This shows that under the different imputation strategies the conclusion of a treatment difference on the EQ5D outcome was altered and the magnitude of effect differed. The mean, minimum and maximum methods showed no treatment difference ($p>0.05$); while BCF, LVCF and regression did ($p<0.05$). The smallest bias in treatment difference occurred with the BCF method ($b = 0.014$). However, the

standard error (SE) was underestimated using this method and the ratio of CI width was 0.62. Although LVCF showed a slightly greater bias, the precision was much better, as the ratio was 0.99 and very close to the desired value of one. Overall it appeared that LVCF would be the better simple imputation method in this case.

In the case of the RQLS, the observed treatment difference was 14.05 (SE=2.29). The imputation method which provided least bias was LVCF. This also showed the best precision value with the ratio = 1.05. The remaining imputation methods were poor for the RQLS and provided bias values of 7.9 upwards. A significant difference on the SF12 PCS was found using mean, maximum, BCF, LVCF and regression imputation methods (Table 7.1). The minimum method did not show a treatment difference. The least biased method was LVCF, which also showed a good value of precision (ratio=1.01).

There was no observed difference in the MCS between treatment groups. This conclusion was preserved under imputation, except for the LVCF method, which provided a significant treatment difference. The direction of effect differed with the minimum method. A negative difference (surgical group had better QoL than medical group) was found compared to the positive difference shown by the other methods. The best method in this case was found to be maximum value imputation.

7.2.2 Multiple imputation of reminder response data

The multiple imputation strategies to be implemented were described in section 7.1.2. As discussed previously, the variables involved in the analysis model should always be included in the imputation model. In addition, any variables which are predictors of both outcome and missingness should be included. No additional covariates were found to be predictors of these, so the variables used in the ANCOVA and previous QoL scores (where available) were used in the imputation models. The ANCOVA results under each of the multiple imputation

methods for each QoL score are shown in Table 7.2, along with the bias and ratio of CI width.

Table 7.2: REFLUX - ANCOVA results after multiple imputation of reminder response data

	12 month treatment comparison					
	Mean Difference	SE	95% CI	p-value	Bias	Ratio of CI width
EQ5D (N=309)						
Observed data	**0.047**	**0.025**	**(-0.003, 0.097)**	**0.07**	**-**	**-**
MCMC for intermittent	0.092	0.034	(0.020, 0.160)	0.02	0.053	1.36
MCMC + regression*	0.060	0.037	(-0.022, 0.141)	0.14	0.024	1.58
MCMC + PMM*	0.055	0.026	(-0.003, 0.107)	0.04	0.008	1.07
MCMC + propensity*	0.031	0.025	(-0.018, 0.080)	0.22	0.017	0.95
MCMC + regression**	0.073	0.038	(-0.011, 0.157)	0.08	0.038	1.63
MCMC + PMM**	0.074	0.038	(-0.011, 0.160)	0.08	0.039	1.66
MCMC + propensity**	0.036	0.034	(-0.041, 0.101)	0.39	0.019	1.38
RQLS (N=276)						
Observed data	**14.05**	**2.29**	**(9.53, 18.6)**	**<0.001**	-	**-**
MCMC for intermittent	12.5	3.29	(5.28, 19.7)	0.03	1.60	1.59
MCMC + regression*	11.7	2.68	(6.34, 17.1)	<0.001	2.39	1.19
MCMC + PMM*	11.5	3.23	(4.63, 18.4)	0.003	2.59	1.52
MCMC + propensity*	6.6	2.74	(1.19, 12.1)	0.018	7.45	1.20
MCMC + regression**	13.7	2.78	(7.94, 19.4)	<0.001	0.39	1.26
MCMC + PMM**	12.4	3.71	(3.99, 20.9)	0.009	1.69	1.93
MCMC + propensity**	6.3	3.03	(0.13, 12.5)	0.046	7.76	1.36
SF12 physical component score (N=299)						
Observed data	**3.51**	**0.88**	**(1.77, 5.25)**	**<0.001**	-	**-**
MCMC for intermittent	4.30	1.03	(2.23, 6.38)	<0.001	0.79	1.19
MCMC + regression*	3.47	1.82	(-0.69, 7.62)	0.09	0.04	2.39
MCMC + PMM*	2.90	1.74	(-1.01, 6.82)	0.13	0.61	2.25
MCMC + propensity*	1.82	1.47	(-1.27, 4.91)	0.23	1.69	1.78
MCMC + regression**	4.02	1.36	(1.07, 6.97)	0.01	0.51	1.70
MCMC + PMM**	2.98	1.95	(-1.72, 7.68)	0.17	0.53	2.70
MCMC + propensity**	1.77	1.17	(-0.53, 4.07)	0.13	1.74	1.32
SF12 mental component score (N=299)						
Observed data	**1.54**	**1.18**	**(-0.78, 3.86)**	**0.19**	-	**-**
MCMC for intermittent	1.71	1.33	(-1.09, 4.50)	0.22	0.18	1.20
MCMC + regression*	0.75	1.27	(-1.82, 3.32)	0.56	0.78	1.11
MCMC + PMM*	1.34	1.48	(-1.77, 4.44)	0.38	0.19	1.34
MCMC + propensity*	0.23	1.36	(-2.54, 3.00)	0.87	1.30	1.19
MCMC + regression**	2.25	1.10	(0.07, 4.43)	0.04	0.72	0.94
MCMC + PMM**	2.05	1.08	(-0.09, 4.19)	0.06	0.52	0.92
MCMC + propensity**	-1.1	1.24	(-3.55, 1.35)	0.38	2.63	1.06

* MI model based on the ANCOVA model and additional covariates;
** MI model based on the ANCOVA model, additional covariates and previous QoL;

The best method for the EQ5D treatment difference was MCMC imputation to make the data monotone followed by a PMM model which included the analytic variables (Table 7.2). This method showed the least bias and showed a good precision value (close to one). In the case of the RQLS, MCMC followed by

regression (analysis model plus previous QoL) was the most accurate with a treatment difference (SE) of 13.7 (2.78) compared to the observed of 14.05 (2.29). The precision of this estimate was not that good with ratio=1.26, but this was not the worst method for this criterion. The conclusion of a significant treatment difference was maintained with $p<0.001$ for both the observed data and that under MCMC regression imputation.

The most accurate multiple imputation method for the PCS was MCMC followed by a regression model (analysis variables only) with b = 0.04. However, the standard error was over double that seen in the observed data, resulting in a precision estimate of 2.29. This altered the conclusion of the significance test, to suggest there was no difference ($p=0.09$) in twelve month PCS between treatment groups, when in fact there was one observed ($p<0.001$). MCMC on all missing data was the least biased when imputing the MCS (Table 7.2), but the estimate of precision was 1.20. The observed data showed no treatment difference ($p=0.19$), which was replicated under MCMC imputation ($p=0.22$).

7.2.3 Imputation of actual missing data

The previous two sub-sections identified the most suitable simple and multiple imputation procedures for reminder responses. This section applies these methods to all the actual missing data and re-analyses the augmented datasets. Table 7.3 shows the results of the ANCOVA analysis under the best simple imputation and also the best multiple imputation method. For the EQ5D score, the result after multiple imputation was reasonably close to that from the observed (complete) data. The results cannot be directly compared as there were a different number of patients involved. The slight differences in treatment difference and standard errors result in slightly different confidence intervals and p-values. However, the conclusion of a non-significant difference in EQ5D score between groups is maintained ($p>0.05$). In the case of the RQLS, under multiple imputation the treatment effect was more conservative at 13.6 (2.78), compared with 14.1 (2.29) for the observed data. The estimate under LVCF was even less,

but both imputation methods found a significant treatment difference, which was seen in the observed data (p<0.001).

Table 7.3: REFLUX - ANCOVA results under imputation of all missing data
12 month treatment comparison

Imputation Method	N	Difference	SE	95% CI	p-value
EQ5D					
Observed Data	309	0.047	0.025	(-0.003, 0.10)	0.07
LVCF	344	0.045	0.024	(-0.002, 0.09)	0.06
MCMC PMM*	355	0.049	0.026	(-0.001, 0.10)	0.06
RQLS					
Observed Data	276	14.1	2.29	(9.53, 18.6)	<0.001
LVCF	327	11.9	2.11	(7.71, 16.0)	<0.001
MCMC + regression**	357	12.7	2.38	(7.95, 17.5)	<0.001
SF12 physical component score (PCS)					
Observed Data	299	3.51	0.88	(1.77, 5.25)	<0.001
LVCF	336	3.29	0.81	(1.69, 4.89)	<0.001
MCMC + regression*	357	3.16	0.89	(1.40, 4.91)	<0.001
SF12 mental component score (MCS)					
Observed Data	299	1.54	1.18	(-0.78, 3.76)	0.19
Maximum	336	2.49	1.31	(-0.09, 5.07)	0.06
MCMC	357	1.48	1.14	(-0.76, 3.71)	0.19

* based on the ANCOVA model and additional covariates;
** based on the ANCOVA model, additional covariates and previous QoL;

For the SF12 PCS the calculated treatment differences under imputation were more conservative than the complete data. However, after both simple and multiple imputation the conclusion of a treatment difference in the 12 month PCS was maintained. Under simple imputation for the MCS, the estimated treatment difference was inflated, which is perhaps to be expected with the maximum method. Under MI the treatment difference was slightly reduced when compared to the observed data, but all three imputation methods still provided no evidence of a treatment difference.

7.2.4 Summary

The choice of simple imputation method had an impact on the accuracy of the estimate of treatment difference. In REFLUX, strategies using previous QoL scores and in particular LVCF seemed appropriate, yet it is well known that this method is not recommended (Carpenter, Kenward 2007). The implications of this will be discussed later (section 7.9). Multiple imputation of reminder responders suggested that for the EQ5D score MCMC followed by a PMM model was the

most accurate. It was MCMC followed by regression for the RQLS and SF12 PCS and MCMC on all missing data for the SF12 MCS that were the most accurate. Using these chosen methods to impute all the actual missing data provided more conservative estimates of treatment difference. The conclusions of the significance tests were not altered. Comparing the best simple imputation method against multiple imputation on the whole showed multiple imputation to be the better choice. Often the treatment difference was closer to that observed with a more appropriate standard error.

7.3 MAVIS

During the MAVIS trial, QoL was measured using the EQ5D and SF12 instruments at baseline, six and 12 months. The trial analysis of the QoL data was an ANCOVA to estimate the mean difference between treatment groups after adjusting for baseline values and the covariates age group, sex and type of housing. There was no statistically significant effect from supplementation on QoL at either six or 12 months (Table 3.6).

7.3.1 Simple imputation of reminder response data

At 12 months, there were 829 (91%) patients with EQ5D scores and 823 (90%) with SF12 scores. Reminder response occurred for 12% of patients at 12 months. Table 7.4 shows the results of the ANCOVA analysis after the reminder-responses had been imputed using simple imputation methods. For the EQ5D data, the least biased simple imputation method was BCF (b = 0.002). This method also showed good precision (ratio = 1.02) and was, therefore, the best imputation method in this instance. The minimum value and maximum value imputation methods were the worst methods. The conclusion of no significant treatment difference in EQ5D scores at 12 months was maintained for all methods ($p>0.05$).

Under imputation the treatment difference for the SF12 PCS differed in direction to that which was observed (except for the maximum value method (Table 7.4)).

Despite this, a conclusion of no significant treatment difference with respect to physical QoL was seen under each imputation strategy. The best method (least biased and good precision) was mean imputation, but this method did show a negative treatment difference.

Table 7.4: MAVIS - ANCOVA results after simple imputation of reminder scores

	12 month treatment comparison					
	Mean Difference	SE	95% CI	p-value	Bias	Ratio of CI width
EQ5D (N=829)						
Observed Data	**-0.019**	**0.011**	**(-0.04, 0.002)**	**0.08**	**-**	**-**
Mean	-0.013	0.010	(-0.03, 0.01)	0.21	0.006	0.95
Maximum	-0.007	0.012	(-0.03, 0.02)	0.52	0.012	1.19
Minimum	-0.037	0.023	(-0.08, 0.01)	0.10	0.018	2.14
BCF	-0.017	0.010	(-0.04, 0.003)	0.10	0.002	1.02
LVCF	-0.011	0.011	(-0.03, 0.01)	0.31	0.008	0.95
Regression*	-0.015	0.010	(-0.03, 0.005)	0.13	0.004	0.83
SF12 physical component score (N=823)						
Observed Data	**0.07**	**0.49**	**(-0.90, 1.03)**	**0.89**	**-**	**-**
Mean	-0.03	0.50	(-1.01, 0.95)	0.95	0.10	1.02
Maximum	0.42	0.66	(-0.88, 1.73)	0.52	0.36	1.35
Minimum	-0.92	0.97	(-2.81, 0.98)	0.34	0.98	1.96
BCF	-0.20	0.47	(-1.12, 0.71)	0.66	0.27	0.95
LVCF	-0.15	0.49	(-1.11, 0.81)	0.75	0.22	0.99
Regression*	-0.14	0.46	(-1.05, 0.76)	0.76	0.21	0.94
SF12 mental component score (N=823)						
Observed Data	**-0.03**	**0.55**	**(-1.11, 1.05)**	**0.96**	**-**	**-**
Mean	0.08	0.52	(-0.95, 1.10)	0.89	0.11	0.95
Maximum	0.54	0.70	(-0.82, 1.91)	0.44	0.57	1.26
Minimum	-0.77	0.96	(-2.65, 1.11)	0.42	0.74	1.74
BCF	0.05	0.52	(-0.98, 1.08)	0.92	0.08	0.95
LVCF	0.13	0.55	(-0.94, 1.20)	0.82	0.16	0.99
Regression*	0.07	0.51	(-0.93, 1.08)	0.89	0.10	0.93

*based on baseline QoL, sex, residence type and age group

A similar phenomenon was seen after imputation of the MCS. In this case, under imputation a positive difference was found by most imputation methods, but the observed data showed a negative difference. The conclusion of the significance test remained non-significant in all cases. The least biased method for imputing the MCS was BCF, as was seen with the EQ5D score. However, the LVCF method provided better precision. Considering bias and precision, the BCF method would have been most appropriate for the MCS.

7.3.2 Multiple imputation of reminder response data

As previously discussed, the analysis model included age group, sex and residence type. In addition to these, the presence of chronic infection at recruitment was found to be associated with both outcome and missingness, so was included in the imputation model. Table 7.5 shows the results of the ANCOVA analysis for the MAVIS trial data after the reminder response data had undergone MI. For the EQ5D score the method displaying the least total bias was MCMC followed by a PMM (covariates only). The precision was reasonable at 1.07 but could be improved upon with the second PMM model that included previous QoL. The significance test showed no treatment difference at the 5% level with p=0.09 for both the observed data and under the MI imputation PMM model.

Table 7.5: MAVIS – ANCOVA results after multiple imputation of reminder response data

	12 month treatment comparison					
	Mean Difference	SE	95% CI	p-value	Bias	Ratio of CI width
EQ5D (N=829)						
Observed data	**-0.019**	**0.011**	**(-0.040, 0.003)**	0.09	-	-
MCMC for intermittent	-0.012	0.012	(-0.036, 0.011)	0.31	0.007	1.09
MCMC + regression*	-0.012	0.014	(-0.040, 0.015)	0.37	0.007	1.28
MCMC + PMM*	-0.020	0.012	(-0.043, 0.003)	0.09	0.001	1.07
MCMC + propensity*	-0.015	0.013	(-0.040, 0.011)	0.26	0.004	1.19
MCMC + regression **	-0.013	0.011	(-0.035, 0.009)	0.24	0.006	1.02
MCMC + PMM**	-0.017	0.011	(-0.038, 0.004)	0.12	0.002	0.98
MCMC + propensity**	-0.015	0.013	(-0.041, 0.010)	0.24	0.004	1.19
SF12 physical component score (N=823)						
Observed data	**0.07**	**0.49**	**(-0.89, 1.03)**	0.89	-	-
MCMC for intermittent	-0.20	0.52	(-1.22, 0.82)	0.70	0.27	1.06
MCMC + regression*	-0.05	0.57	(-1.18, 1.07)	0.93	0.12	1.17
MCMC + PMM*	-0.21	0.57	(-1.34, 0.42)	0.72	0.28	0.92
MCMC + propensity*	-0.08	0.60	(-1.25, 1.09)	0.89	0.15	1.22
MCMC + regression **	-0.13	0.50	(-1.11, 0.86)	0.80	0.19	1.03
MCMC + PMM**	-0.05	0.55	(-1.15, 1.05)	0.93	0.12	1.15
MCMC + propensity**	-0.24	0.73	(-1.73, 1.26)	0.75	0.31	1.56
SF12 mental component score(N=823)						
Observed data	**-0.030**	**0.55**	**(-1.11, 1.05)**	**0.96**	-	-
MCMC for intermittent	0.222	0.57	(-0.89, 1.33)	0.69	0.25	1.03
MCMC + regression*	0.033	0.64	(-1.25, 1.31)	0.96	0.06	1.19
MCMC + PMM*	-0.033	0.58	(-1.17, 1.11)	0.96	0.003	1.06
MCMC + propensity*	-0.061	0.67	(-1.39, 1.27)	0.93	0.03	1.23
MCMC + regression **	0.216	0.59	(-0.95, 1.38)	0.72	0.25	1.08
MCMC + PMM**	0.308	0.57	(-0.81, 1.42)	0.59	0.34	1.02
MCMC + propensity**	0.067	0.63	(-1.18, 1.31)	0.92	0.10	1.15

* imputation model – age group, sex, residence type and chronic infection

** imputation model - age group, sex, residence type and chronic infection plus previous QoL

The least biased method for the PCS was MCMC, followed by a PMM model (Table 7.5). A more precise estimate was seen with the regression method, but the bias was greater. All the imputation strategies produced a treatment effect that was negative and non-significant. However, in the observed data, a positive difference was identified, but also not significantly different between treatment groups. For the MCS, the most accurate method was the MCMC, followed by a PMM model (covariates only). The precision was good and although this was improved upon when the QoL was added to the imputation model, this worsened the bias. The other MI methods provided treatment estimates in the opposite direction, although the result of the significance test remained the same.

7.3.3 Imputation of actual missing data

Baseline carried forwards was the most accurate simple imputation for two of the three QoL scores. MCMC followed by a PMM model was the most suitable MI method. Using these methods, the actual missing data was imputed and the ANCOVA carried out. The results are shown in Table 7.6.

Table 7.6: MAVIS - ANCOVA results under imputation of all missing data

		12 month treatment comparison			
Imputation Method	N	Difference	SE	95% CI	p-value
EQ5D					
Observed Data	829	-0.019	0.011	(-0.040, 0.003)	0.09
BCF	908	-0.019	0.010	(-0.038, 0.001)	0.06
MCMC + PMM*	910	-0.016	0.010	(-0.038, 0.006)	0.16
SF12 physical component score (PCS)					
Observed Data	823	0.07	0.49	(-0.89, 1.03)	0.89
Mean	906	0.23	0.48	(-0.71, 1.17)	0.63
MCMC + PMM**	910	0.30	0.54	(-0.76, 1.37)	0.58
SF12 mental component score (MCS)					
Observed Data	823	-0.004	0.55	(-1.09, 1.08)	0.99
BCF	906	-0.06	0.51	(-1.06, 0.94)	0.90
MCMC + PMM*	910	-0.04	0.54	(-1.09, 1.02)	0.95

* age group, sex, residence type and chronic infection ; ** age group, sex, residence type and chronic infection +QoL; PCS – physical component score; MCS – mental component score

Using baseline carried forwards for the missing EQ5D data gave very similar results for 908 patients to that which was observed for the 829 patients. Results for the two SF12 component scores were not as consistent. Using imputation on all

missing data increased the treatment difference in both cases. However, the conclusion of the significance test for treatment difference remained unchanged (non-significant).

7.3.4 Summary

In general, imputation (simple or multiple) was reasonable for the EQ5D data but was erratic for the SF12 data. Baseline carried forwards was the most accurate simple imputation method for the reminder data, while a PMM model was the best MI method.

7.4 RECORD

The RECORD trial utilised the EQ5D and SF12 QoL instruments. As discussed in section 3.4, data was collected at five time points, but only the first three were used here (4 months (baseline), 12 and 24 months (follow up)). QoL outcomes at 24 months were examined by ANCOVA to detect treatment difference, with adjustment for 4-month quality of life scores, age group (<80, 80+), sex, time since fracture (<90 days, 90 days+) and type of fracture (proximal femur, distal forearm, clinical vertebral as indicator variables). There were four treatment groups resulting in two comparisons: calcium versus no calcium and vitamin D versus no vitamin D. The accuracy of imputation with regard to each of these treatment comparisons will be assessed separately.

7.4.1 Simple imputation of reminder response data

QoL scores for those patients who were reminder-responders at 24 months were imputed. The ANCOVA was carried out and the results are shown for the calcium comparison in Table 7.7 and the Vitamin D comparison in Table 7.8. The direction of the calcium effect on the EQ5D score was positive for all methods. The least biased imputation method was LVCF (b = 0.002) with ratio=0.87. For the Vitamin D comparison, the direction of treatment difference differed between methods, with minimum and LVCF preserving the negative difference observed, and LVCF producing the least bias (b = 0.001).

For the SF12 PCS, all methods provided a treatment difference in the observed direction for Vitamin D, but the maximum method showed a negative difference for the calcium comparison. BCF was the method showing the least bias for the calcium comparison, while the minimum method showed least bias of the vitamin D comparison. The ratio of the CI width was closest to one with the LVCF method. For the MCS, the magnitude and direction of effect varied under different imputation strategies for both treatment comparisons. LVCF was the best simple imputation method for both the calcium comparison and the vitamin D comparison, as it showed the least bias and a precision value close to one.

Table 7.7: RECORD – ANCOVA for calcium comparison after simple imputation of reminder score

	24 month calcium comparison					
Method	Difference	SE	95% CI	p-value	Bias	Ratio of CI width
EQ5D (n=2879)						
Observed data	**0.015**	**0.008**	**(0.00, 0.030)**	**0.05**	**-**	**-**
Mean	0.010	0.007	(-0.004, 0.024)	0.15	0.005	0.93
Maximum	0.005	0.008	(-0.011, 0.021)	0.51	0.010	1.07
Minimum	0.035	0.021	(-0.007, 0.076)	0.10	0.020	2.77
Baseline CF (4m)	0.011	0.007	(-0.002, 0.024)	0.11	0.004	0.87
Regression*	0.011	0.007	(-0.002, 0.024)	0.10	0.004	0.87
LVCF	0.013	0.007	(-0.001, 0.027)	0.08	0.002	0.93
SF 12 physical component score(N=2695)						
Observed data	**0.44**	**0.31**	**(-0.17, 1.05)**	**0.16**	**-**	**-**
Mean	0.12	0.30	(-0.47, 0.70)	0.70	0.32	0.96
Maximum	-0.51	0.53	(-1.55, 0.52)	0.33	0.95	1.70
Minimum	0.93	0.60	(-0.25, 2.11)	0.12	0.49	1.93
Baseline CF (4m)	0.37	0.28	(-0.17, 0.91)	0.19	0.07	0.89
Regression*	0.33	0.27	(-0.21, 0.86)	0.23	0.11	0.88
LVCF	0.26	0.30	(-0.33, 0.84)	0.39	0.18	0.96
SF12 mental component score (N=2695)						
Observed data	**0.03**	**0.33**	**(-0.63, 0.68)**	**0.94**	**-**	**-**
Mean	0.20	0.30	(-0.39, 0.79)	0.51	0.17	0.90
Maximum	-0.26	0.47	(-1.18, 0.67)	0.59	0.29	1.41
Minimum	1.11	0.75	(0.35, 2.58)	0.14	1.08	1.70
Baseline CF (4m)	-0.06	0.30	(-0.64, 0.53)	0.85	0.09	0.89
Regression*	0.09	0.29	(-0.47, 0.65)	0.76	0.06	0.85
LVCF	0.10	0.32	(-0.54, 0.73)	0.76	0.07	0.97

* On baseline QoL, sex, time since recruiting fracture, fracture type and age group

Table 7.8: RECORD – ANCOVA for Vitamin D comparison after simple imputation of reminder score

	24 month Vitamin D comparison					
Method	Difference	SE	95% CI	p-value	Bias	Ratio of CI width
EQ5D (n=2879)						
Observed data	**-0.002**	**0.008**	**(-0.017, 0.013)**	**0.81**	**-**	**-**
Mean	0.009	0.007	(-0.005, 0.022)	0.22	0.011	0.90
Maximum	0.011	0.008	(-0.005, 0.027)	0.17	0.013	1.07
Minimum	-0.005	0.021	(-0.046, 0.037)	0.82	0.003	2.77
Baseline CF (4m)	0.003	0.007	(-0.01, 0.016)	0.64	0.005	0.87
Regression*	0.005	0.007	(-0.008, 0.018)	0.43	0.007	0.87
LVCF	-0.001	0.007	(-0.016, 0.013)	0.87	0.001	0.97
SF 12 physical component score (N=2695)						
Observed data	**0.16**	**0.31**	**(-0.45, 0.77)**	**0.61**	**-**	**-**
Mean	0.39	0.30	(-0.19, 0.98)	0.19	0.23	0.96
Maximum	0.48	0.53	(-0.55, 1.52)	0.36	0.32	1.66
Minimum	0.28	0.60	(-0.25, 2.11)	0.64	0.12	1.93
Baseline CF (4m)	0.33	0.28	(-0.21, 0.87)	0.23	0.17	0.89
Regression*	0.35	0.27	(-0.18, 0.88)	0.20	0.19	0.87
LVCF	0.39	0.30	(-0.20, 0.97)	0.20	0.23	0.96
SF12 mental component score (N=2695)						
Observed data	**-0.02**	**0.33**	**(-0.68, 0.68)**	**0.94**	**-**	**-**
Mean	0.24	0.30	(-0.35, 0.83)	0.43	0.26	0.87
Maximum	0.30	0.47	(-0.62, 1.22)	0.52	0.32	1.35
Minimum	0.12	0.75	(-1.34, 1.58)	0.87	0.14	2.15
Baseline CF (4m)	-0.07	0.30	(-0.65, 0.52)	0.83	0.05	0.86
Regression*	0.07	0.29	(-0.49, 0.64)	0.80	0.09	0.56
LVCF	-0.06	0.32	(-0.69, 0.58)	0.86	0.04	0.93

* On baseline QoL, sex, time since recruiting fracture, fracture type and age group

7.4.2 Multiple imputation of reminder response data

Table 7.9 shows the results of the ANCOVA for the calcium treatment difference for each of the QoL scores and treatment comparisons after multiple imputation of the reminder response data. The MCMC approach for all intermittent missingness provided the least bias for both treatment comparisons of the EQ5D scores, yet this method did not display the best precision values. Under MI, the result of the significance test changed for the calcium comparison. The observed p-value was borderline at p=0.05, but under MI it was p=0.10 and above. Under MI, the direction of the difference for the Vitamin D comparison was reversed. There was no significant difference found in EQ5D scores between those who did and did not receive Vitamin D supplementation.

MCMC on all missing data provided the least bias for the calcium treatment effect on the SF12 physical component scores (Table 7.9).

Table 7.9: RECORD – ANCOVA for calcium comparison after multiple imputation of reminder response data

	24 month calcium comparison					
Method	Difference	SE	95% CI	p-value	Bias	Ratio of CI width
EQ5D (n=2879)						
Observed data	**0.015**	**0.008**	**(0.0001, 0.030)**	**0.05**	**-**	**-**
MCMC for intermittent	0.013	0.008	(-0.002, 0.027)	0.10	0.002	0.96
MCMC + regression*	0.012	0.009	(-0.005, 0.029)	0.17	0.003	1.13
MCMC + PMM*	0.013	0.01	(-0.006, 0.032)	0.18	0.002	1.26
MCMC + propensity*	0.018	0.009	(-0.001, 0.035)	0.04	0.003	1.20
MCMC + regression**	0.019	0.008	(0.004, 0.034)	0.02	0.004	1.00
MCMC + PMM**	0.018	0.008	(0.001, 0.035)	0.04	0.003	1.20
MCMC + propensity**	0.018	0.009	(0.001, 0.035)	0.04	0.003	1.20
SF12 physical component score (N=2695)						
Observed data	**0.44**	**0.31**	**(-0.18, 1.05)**	**0.16**	**-**	**-**
MCMC for intermittent	0.43	0.34	(-0.24, 1.10)	0.21	0.01	1.09
MCMC + regression*	0.38	0.38	(-0.38, 1.14)	0.33	0.06	1.24
MCMC + PMM*	0.95	0.35	(0.27, 1.62)	0.01	0.51	1.10
MCMC + propensity*	0.27	0.38	(-0.48, 1.03)	0.47	0.17	1.23
MCMC + regression**	0.29	0.42	(-0.59, 1.18)	0.49	0.15	1.44
MCMC + PMM**	0.37	0.37	(-0.37, 1.11)	0.32	0.07	1.20
MCMC + propensity**	1.08	0.37	(0.30, 1.78)	0.01	0.64	1.20
SF12 mental component score (N=2695)						
Observed data	**0.03**	**0.33**	**(-0.63, 0.68)**	**0.94**	**-**	**-**
MCMC for intermittent	0.14	0.37	(-0.61, 0.88)	0.72	0.11	1.14
MCMC + regression*	0.39	0.38	(-0.37, 1.14)	0.32	0.36	1.15
MCMC + PMM*	0.29	0.46	(-0.67, 1.25)	0.54	0.26	1.47
MCMC + propensity*	0.23	0.39	(-0.55, 1.02)	0.56	0.21	1.20
MCMC + regression**	-0.06	0.44	(-0.98, 0.87)	0.80	0.08	1.41
MCMC + PMM**	0.02	0.35	(-0.67, 0.70)	0.96	0.01	1.05
MCMC + propensity**	0.74	0.38	(-0.02, 1.49)	0.06	0.71	1.15

*age group, time since recruiting fracture, residence type after fracture, type of fracture, treatment.
** same as * plus previous QoL

For the Vitamin D treatment comparison, the least biased method used was MCMC to make the data monotone, followed by a propensity model (covariates and previous QoL). However, better precision was obtained with a PMM model. MCMC to make the data monotone, followed by a PMM model (covariates and previous QoL) was the least biased method for the calcium treatment comparison of the SF12 MCS. The remaining methods were poor, with a bias of at least 0.08 in all cases, ranging to a maximum of 0.71. For the Vitamin D comparison of the MCS, all methods except MCMC for intermittent data showed a treatment effect of positive magnitude rather than the negative difference observed. Again, however,

the resulting p-values were non-significant indicating no evidence of a difference in MCS at 24 months between those taking Vitamin D supplementation and those not. The least biased method in this instance was MCMC imputation, followed by a regression model, which included the covariates and previous QoL. The models which showed best precision were MCMC on all missing data and the PMM.

Table 7.10: RECORD – ANCOVA for Vitamin D comparison after multiple imputation of reminder response data

	24 month Vitamin D comparison					
Method	Difference	SE	95% CI	p-value	Bias	Ratio of CI width
EQ5D (N=2879)						
Observed data	**-0.002**	**0.008**	**(-0.017, 0.013)**	**0.77**	**-**	**-**
MCMC for intermittent	0.002	0.008	(-0.015, 0.019)	0.79	0.004	1.13
MCMC + regression*	0.008	0.011	(-0.015, 0.032)	0.46	0.010	1.57
MCMC + PMM*	0.007	0.009	(-0.011, 0.025)	0.46	0.009	1.20
MCMC + propensity*	0.005	0.009	(-0.013, 0.023)	0.57	0.007	1.20
MCMC + regression**	0.003	0.008	(-0.013, 0.018)	0.73	0.005	1.03
MCMC + PMM**	0.005	0.008	(-0.012, 0.021)	0.57	0.007	1.10
MCMC + propensity**	0.005	0.009	(-0.013, 0.023)	0.57	0.007	1.20
SF12 physical component score (N=2695)						
Observed data	**0.15**	**0.31**	**(-0.46, 0.76)**	**0.64**	-	-
MCMC for intermittent	0.28	0.35	(-0.41, 0.97)	0.42	0.13	1.13
MCMC + regression*	0.35	0.40	(-0.44, 1.14)	0.38	0.20	0.48
MCMC + PMM*	0.64	0.44	(-0.27, 1.55)	0.16	0.49	1.49
MCMC + propensity*	0.37	0.43	(-0.51, 1.25)	0.40	0.22	1.44
MCMC + regression**	0.27	0.32	(-0.37, 0.91)	0.40	0.12	1.05
MCMC + PMM**	0.30	0.35	(-0.41, 1.01)	0.40	0.15	1.16
MCMC + propensity**	0.19	0.36	(-0.53, 0.91)	0.60	0.04	1.18
SF12 mental component score (N=2695)						
Observed data	**-0.024**	**0.33**	**(-0.68, 0.63)**	**0.94**	-	-
MCMC for intermittent	-0.084	0.33	(-0.74, 0.57)	0.80	0.06	1.00
MCMC + regression*	0.230	0.39	(-0.55, 1.01)	0.56	0.25	1.19
MCMC + PMM*	0.64	0.43	(-0.22, 1.50)	0.14	0.66	1.31
MCMC + propensity*	0.24	0.37	(-0.48, 0.97)	0.51	0.26	1.11
MCMC + regression**	0.025	0.34	(-0.64, 0.69)	0.94	0.05	1.02
MCMC + PMM**	0.109	0.33	(-0.54, 0.76)	0.74	0.13	0.99
MCMC + propensity**	0.171	0.49	(-0.86, 1.20)	0.73	0.20	1.57

*age group, time since recruiting fracture, residence type after fracture, type of fracture, treatment.
** same as * plus previous QoL

7.4.3 Imputation of actual missing data

The previous two subsections utilised the reminder data to identify the most suitable simple and multiple imputation methods for missing QoL data in the RECORD trial. Table 7.11 shows the results of using these methods to impute all the actual missing data, rather than that which was collected via reminder. It is difficult to make a direct comparison between the results shown in Table 7.11

because of the different sample sizes involved. It is seen that using LVCF on the missing EQ5D outcomes at 24 months, finds a significant treatment difference. This difference was shown to be borderline significant in the observed data and is non-significant under MCMC imputation. Despite the variation in calculated treatment differences for the SF12 component scores, the conclusion of no significant treatment difference was consistent.

Table 7.11: RECORD – ANCOVA results after imputation on all missing data

			24 month treatment comparison			
Comparison	Method	N	Difference	SE	95% CI	p-value
		EQ5D				
Calcium	Observed Data	2879	0.015	0.008	(0.00, 0.30)	0.05
Calcium	LVCF	3906	0.016	0.006	(0.004, 0.028)	0.01
Calcium	MCMC	5291	0.010	0.01	(-0.003, 0.023)	0.14
Vitamin D	Observed Data	2879	-0.002	0.008	(-0.017, 0.013)	0.81
Vitamin D	LVCF	3906	0.003	0.006	(-0.009, 0.015)	0.62
Vitamin D	MCMC	5291	-0.0002	0.01	(-0.014, 0.013)	0.97
		SF 12 Physical component score				
Calcium	Observed Data	2695	0.44	0.31	(-0.17, 1.05)	0.16
Calcium	BCF	3643	0.35	0.24	(-0.11, 0.81)	0.14
Calcium	MCMC	5291	0.31	0.33	(-0.40, 1.02)	0.36
Vitamin D	Observed Data	2695	0.16	0.31	(-0.45, 0.77)	0.61
Vitamin D	Minimum	3643	0.42	0.54	(-0.63, 1.47)	0.44
Vitamin D	BCF	3643	0.09	0.24	(-0.37, 0.55)	0.71
Vitamin D	MCMC + propensity**	5291	-0.37	0.32	(-1.04, 0.29)	0.26
		SF 12 Mental component score				
Calcium	Observed Data	2695	0.03	0.33	(-0.63, 0.68)	0.94
Calcium	LVCF	3643	0.0001	0.28	(-0.55, 0.55)	0.99
Calcium	MCMC + PMM**	5291	-0.016	0.27	(-0.56, 0.53)	0.96
Vitamin D	Observed Data	2695	-0.02	0.33	(-0.68, 0.68)	0.94
Vitamin D	LVCF	3643	0.11	0.28	(-0.45, 0.66)	0.71
Vitamin D	MCMC + regression**	5291	-0.19	0.34	(-0.91, 0.53)	0.58

** age group, time since recruiting fracture, residence type after fracture, type of fracture, treatment, previous QoL; MCS – mental component score; PCS – physical component score

7.4.4 Summary

The simple imputation method most accurate for missing data in the RECORD trial was LVCF for both the EQ5D and SF12 mental scores. It is known that this method is not recommended, yet this is a prime example of why some researchers would use it as it has provided reasonable results in this instance. Baseline carried forward was most suitable for the PCS, but has similar problems to LVCF. MCMC imputation on all missing data was usually the best MI choice. However, when these methods were applied to the actual missing data, they did not perform so

well. Estimates of treatment differences were altered, yet on all but one occasion, the result of the significance test was unchanged. Using LVCF on the missing EQ5D outcomes would have altered the observed conclusion and found that there was a significant difference in the EQ5D scores at 24 months between those taking calcium supplementation and those not (Table 7.11).

7.5 KAT

The KAT trial consisted of three treatment comparisons as described in section 2.5. The results for the second treatment comparison (patella resurfacing versus no patella resurfacing) will be presented here. Trial analysis reported treatment difference at both one and two years, however to avoid repetition only the two-year comparison will be presented.

7.5.1 Simple imputation of reminder response data

As before reminder responses were removed, imputed using several methods and the ANCOVA applied to the augmented datasets. The results of the trial analysis for comparison B (patella resurfacing vs. no patella resurfacing) under simple imputation are shown in Table 7.12. LVCF was the method that showed the least bias for three of the four QoL scores. For the remaining score (SF12 MCS), BCF was the least biased method. The other simple imputation methods were fairly poor for the SF12 component scores. Despite LVCF showing least bias for the EQ5D, BCF provided a better estimate of precision.

Table 7.12: KAT – ANCOVA results after simple imputation of reminder response data

	Two year treatment comparison					
Imputation method	Difference	SE	95% CI	p-value	Bias	Ratio of CI width
EQ5D (N=1321)						
Observed data	**0.013**	**0.013**	**(-0.01, 0.04)**	**0.35**	**-**	**-**
Mean	0.004	0.012	(-0.02, 0.03)	0.76	0.007	1.00
Minimum	0.018	0.023	(-0.03, 0.06)	0.43	0.005	1.80
Maximum	-0.001	0.013	(-0.03, 0.03)	0.97	0.014	1.20
BCF	0.005	0.015	(-0.02, 0.03)	0.73	0.008	1.00
LVCF	0.009	0.013	(-0.02, 0.04)	0.49	0.004	1.20
Regression*	0.005	0.012	(-0.02, 0.03)	0.70	0.008	1.00
SF12 physical component score (N=1294)						
Observed	**-0.10**	0.56	**(-1.19, 1.00)**	0.86	-	-
Mean	0.03	0.52	(-0.98, 1.05)	0.95	0.13	0.93
Minimum	0.70	0.88	(-1.03, 2.44)	0.43	0.80	1.57
Maximum	-0.44	0.72	(-1.84, 1.00)	0.54	0.34	1.30
BCF	0.24	0.56	(-0.86, 1.35)	0.67	0.34	1.01
LVCF	-0.13	0.57	(-1.24, 0.98)	0.82	0.03	1.01
Regression*	0.05	0.52	(-0.97, 1.06)	0.92	0.15	0.93
SF12 mental component score (N=1294)						
Observed	**0.29**	0.52	**(-0.74, 1.31)**	0.58	-	-
Mean	0.001	0.48	(-0.95, 0.95)	0.99	0.29	0.93
Minimum	0.70	0.94	(-1.15, 2.55)	0.46	0.41	1.80
Maximum	-0.44	0.70	(-1.81, 0.93)	0.53	0.73	1.34
BCF	0.30	0.50	(-0.67, 1.27)	0.54	0.01	0.95
LVCF	0.07	0.53	(-0.97, 1.11)	0.90	0.22	1.01
Regression*	0.001	0.48	(-0.95, 0.95)	0.99	0.29	0.93
OKS (N=1091)						
Observed	**0.27**	0.57	**(-0.86, 1.39)**	0.64	-	-
Mean	0.13	0.52	(-0.89, 1.15)	0.81	0.14	0.91
Minimum	0.86	0.91	(-0.93, 2.64)	0.35	0.59	1.59
Maximum	-0.16	0.60	(-1.33, 1.01)	0.79	0.43	1.04
BCF	0.52	0.66	(-0.78, 1.81)	0.44	0.25	1.15
LVCF	0.21	0.58	(-0.93, 1.36)	0.72	0.06	1.02
Regression*	0.15	0.52	(-0.88, 1.17)	0.78	0.12	0.91

*On age, sex and ASA grade

7.5.2 Multiple imputation of reminder response data

The results of the ANCOVA analysis after multiple imputation are shown in Table 7.13. For the EQ5D outcome, MCMC on all missing data was the least biased but the precision was not as good as other methods. This was also the best method for the OKS outcome as an unbiased estimate was produced. MCMC followed by

PMM (including QoL) was the least biased for the SF12 physical component scores and also showed good precision as ratio = 1.03.

Table 7.13: KAT – ANCOVA results under multiple imputation of reminder response data

	Two year treatment comparison					
	Difference	SE	95% CI	p-value	Bias	Ratio of CI width
EQ5D (N=1321)						
Observed data	**0.013**	**0.013**	**(-0.014, 0.039)**	**0.35**	**-**	**-**
MCMC all	0.011	0.015	(-0.019, 0.042)	0.47	0.002	1.15
MCMC +regression*	0.005	0.015	(-0.024, 0.034)	0.74	0.008	1.09
MCMC + regression**	0.014	0.016	(-0.019, 0.043)	0.42	0.001	1.00
MCMC + PMM*	0.004	0.014	(-0.024, 0.032)	0.79	0.009	1.06
MCMC + PMM**	0.010	0.015	(-0.019, 0.039)	0.48	0.003	1.09
MCMC + propensity*	0.007	0.019	(-0.033, 0.047)	0.71	0.006	1.51
MCMC + propensity**	0.004	0.014	(-0.022, 0.031)	0.74	0.009	1.00
SF12 physical component score (N=1294)						
Observed data	**-0.10**	**0.56**	**(-1.19, 1.00)**	**0.86**	-	-
MCMC all	-0.03	0.62	(-1.25, 1.20)	0.97	0.07	1.12
MCMC +regression*	-0.19	0.69	(-1.58, 1.21)	0.79	0.09	1.27
MCMC + regression**	-0.16	0.66	(-1.47, 1.16)	0.81	0.06	1.20
MCMC + PMM*	0.29	0.61	(-0.92, 1.49)	0.64	0.39	1.10
MCMC + PMM**	-0.07	0.57	(-1.19, 1.06)	0.91	0.03	1.03
MCMC + propensity*	0.03	0.57	(-1.09, 1.16)	0.95	0.13	1.03
MCMC + propensity**	0.28	0.64	(-1.00, 1.55)	0.66	0.38	1.16
SF12 mental component score (N=1294)						
Observed data	**0.29**	**0.52**	**(-0.74, 1.31)**	**0.58**	-	-
MCMC all	0.02	0.60	(-1.18, 1.22)	0.98	0.27	1.17
MCMC +regression*	-0.14	0.66	(-1.47, 1.19)	0.83	0.44	1.30
MCMC + regression**	0.11	0.63	(-1.15, 1.37)	0.86	0.18	1.23
MCMC + PMM*	-0.19	0.60	(-1.39, 1.01)	0.76	0.48	1.17
MCMC + PMM**	0.25	0.53	(-0.80, 1.30)	0.64	0.04	1.02
MCMC + propensity*	-0.05	0.57	(-1.17, 1.07)	0.93	0.34	1.09
MCMC + propensity**	0.27	0.63	(-0.99, 1.53)	0.67	0.02	1.23
Oxford Knee Score (N=1091)						
Observed data	**0.27**	**0.57**	**(-0.86, 1.39)**	**0.64**	**-**	**-**
MCMC +regression*	0.27	0.62	(-0.94, 1.49)	0.66	0.00	1.08
MCMC + regression**	-0.02	0.63	(-1.27, 1.23)	0.98	0.29	1.11
MCMC + PMM*	0.19	0.61	(-1.01, 1.39)	0.75	0.08	1.07
MCMC + PMM**	-0.05	0.59	(-1.21, 1.11)	0.93	0.33	1.03
MCMC + propensity*	0.16	0.6	(-1.01, 1.33)	0.79	0.11	1.04
MCMC + propensity**	-0.05	0.69	(-1.44, 1.35)	0.96	0.32	1.24
MCMC +regression*	0.10	0.63	(-1.13, 1.33)	0.87	0.17	1.09

* any-readmissions, place of arthritis, further knee surgery; ** same as * plus previous QoL

For the mental component score, a propensity model (including QoL) following MCMC showed the smallest bias, but the precision was not very good. A better method was a PMM (including QoL) as although there was a slight increase in bias, the precision ratio was 1.02 and, therefore, much closer to the desired value of one. For both scores a number of methods provided estimates in the opposite direction, but no statistical difference between treatment groups was found.

7.5.3 Imputation of actual missing data

The best simple and multiple imputation methods were identified in the previous sections. Table 7.14 presents the results of the ANCOVA having imputed the actual missing data using these identified best methods.

Table 7.14: KAT - ANCOVA results after imputation on all missing data

		Two year treatment comparison			
Method	N	Mean Difference	SE	95% CI	p-value
EQ5D					
Observed Data	1321	0.013	0.013	(-0.014, 0.039)	0.35
LVCF	1598	0.014	0.013	(-0.012, 0.039)	0.29
MCMC + regression**	1715	0.015	0.013	(-0.01, 0.04)	0.26
SF12 physical component score (PCS)					
Observed Data	1294	-0.10	0.56	(-1.19, 1.00)	0.86
LVCF	1572	0.16	0.51	(-0.83, 1.16)	0.75
MCMC + PMM**	1715	0.09	0.52	(-0.93, 1.11)	0.87
SF12 mental component score (MCS)					
Observed Data	1294	0.29	0.52	(-0.74, 1.31)	0.58
Baseline CF	1572	0.34	0.45	(-0.55, 1.22)	0.45
MCMC then PMM**	1715	0.38	0.51	(-0.62, 1.38)	0.46
OKS					
Observed Data	1091	0.27	0.57	(-0.86, 1.39)	0.64
LVCF	1591	0.40	0.51	(-0.60, 1.40)	0.43
MCMC for intermittent	1715	0.40	0.50	(-0.58, 1.38)	0.42

** any-readmissions, place of arthritis, further knee surgery, previous QoL;

The results under the different imputation strategies are reasonably consistent between the observed data, LVCF/BCF and MCMC multiple imputation. Although the conclusion of the significant test is not altered between strategies for the remaining QoL scores, there is more variation in calculated treatment difference and standard errors. For the PCS, the observed treatment difference was negative, but under imputation it was positive.

7.5.4 Summary

The LVCF method was the most suitable simple imputation procedure for three of the four QoL scores. For the PCS, baseline carried forwards was preferred. MCMC for intermittent missingness was the best MI method for the EQ5D and OKS outcomes, while MCMC to make the data monotone followed by a PMM model was better for the SF12 physical and mental component scores respectively. The best MI method tended to provide better point estimates, SE's and subsequent confidence intervals than the best simple imputation method, when comparing to the observed data. Undertaking imputation on the actual missing data provided consistent results to that observed for the EQ5D data, but not for the SF12 and OKS scores.

7.6 PRISM

The SF36 physical and mental component scores, Arthritis Index and EQ5D were the QoL scores collected at baseline and yearly up to four years. Analysis was conducted to assess treatment difference at two years adjusting for the minimisation variables (using binary indicators) and baseline QoL. The binary indicators were: serum alkaline phosphatase level (normal, elevated or greatly elevated); previous treatment with bisphosphonates; Paget's in a weight bearing limb; Paget's in the skull; deformity due to Paget's; pain perceived due to Paget's.

7.6.1 Simple imputation of reminder response data

The trial analysis looked at two year treatment differences. Therefore, although data was collected at years three and four, no imputation was carried out for these assessments. If data was present at these assessments, it could be used as part of an imputation procedure for the 24 month assessment. The results of the ANCOVA analysis after simple imputation are shown in Table 7.15. For the EQ5D score maximum value imputation showed the least bias but was not the best procedure in terms of precision. LVCF or NVCB were good for both bias and precision.

Table 7.15: PRISM – ANCOVA results after simple imputation for reminder response data

	Two year treatment difference					
	Difference	SE	95% CI	p-value	Bias	Ratio of CI width
EQ5D (N=845)						
Observed Data	**0.015**	**0.017**	**(-0.019, 0.049)**	**0.38**	**-**	**-**
Mean	0.003	0.016	(-0.029, 0.034)	0.86	0.012	0.93
Maximum	0.019	0.019	(-0.019, 0.057)	0.32	0.004	1.12
Minimum	-0.045	0.035	(-0.113, 0.023)	0.19	0.060	2.00
BCF	0.005	0.016	(-0.026, 0.036)	0.77	0.010	0.91
LVCF	0.006	0.017	(-0.027, 0.039)	0.70	0.009	0.97
NVCB (N=741)	0.010	0.018	(-0.026, 0.045)	0.60	0.005	1.04
Regression*	0.004	0.015	(-0.027, 0.034)	0.82	0.011	0.90
SF36 Physical component score (N=798)						
Observed Data	**0.14**	**0.53**	**(-0.90, 1.17)**	**0.80**	**-**	**-**
Mean	0.17	0.53	(-0.86, 1.21)	0.74	0.03	1.00
Maximum	0.99	0.85	(-0.68, 2.66)	0.24	0.85	1.61
Minimum	-0.88	1.00	(-2.83, 1.08)	0.38	1.02	1.89
BCF	-0.14	0.49	(-1.10, 0.82)	0.78	0.28	0.93
LVCF	-0.19	0.52	(-1.21, 0.82)	0.71	0.33	0.98
NVCB (N=696)	0.014	0.57	(-1.10, 1.12)	0.98	0.126	1.07
Regression*	-0.071	0.48	(-1.02, 0.87)	0.88	0.211	0.91
SF36 mental component score (N=798)						
Observed Data	**0.69**	**0.68**	**(-0.65, 2.04)**	**0.31**	**-**	**-**
Mean	0.46	0.65	(-0.83, 1.72)	0.49	0.23	0.95
Maximum	1.20	0.90	(-0.58, 2.97)	0.19	0.51	1.32
Minimum	-0.70	1.11	(-2.88, 1.49)	0.53	1.39	1.62
BCF	0.71	0.64	(-0.54, 1.96)	0.26	0.02	0.93
LVCF	0.47	0.68	(-0.86, 1.81)	0.49	0.22	0.99
NVCB (N=696)	0.97	0.74	(-0.48, 2.42)	0.19	0.28	1.08
Regression*	0.57	0.62	(-0.65, 1.80)	0.36	0.12	0.91
Arthritis Index (N=798)						
Observed Data	**-0.04**	**0.60**	**(-1.22, 1.14)**	**0.95**	**-**	**-**
Mean	-0.10	0.60	(-1.27, 1.08)	0.87	0.06	0.99
Maximum	0.80	0.94	(-1.05, 2.64)	0.40	0.84	1.56
Minimum	-0.94	0.90	(-2.71, 0.82)	0.30	0.90	1.50
BCF	-0.26	0.55	(-1.35, 0.83)	0.64	0.22	0.92
LVCF	-0.39	0.59	(-1.54, 0.76)	0.51	0.35	0.97
NVCB (N=696)	-0.12	0.64	(-1.38, 1.14)	0.85	0.08	1.07
Regression*	-0.23	0.55	(-1.31, 0.84)	0.67	0.19	0.91

*On baseline QoL, age and whether or not a patient had previous bisphosphonate treatment

For both the PCS and Arthritis Index, mean value imputation was least biased and showed good comparative precision. Finally for the MCS, baseline carried forwards was the best imputation method. The treatment difference calculated

under imputation varied in direction for the PCS, although the conclusion of no treatment difference was consistent.

7.6.2 Multiple imputation of reminder response data

In addition to the variables in the analytic model, age was found to be associated with both outcome and missingness and therefore included in the imputation model. The results of the ANCOVA after various MI procedures are shown in Table 7.16. For each of the EQ5D and MCS, the least biased MI method was MCMC to make the data monotone, followed by a PMM model on the remaining missing data. The analysis variables and age were required for the imputation model. Despite a greater bias, the precision was improved using QoL in the imputation model. In the case of the PCS and Arthritis index, the least biased MI method was MCMC followed by a propensity model (including both analytic and previous QoL terms). However, the precision was greater for this model than for the other MI methods.

Table 7.16: PRISM – ANCOVA results after MI of reminder response data

	Two year treatment comparison					
	Difference	SE	95% CI	p-value	Bias	Ratio of CI width
EQ5D (N=845)						
Observed data	**0.015**	**0.017**	**(-0.019, 0.049)**	**0.38**	-	-
MCMC for intermittent	0.006	0.018	(-0.030, 0.042)	0.74	0.009	1.06
MCMC + regression*	0.005	0.019	(-0.032, 0.042)	0.78	0.010	1.09
MCMC + regression**	0.002	0.018	(-0.033, 0.038)	0.90	0.013	1.04
MCMC + PMM*	0.012	0.021	(-0.031, 0.055)	0.57	0.003	1.26
MCMC + PMM**	0.009	0.018	(-0.023, 0.044)	0.64	0.006	0.99
MCMC + propensity*	0.004	0.019	(-0.033, 0.041)	0.84	0.011	1.09
MCMC + propensity**	0.006	0.018	(-0.030, 0.042)	0.73	0.009	1.06
SF36 physical component score (N=798)						
Observed data	**0.14**	**0.53**	**(-0.89, 1.17)**	**0.80**	-	-
MCMC for intermittent	-0.22	0.56	(-1.32, 0.89)	0.70	0.36	1.07
MCMC + regression*	0.31	0.73	(-1.17, 1.80)	0.67	0.17	1.44
MCMC + regression**	-0.05	0.59	(-1.22, 1.12)	0.94	0.19	1.14
MCMC + PMM*	0.39	0.61	(-0.81, 1.59)	0.52	0.25	1.17
MCMC + PMM**	-0.01	0.59	(-1.18, 1.16)	0.99	0.15	1.14
MCMC + propensity*	0.03	0.78	(-1.57, 1.63)	0.97	0.11	1.55
MCMC + propensity**	0.29	0.62	(-0.93, 1.50)	0.64	0.15	1.18
SF36 Mental component score (N=798)						
Observed data	**0.70**	**0.69**	**(-0.65, 2.04)**	**0.31**	-	-
MCMC for intermittent	0.87	0.70	(-0.50, 2.23)	0.21	0.17	1.01
MCMC + regression*	0.77	0.97	(-1.24, 2.77)	0.44	0.07	1.49
MCMC + regression**	0.56	0.74	(-0.89, 2.01)	0.45	0.14	1.08
MCMC + PMM*	0.65	0.75	(-0.83, 2.12)	0.39	0.05	1.10
MCMC + PMM**	0.88	0.72	(-0.53, 2.29)	0.22	0.18	1.05
MCMC + propensity*	0.40	0.75	(-1.07, 1.88)	0.59	0.30	1.10
MCMC + propensity**	0.30	0.81	(-1.31, 1.90)	0.72	0.40	1.19
Arthritis Index (N=798)						
Observed data	**-0.04**	**0.60**	**(-1.22, 1.14)**	**0.95**	-	-
MCMC for intermittent	-0.51	0.67	(-1.83, 0.82)	0.45	0.47	1.12
MCMC + regression*	0.23	0.92	(-1.68, 2.15)	0.80	0.27	1.62
MCMC + regression**	-0.48	0.64	(-1.74, 0.79)	0.46	0.44	1.07
MCMC + PMM*	0.38	0.75	(-1.10, 1.86)	0.61	0.42	1.25
MCMC + PMM**	-0.28	0.63	(-1.52, 0.96)	0.66	0.24	1.05
MCMC + propensity*	-0.19	0.74	(-1.66, 1.28)	0.80	0.15	1.25
MCMC + propensity**	-0.16	0.84	(-1.88, 1.55)	0.85	0.12	1.45

* MI model based on the ANCOVA model and age;

** MI model based on the ANCOVA model, age and previous QoL;

7.6.3 Imputation of actual missing data

The least biased simple and multiple imputation methods were identified after imputation of reminder response data. Using these methods, the actual missing data was imputed and the ANCOVA carried out. The results are shown in Table 7.17.

Table 7.17: PRISM – ANCOVA after imputation on all missing data

		Two year treatment comparison			
Method	N	Difference	SE	95% CI	p-value
		EQ5D			
Observed Data	845	0.015	0.017	(-0.019, 0.049)	0.38
Maximum value	1250	0.018	0.017	(-0.015, 0.051)	0.29
MCMC + PMM*	1324	0.015	0.019	(-0.023, 0.052)	0.44
		PCS			
Observed Data	798	0.14	0.53	(-0.90, 1.17)	0.80
Mean	1198	0.05	0.41	(-0.76, 0.86)	0.90
MCMC + propensity*	1324	0.13	0.55	(-0.96, 1.21)	0.82
		MCS			
Observed Data	798	0.69	0.68	(-0.65, 2.04)	0.31
Baseline CF	1198	0.55	0.47	(-0.38, 1.48)	0.24
MCMC + PMM*	1324	0.61	0.78	(-0.99, 2.20)	0.44
		Arthritis Index			
Observed Data	798	-0.04	0.60	(-1.22, 1.14)	0.95
Mean	1198	0.025	0.47	(-0.90, 0.95)	0.96
MCMC + propensity*	1324	-0.18	0.82	(-1.95, 1.58)	0.83

* MI model based on the ANCOVA model and age; PCS –physical component score; MCS – mental components score

Despite the differing sample sizes, the results under imputation are fairly close to that which was observed for the EQ5D score. The results are slightly less consistent in terms of the calculated treatment difference for the SF12 MCS and least consistent for the Arthritis Index. However, despite the variation in calculated treatment difference and standard error under each data scenario, there was no significant treatment differences found in the QoL scores at two years.

7.6.4 Summary

There was no single simple imputation method preferred for the four QoL scores. MCMC imputation to make the data monotone followed by a PMM was the most suitable MI procedure. On the whole, simple imputation performed better than MI when imputing reminder responses. Undertaking imputation on the actual missing data resulted in fairly consistent results to that which was observed and

maintained the conclusion of no treatment difference with respect to the two year QoL data.

7.7 TOMBOLA

The QoL measure in the TOMBOLA trial was the EQ5D. As previously mentioned there were two treatment comparisons within the TOMBOLA trial: comparison one - colposcopy versus cytological surveillance; comparison two within the colposcopy arm – immediate treatment versus biopsy and recall. The trial itself did not perform an analysis of the EQ5D scores, but for illustration an ANCOVA adjusting for baseline score, age group (20-29,30-39,40-49,50-59), trial centre (Grampian, Tayside, Nottingham), eligible smear status (mild, BNA), and HPV status (negative, positive, no sample) was carried out. Table 3.22 showed the results from these analyses. There was a significant difference in the 12 month EQ5D scores between those who received immediate treatment after colposcopy and those who were selective recall after their colposcopy (p=0.03).

7.7.1 Simple imputation of reminder response data

Patients who responded at both the endpoints (12 or 30 months) and at baseline were included in the analysis. Table 7.18 shows the results of the ANCOVA analysis after simple imputation of the reminder response data. There were 2294 (67%) patients involved in the first comparison (colposcopy versus cytological surveillance), who provided scores at baseline and 12 months, with 601 (26%) responding by reminder. The least biased methods were LVCF and BCF. These methods showed reasonable precision, although they did under-estimate the standard error. At 30 months, there were 1825 (54%) patients who provided scores, with 476 (26%) responding by reminder. The observed data showed no treatment difference, which was maintained by all methods (Table 7.18). The least biased methods were baseline carried forwards and regression (b = 0.001), with BCF having the better precision value.

Table 7.18: TOMBOLA - ANCOVA results after simple imputation of reminder response data

Method	Treatment Difference	SE	95% CI	p-value	Bias	Ratio of CI width
12 months: Colposcopy vs. cytological surveillance (N=2294)						
Observed Data	**-0.0012**	**0.005**	**(-0.011, 0.009)**	**0.81**	**-**	**-**
Mean	0.0001	0.004	(-0.009, 0.009)	0.98	0.0013	0.90
Maximum	0.0006	0.005	(-0.009, 0.010)	0.90	0.0018	0.95
Minimum	-0.005	0.018	(-0.040, 0.031)	0.80	0.0038	3.55
BCF	-0.001	0.004	(-0.009, 0.008)	0.88	0.0002	0.85
LVCF	-0.001	0.005	(-0.010, 0.008)	0.87	0.0002	0.90
NVCB (N=2187)	-0.0008	0.005	(-0.011, 0.009)	0.87	0.0004	1.00
Regression*	-0.0008	0.004	(-0.009, 0.007)	0.84	0.0004	0.80
30 months: Colposcopy vs. cytological surveillance (N=1825)						
Observed Data	**0.004**	**0.006**	**(-0.007, 0.015)**	**0.49**	**-**	**-**
Mean	0.003	0.005	(-0.007, 0.013)	0.54	0.001	0.91
Maximum	0.001	0.006	(-0.010, 0.012)	0.85	0.003	1.00
Minimum	0.021	0.019	(-0.015, 0.058)	0.25	0.017	3.32
BCF	0.004	0.005	(-0.006, 0.014)	0.43	0.000	0.91
LVCF	0.001	0.006	(-0.010, 0.013)	0.80	0.003	1.05
Regression*	0.004	0.005	(-0.006, 0.013)	0.43	0.000	0.86
12 months: Immediate treatment vs. biopsy and recall (N=709)						
Observed Data	**0.019**	**0.008**	**(0.002, 0.035)**	**0.03**	**-**	**-**
Mean	0.012	0.007	(-0.003, 0.027)	0.11	0.007	0.81
Maximum	0.014	0.008	(-0.002, 0.030)	0.09	0.005	0.86
Minimum	-0.004	0.032	(-0.067, 0.059)	0.90	0.023	3.41
BCF	0.014	0.007	(-0.001, 0.028)	0.06	0.005	0.78
LVCF	0.014	0.008	(-0.001, 0.029)	0.07	0.005	0.81
NVCB (N=681)	0.015	0.008	(-0.002, 0.032)	0.08	0.004	0.92
Regression*	0.012	0.007	(-0.002, 0.026)	0.09	0.007	0.76
30 months: Immediate treatment vs. biopsy and recall (N=584)						
Observed Data	**0.006**	**0.010**	**(-0.014, 0.026)**	**0.56**	**-**	**-**
Mean	0.010	0.009	(-0.007, 0.028)	0.25	0.004	0.88
Maximum	0.010	0.010	(-0.010, 0.029)	0.31	0.004	0.98
Minimum	0.015	0.032	(-0.048, 0.079)	0.64	0.009	3.18
BCF	0.010	0.009	(-0.008, 0.027)	0.27	0.004	0.88
LVCF	0.009	0.010	(-0.010, 0.028)	0.35	0.003	0.95
Regression*	0.009	0.009	(-0.008, 0.026)	0.30	0.003	0.85

* Baseline QoL, age group, smoking status, employment status and marital status

In the second comparison (immediate treatment versus biopsy and selective recall) at 12 months, 709 patients provided a score with 178 (25%) after reminder. The least biased method was NVCB. This had precision ratio equal to 0.92 (Table 7.18). In the observed data, a treatment difference was found (p=0.027), but this was altered under imputation to borderline evidence (p=0.075 for NVCB). At 30 months, 584 patients provided an EQ5D score with 143 (24%) responding by reminder. The least biased method was LVCF (b = 0.003) with a good precision value (ratio=0.95). Across each of the treatment comparisons and across both

endpoints, the least biased method involved other QoL scores observed for each patient.

7.7.2 Multiple imputation of reminder response data

Table 7.19 shows the results of the ANCOVA analyses after multiple imputation.

Table 7.19: TOMBOLA - ANCOVA results after multiple imputation of reminder response data

EQ5D	Treatment Difference	SE	95% CI	p-value	Bias	Ratio of CI width
12 months: Colposcopy vs. cytological surveillance (N=2294)						
Observed	**-0.0012**	**0.005**	**(-0.011, 0.008)**	**0.81**	**-**	**-**
MCMC	0.0008	0.006	(-0.011, 0.013)	0.89	0.0020	1.26
MCMC +regression*	0.001	0.006	(-0.011, 0.012)	0.91	0.0022	1.21
MCMC + regression**	0.001	0.006	(-0.011, 0.013)	0.91	0.0022	1.26
MCMC + PMM*	-0.001	0.005	(-0.011, 0.010)	0.88	0.0002	1.11
MCMC + PMM**	0.002	0.006	(-0.012, 0.015)	0.80	0.0032	1.42
MCMC + propensity*	0.001	0.005	(-0.009, 0.012)	0.76	0.0022	1.11
MCMC + propensity**	-0.0001	0.006	(-0.012, 0.012)	0.98	0.0011	1.26
30 months: Colposcopy vs. cytological surveillance (N=1825)						
Observed	**0.004**	**0.006**	**(-0.007, 0.015)**	**0.49**	**-**	**-**
MCMC	0.006	0.006	(-0.007, 0.018)	0.38	0.002	1.14
MCMC +regression*	0.010	0.008	(-0.008, 0.027)	0.26	0.006	1.59
MCMC + regression**	0.003	0.006	(-0.009, 0.015)	0.66	0.001	1.09
MCMC + PMM*	0.001	0.006	(-0.012, 0.013)	0.90	0.003	1.14
MCMC + PMM**	0.003	0.006	(-0.009, 0.016)	0.60	0.001	1.14
MCMC + propensity*	0.003	0.007	(-0.011, 0.016)	0.70	0.001	1.23
MCMC + propensity**	0.006	0.009	(-0.012, 0.024)	0.51	0.002	1.64
12 months: Immediate treatment vs. biopsy and recall (N=709)						
Observed	**0.019**	**0.008**	**(0.002, 0.035)**	**0.03**	**-**	**-**
MCMC	0.016	0.007	(-0.004, 0.035)	0.12	0.003	1.18
MCMC +regression*	0.011	0.009	(-0.008, 0.029)	0.25	0.008	1.12
MCMC + regression**	0.015	0.010	(-0.006, 0.035)	0.15	0.004	1.24
MCMC + PMM*	0.018	0.010	(-0.011, 0.028)	0.39	0.001	1.18
MCMC + PMM**	0.012	0.009	(-0.006, 0.030)	0.20	0.007	1.09
MCMC + propensity*	0.011	0.009	(-0.008, 0.029)	0.26	0.008	1.12
MCMC + propensity**	0.009	0.010	(-0.011, 0.029)	0.38	0.010	1.21
30 months: Immediate treatment vs. biopsy and recall (N=584)						
Observed	**0.006**	**0.010**	**(-0.014, 0.026)**	**0.56**	-	-
MCMC	0.014	0.011	(-0.007, 0.035)	0.18	0.008	1.05
MCMC +regression*	0.009	0.013	(-0.018, 0.037)	0.48	0.003	1.38
MCMC + regression**	0.007	0.011	(-0.015, 0.029)	0.54	0.007	1.10
MCMC + PMM*	0.015	0.011	(-0.006, 0.037)	0.17	0.009	1.08
MCMC + PMM**	0.011	0.011	(-0.010, 0.032)	0.29	0.005	1.05
MCMC + propensity*	0.011	0.013	(-0.015, 0.037)	0.38	0.005	1.30
MCMC + propensity**	0.009	0.011	(-0.014, 0.031)	0.44	0.003	1.13

*treatment, age group, trial centre, eligible smear, HPV; ** same as * plus previous QoL

The least biased MI method for the 12-month treatment comparison (colposcopy versus cytological surveillance) was MCMC to make the data monotone, followed by a PMM model (treatment group, age group, trial centre, eligible smear status

and HPV status). The bias was very small (0.0002) and the method showed reasonable precision compared to the observed data (ratio = 1.11). At 30 months, MCMC imputation followed by a regression model (covariates plus QoL) was the best method. In the second comparison (immediate treatment versus biopsy and selective recall), a PMM model was best at 12 months. Including QoL increased the bias slightly, but gave a better estimate of precision. At 30 months, MCMC, followed by a PMM (covariates and QoL) was the best method. Although not the least biased, it was the most precise method.

7.7.3 Imputation of actual missing data

The least biased simple imputation and multiple imputation methods were identified in the previous two sections using the data collected after reminder. These methods were used to impute all missing data and the results are shown in Table 7.20.

Table 7.20: TOMBOLA – ANOCVA results after imputation of all missing data

Method	Treatment Difference	SE	95% CI	p-value	N
12 months: Colposcopy versus cytological surveillance					
Observed Data	-0.0012	0.005	(-0.011, 0.009)	0.81	2294
BCF	-0.002	0.004	(-0.009, 0.004)	0.48	3300
MCMC + PMM*	0.002	0.006	(-0.010, 0.014)	0.73	3381
30 months: Colposcopy versus cytological surveillance					
Observed Data	0.004	0.006	(-0.007, 0.015)	0.49	1825
BCF	0.0006	0.003	(-0.006, 0.007)	0.85	3300
MCMC + regression **	0.005	0.006	(-0.007, 0.018)	0.36	3381
12 months: Immediate treatment vs. biopsy and recall					
Observed Data	0.019	0.008	(0.002, 0.035)	0.03	709
NVCB	0.015	0.008	(-0.001, 0.031)	0.07	813
MCMC + PMM**	0.019	0.008	(0.002, 0.035)	0.02	986
30 months: Immediate treatment vs. biopsy and recall					
Observed Data	0.006	0.010	(-0.014,0.026)	0.56	584
LVCF	0.005	0.008	(-0.010,0.020)	0.51	966
MCMC + PMM**	0.007	0.010	(-0.013,0.026)	0.51	986

*treatment, age group, trial centre, eligible smear, HPV; ** same as * plus previous QoL

The slightly reduced N arose because age group was missing for 18 patients in total, which meant they could not be included in the analysis. It is seen that different data strategies have an impact on the calculated treatment difference. For the 12 month colposcopy versus cytological surveillance comparison, there

was no change in conclusion of the significance test under the different analysis strategies ($p>0.05$). The same occurs for the second treatment comparison, where no significant difference between treatment groups was found.

The comparison of the immediate treatment versus biopsy and selective recall at 12 months was the most interesting. In the observed data, a significant treatment difference ($p=0.03$) was found. Under MI, using a PMM model this difference was maintained ($p<0.05$). However, using NVCB changed the conclusion to be borderline evidence ($p=0.07$) of a difference in QoL between treatment groups. The identified treatment difference was reduced although the standard error was the same as that which was observed.

7.7.4 Summary

The best simple and best multiple imputation methods performed equally. This was perhaps because the MI models did not include any other additional variables other than those in the analysis models. They are of most use when additional variables relating to outcome and missingness are used. LVCF or NVCB were the best simple imputation methods and MCMC for all missing data was best in both treatment comparisons at 30 months. Using imputation on all the actual missing data had the biggest impact at 12 months for the second treatment comparison.

7.8 Norwegian Palliative Care (NPC) Trial

The NPC trial employed the QLQ-C30 questionnaire. Of particular interest were the pain, physical functioning and emotional functioning dimension scores. The assessment of most interest was at four months adjusting for baseline scores, sex, age group, cluster pair and baseline Karnofsky performance index. No significant differences were found between treatment groups for any of the three QoL dimensions at each of the follow up assessments (Table 3.26).

7.8.1 Simple imputation of reminder response data

Table 7.21 shows the results of the ANCOVA to determine the treatment difference at four months for the three QoL dimensions. The 'best' method for the pain score was mean value imputation. Although the correct direction (and non-significance) was retained under imputation of the physical functioning scores, the magnitude of the estimate varied considerably. The least biased method was mean imputation for both the physical functioning and emotional functioning scores. Although not perfect, the precision under mean imputation was reasonable with the ratio equal to 0.91 and 0.95 respectively.

Table 7.21: NPC Trial: ANCOVA results QoL score at four months after simple imputation of reminder response data

	Four month treatment comparison					
	Difference	SE	95%CI	p-value	Bias	Ratio of CI width
Pain (N=136)						
Observed Data	**1.49**	**4.86**	**(-8.12, 11.1)**	**0.76**	-	-
Mean	1.79	4.57	(-7.26, 10.8)	0.70	0.30	0.94
Maximum	-1.75	6.49	(-14.6, 11.1)	0.79	3.24	1.34
Minimum	4.34	5.27	(-6.1, 14.8)	0.41	2.85	1.09
BCF	-1.29	4.59	(-10.4, 7.8)	0.78	2.78	0.95
LVCF	0.37	5.00	(-9.53, 10.3)	0.94	1.12	1.03
Regression*	0.44	4.41	(-8.29, 9.16)	0.92	1.05	0.91
Physical Functioning (N=135)						
Observed Data	**-6.11**	**4.77**	**(-15.6, 3.33)**	**0.20**	-	-
Mean	-5.25	4.34	(-13.8, 3.35)	0.23	0.86	0.91
Maximum	-9.68	5.67	(-20.9, 1.54)	0.09	3.57	1.19
Minimum	-0.08	5.79	(-11.5, 11.4)	0.99	6.03	1.21
BCF	-1.90	4.17	(-10.2, 6.35)	0.65	4.21	0.87
LVCF	-4.62	4.59	(-13.7, 4.47)	0.32	1.49	0.96
Regression*	-2.49	4.05	(-10.5, 5.53)	0.54	3.62	0.85
Emotional Functioning (N=135)						
Observed Data	-3.50	3.16	(-9.74, 2.75)	0.27	-	-
Mean	-2.53	3.00	(-8.47, 3.41)	0.40	0.97	0.95
Maximum	-4.78	3.66	(-12.0, 2.46)	0.19	1.28	1.16
Minimum	4.26	6.40	(-8.41, 16.9)	0.51	7.76	2.03
BCF	0.96	2.81	(-4.60, 6.52)	0.73	4.46	0.89
LVCF	-1.74	3.10	(-7.90, 4.40)	0.58	1.76	0.98
Regression*	-0.38	2.70	(-5.73, 4.97)	0.89	3.12	0.86

* sex, age group, cluster pair, baseline QoL and Karnofsky performance status

7.8.2 Multiple imputation of reminder response data

In addition to those variables in the ANCOVA model, no other variables were found to be associated with missingness and outcome. Treatment and performance status were already included as part of the analysis model. Table

7.22 shows the results of the ANCOVA for the three QoL dimension scores after various MI strategies. The least biased and most precise method for the pain score at four months was MCMC to make the data monotone, followed by a PMM model (analysis variables only). The least biased MI method for the physical functioning scores was a predictive mean match model, yet MCMC imputation provided a more suitable measure of precision. The same phenomenon was seen for the emotional functioning score. There did not appear to be one MI method which was best for all three QoL dimensions.

Table 7.22: NPC Trial: ANCOVA for each QoL score after multiple imputation of reminder-responses

	Four month treatment comparison					
	Difference	SE	95%CI	p-value	Bias	Ratio of CI width
Pain (N=136)						
Observed	**1.49**	**4.86**	**(-8.13, 11.1)**	**0.76**	-	-
MCMC	2.91	6.2	(-9.59, 15.4)	0.64	1.42	1.30
MCMC + regression*	1.96	8.14	(-15.4, 19.3)	0.81	0.47	1.80
MCMC +PMM*	1.64	5.66	(-9.48, 12.8)	0.77	0.15	1.16
MCMC + propensity*	1.25	6.24	(-11.2, 13.7)	0.84	0.24	1.29
MCMC + regression**	3.36	6.28	(-9.28, 16.0)	0.60	1.87	1.31
MCMC + PMM**	1.69	5.97	(-10.2, 13.5)	0.78	0.20	1.23
MCMC + propensity**	3.49	6.15	(-8.80, 15.8)	0.57	2.00	1.28
Physical functioning (N=135)						
Observed	**-6.11**	**4.77**	**(-15.6, 3.33)**	**0.20**	**-**	**-**
MCMC	-6.55	4.8	(-16.0, 2.87)	0.17	0.44	0.99
MCMC + regression*	-7.36	7.65	(-23.8, 9.06)	0.35	1.25	1.74
MCMC +PMM*	-5.01	5.03	(-14.9, 4.86)	0.32	1.10	1.04
MCMC + propensity*	-8.24	6.57	(-21.6, 5.15)	0.22	2.13	1.41
MCMC + regression**	-4.73	5.12	(-14.9, 5.39)	0.36	1.38	1.07
MCMC + PMM**	-5.92	5.76	(-17.4, 5.52)	0.31	0.19	1.21
MCMC + propensity**	-3.81	5.73	(-15.2, 7.59)	0.51	2.30	1.20
Emotional functioning (N=135)						
Observed	**-3.50**	**3.16**	**(-9.74, 2.75)**	**0.27**	**-**	**-**
MCMC	-1.76	3.32	(-8.30, 4.79)	0.60	1.74	1.05
MCMC + regression*	-6.89	4.44	(-15.8, 1.99)	0.13	3.39	1.42
MCMC +PMM*	-2.62	3.84	(-10.3, 5.05)	0.50	0.88	0.99
MCMC + propensity*	-1.07	4.04	(-9.06, 6.91)	0.79	2.43	1.28
MCMC + regression**	-1.32	3.7	(-8.68, 6.04)	0.72	2.18	1.18
MCMC + PMM**	-3.38	4.32	(-12.0, 5.25)	0.44	0.12	1.38
MCMC + propensity**	-0.06	4.42	(-8.88, 8.99)	0.99	3.44	1.43

* sex, age group, cluster pair, baseline Karnofsky performance index; ** same as * plus previous QoL

7.8.3 Imputation of actual missing data

In this section, the identified 'best' simple and MI method for each QoL score were used to impute all the missing data, rather than the reminder data. Table 7.23 shows the results of the ANCOVA under three different analysis strategies: the observed data; observed data plus simple imputation (best method) of the missing data; observed data plus multiple imputation (best method) of the missing data. Using mean imputation for all missing data changed the result to a significant treatment difference in most cases. Multiple imputation tended to provide more consistent results when compared to the observed data, but cannot be directly compared due to the differing numbers involved.

Table 7.23: NPC Trial – ANCOVA after imputation of all missing data

		Four month treatment comparison			
	N	**Difference**	**SE**	**95%CI**	**p-value**
Pain					
Observed data	136	1.49	4.86	(-8.13, 11.1)	0.76
Mean value	433	5.77	1.77	(2.30, 9.24)	0.001
MCMC + PMM*	433	3.75	6.17	(-10.9, 18.4)	0.56
Physical Functioning					
Observed data	135	-6.11	4.77	(-15.6, 3.33)	0.20
Mean value	433	6.06	1.98	(2.17, 9.96)	0.002
MCMC + PMM*	434	-9.79	2.98	(-15.7, -3.92)	0.001
Emotional functioning					
Observed data	132	-6.11	4.77	(-15.6,3.33)	0.20
Mean value	433	9.32	1.78	(5.82,12.8)	<0.001
MCMC + PMM**	434	-1.09	3.8	(-10.3,8.08)	0.78

* sex, age group, cluster pair, baseline Karnofsky performance index; ** same as * plus previous QoL

7.8.4 Summary

In this dataset, neither simple nor multiple imputation was consistently better for the imputation of the reminder data. When the best SI or best MI method were applied to all missing data, the conclusion of treatment difference changed. Simple imputation on the whole suggested a significant difference in QoL scores between treatment groups while the observed data and that under MI did not. This highlights the issue of why imputation should be used with caution. One major flaw in the imputation methods considered above is that all patients are

assumed to be alive and thus, have QoL (however good or bad). In this particular sample, many had died during the course of the trial and therefore should perhaps be treated differently. One possibility is to restrict analysis to those who were still alive at four months, but did not provide questionnaires. Alternatively, one could impute a value of zero for death but there is concern over whether this is appropriate. This was not considered further here, but it would be of interest to investigate this issue in future work. The large number of deaths was considered further in the model-based strategies in chapter nine.

7.9 Discussion

The aim of the work presented in this chapter was to determine the accuracy of a number of imputation strategies for missing QoL outcomes. In each of the seven datasets and for each of the QoL scores, a range of simple and multiple imputation procedures were carried out. Initially, to allow the accuracy of imputation to be assessed, it was the scores obtained through reminders which were imputed, rather than the actual missing values. The analysis method originally used by the trial researchers was then implemented on the imputed datasets and an estimate of treatment difference obtained. This was then compared to the observed result from the data which included all the responders (both immediate and reminder). A measure of bias and precision were used to assess which imputation was most accurate.

Table 7.24: Summary of the 'best' imputation methods for each trial

Trial	QoL Measure	Simple imputation	Multiple imputation
REFLUX	RQLS	LVCF	MI – regression
	EQ5D	LVCF	MI – predictive mean match
	SF12	LVCF	MI - regression
MAVIS	EQ5D	BCF	MI – predictive mean match
	SF12	Mean/BCF	MI – predictive mean match
RECORD	EQ5D	LVCF	MI- MCMC
	SF12	LVCF/BCF	MI - MCMC
KAT	OKS	LVCF	MI - MCMC
	EQ5D	LVCF	MI - regression
	SF12	LVCF	MI – predictive mean match
PRISM	EQ5D	Max	MI – predictive mean match
	SF36	Mean	MI – predictive mean match
TOMBOLA	EQ5D	BCF/LVCF	MI – predictive mean match
NPC Trial	QLQ-C30	Mean	MI – predictive mean match

Table 7.24 summarises the best simple and multiple imputation methods identified for each trial. The best simple imputation method tended to be one of mean imputation, LVCF or BCF. This was surprising, as it is well known in the literature that LVCF/BCF are not recommended, as they make strong assumptions about the stability of the QoL (Carpenter, Kenward 2007, Gadbury, Coffey & Allison 2003). Mean imputation is also known to be problematic in that it reduces the standard error and ultimately impacts on confidence intervals and p-values. In the majority of cases, MCMC imputation or predictive mean match model following MCMC imputation to make the data monotone were the most accurate MI methods.

Comparing between simple and multiple imputation showed that there was no one overall best method. The most appropriate choice will be determined by the data in question and what other variables are available to put into the imputation model. Within a particular trial dataset, although simple imputation might be the better choice for some of the QoL scores, on the whole MI was the better option. The standard errors under MI tended to reflect those calculated in the observed dataset, meaning the confidence interval under imputation showed equivalent width to that of the observed data. This resulted in a value of the precision estimate closer to the desired value of one. Appropriate standard error values are very important when calculating the confidence intervals and p-values for significance of treatment difference (Molenberghs, Kenward 2007).

The REFLUX trial contained the most missing data (or in this case data collected via reminder-response). MCMC to make the data monotone followed by a regression model (including covariates and previous QoL) was clearly the most superior over the simple imputation methods. The difference between simple imputation and MI is less obvious, due to the fact that the amount of data being imputed is reduced. For example, in the MAVIS trial only 12% of data is undergoing imputation and in this trial one of each of the simple and multiple imputation methods were equivalent in providing the smallest bias.

The simple imputation methods assume the data is missing completely at random (MCAR). That is to say, the data are missing for a reason unrelated to anything you have observed. MI methods assume missing at random (MAR), which assumes missing data, are related to observed data (covariates and/or outcome) (Fairclough 2002). If there is no evidence for MCAR, then simple imputation methods should not be used and in the case that they are used, this should be done so with caution.

Chapter five showed that the missing data was more likely not MCAR, suggesting a reason why the simple imputation methods were not as good in these trials. In some cases, the mechanism of missing data was shown to be MCAR. It was under these circumstances that a simple imputation method was more accurate than an MI method. For example, in the TOMBOLA trial there was no evidence against the MCAR mechanism and BCF or LVCF was found to be favoured over an MI procedure. In contrast, in the case of the RECORD and KAT trials, there was definitive evidence against the MCAR assumption in favour of the more plausible alternative of MAR or MNAR. In these trials, it was one of the MI procedures which was the least biased and most precise. It is unlikely that QoL data is missing for a reason unrelated to QoL. Under-going MI seems like a better alternative, because the MAR assumption is more plausible in this instance.

Molenberghs and Kenward discuss the merits of the different MI approaches (Molenberghs, Kenward 2007). They promote the use of a regression or predictive mean match model for the longitudinal setting. An advantage of PMM over regression is that imputed values are always within the range of the data (Fairclough 2002). In situations of monotone missingness, it is expected that the MCMC approach and the regression method should lead to very similar answers. The difference is due largely to the different prior distribution used (Molenberghs, Kenward 2007). Regression and PMM imputation have been shown to be the most accurate in this current situation, with the propensity score model on the whole performing poorly in comparison. A reason for this is given by Molenberghs and Kenward:

> "The propensity score method uses only the covariate information associated with whether the imputed values are missing. It does not use associations among variables; As a consequence, while it can be effective for inferences about the distributions of individual imputed variables, it is not appropriate for analyses involving relationships among variables." (Molenberghs, Kenward 2007, pg144)

If the analysis and imputation model are the same, the resulting estimates under imputation will be equivalent to those obtained by maximum likelihood, for example, using a repeated measure design (Carpenter, Kenward 2007). In contrast when additional information about the patient, which is potentially related to their QoL (when the observation is missing), is added to the imputation model, the estimates may be dramatically different.

Using simulation studies, a number of authors have shown that multiple imputation provides more robust treatment estimates and appropriate standard errors (Myers 2000; Hunsberger et al. 2001; Cook 1997; Donders et al. 2006; Huson, Chung & Salgo 2007; Liu, Gould 2002; Morita et al. 2005; Patrician 2002; Tang et al. 2005). Each of these authors has compared multiple imputation to a simpler alternative or complete case strategy. In some situations, multiple imputation was shown to perform at least as well in terms of treatment effects, but on the whole provided a more realistic standard error. The rationale underlying the approach used in this thesis is that the 'reminder-responders' are likely to be representative of those who do not respond at all. Thus, we were able to identify potentially suitable imputation methods. This method could then be used on the actual missing data in order to allow more patients (and data) to be included in the final analysis. The effects of which were seen in the sections above.

On the whole although the magnitude (and in some cases the direction of effect) differed, the conclusion from the resultant significance test remained unchanged. However, in RECORD, if the LVCF procedure had been used the resultant p-value would have been p=0.01 compared to p=0.05 from the published complete-case analysis. Therefore, instead of showing a

borderline significant difference between treatment groups more definite evidence would have been found. Similarly, with the NPC trial, the use of mean imputation would have a shown a significant treatment difference in the three dimension scores, rather than the no difference which was observed. It is then perhaps more appropriate to use imputation as a tool to assess the sensitivity of results, rather than as part of the primary analysis. This conclusion is also recommended by a number of authors on this topic (Fairclough 2002; Fayers, Machin 2007; Carpenter, Kenward 2007).

Imputation should always be used with caution. There is no substitute for real data. Data collected by reminder will always be preferable to data that has been imputed, as reminder-responses are reflective of the truth. The benefits of imputation are that potentially all participants can be included in analysis and that it may provide a cheaper alternative than repeated reminders. However, this chapter has shown that imputation will still introduce bias into the result, whatever the method. The results of this are utilised later in chapter ten to investigate the cost-effectiveness of the different data collection strategies (imputation and/or reminders).

As Huson *et al.* discuss, there is no one imputation technique that is applicable for all possible missing data patterns and missing data mechanisms (Huson, Chung & Salgo 2007). However, here it has been shown that multiple imputation was more suitable then simple imputation methods, and in particular when missing data was found to be informative. Multiple imputation models the uncertainty in the missing data and is based on the MAR assumption, which is more plausible in the QoL setting. When deciding on the best model for imputation, it is recommended that all the variables in the analysis model are included, plus any additional variables which are related to both outcome and missingness. MI is the difficult way to analyze data where missingness is MAR and will only provide a benefit when the analyst has additional information that is related to QoL, both when the response is observed and when it is missing (Fairclough 2002).

7.9.1 Conclusion

The advantage of the work presented in this thesis over the current literature, is the use of the reminder data in identifying suitable imputation procedures. The reminder data not only increases the sample size for analysis, but also provides a basis for an investigation into which is likely to be the best, most accurate imputation method. This 'best' choice can then be used on the actual missing data. This allows more patients to be included in a sensitivity analysis. It is recommended that researchers consider carefully which imputation method (if any), will be suitable and that they take into account the missing data mechanism inherent in the trial. The reminder-responses will help in this, as has been outlined throughout the chapter. However, imputation is not the only solution to missing data. Some alternative model-based techniques are discussed in the next chapter.

Chapter 8 Model-based procedures for missing data

8.1 Introduction

Chapter three described that the analysis method chosen by the trial researchers was an analysis of covariance (ANCOVA) on a single endpoint (usually final), adjusting for some baseline patient characteristics. This was undertaken as a complete-case analysis and any patient with missing baseline or final QoL assessment was ignored. Chapter six and seven detailed a number of imputation procedures as one way of dealing with the missing data. This showed that multiple imputation was preferred to simple imputation when the missing data were not missing completely at random (MCAR). An alternative to the ANCOVA strategy (with or without imputation) is to utilise the longitudinal nature of the follow up data and employ some other more sophisticated model-based methods. This chapter aims to outline some of these available methods for longitudinal data (continuous outcomes). These methods will then be applied to the trial datasets in chapter nine.

8.2 Models for longitudinal data

This section outlines the analysis of longitudinal data in the context of QoL outcomes. There are two types of longitudinal design: event-or-condition driven designs and time-driven designs (Fairclough 2002). The first occurs when the study objective is to compare outcomes in subjects experiencing the same condition, and when assessments are planned to occur at clinically relevant times, or to correspond with phases of the intervention. When a design is event-driven, or there are very few assessments (4 or less) that are at distinct points in time, a repeated measures model is most appropriate. Time-driven designs involve a more prolonged time period or the phases of treatment are not distinct. In this instance, with a large number of assessments or if they are spread over time, a growth curve model is appropriate.

8.2.1 Model structure

Using the notation that was originally presented in chapter four, consider a study of J QoL assessments. Then y_{ij} indicates the jth observation of QoL on the ith individual. The general linear model for QoL outcomes can be expressed as

$$Y_{ij} = X_{ij}\beta + \varepsilon_{ij} \text{ or in matrix notation } Y_i = X_i\beta + \varepsilon_i,$$

where

Y_i = complete data vector of planned observation of QoL outcome for the ith subject, including the observed data Y_i^{obs} and missing data Y_i^{miss}

X_i = the design matrix of covariates

β = the vector of fixed effect parameters

ε_i = the vector of residual errors

Σ_i = covariance of the complete data.

This general linear model can be formed as either a repeated measures model or a growth curve model. The model building process starts by defining a fully parameterised structure for the means ($X_i\beta$), then identifying the structure of Σ_i and finally, simplifying the mean structure (Fairclough 2002).

To compare models a maximum likelihood (ML) ratio test or restricted maximum likelihood (REML) ratio test can be carried out (Diggle et al. 2002). The deviance is constructed as the difference in the values of -2 times the log likelihood for the two models:

$$\text{Deviance} = -2(\log L_1 - \log L_2) \sim \chi^2_{df1-df2}.$$

This statistic is then compared to a χ^2 distribution with degrees of freedom equal to the difference in the number of parameters in the two covariance structures. Tests based on REML are valid as long as the fixed effects are the same. Either ML or REML ratio tests are suitable when comparing nested covariance structures. Likelihood tests for nested fixed effect models must be limited to the use of ML, because the restricted likelihood adjustment depends on the fixed effects design matrix. To identify nesting, it is necessary to decide whether the set of restrictions in the parameters of one model can be used to define the other model. Table 8.2 gives examples of the covariance structures. It is easily seen that a restriction on

the toeplitz structure defines compound symmetry and all homogenous structures are nested within the respective heterogeneous structure. Structures such as toeplitz and first order autoregressive (AR(1)) cannot be directly compared.

8.2.2 Repeated measures models

The general linear model has been defined in the last section. Using this as a base, the structure of a repeated measures model can be explained. The mean structure for repeated measures models will be described using the cell means model. This will be illustrated using two treatments (control and intervention) and three assessments (baseline, 3 months and 6 months post intervention) as shown in Table 8.1.

Table 8.1: Repeated measures cell means model – two treatments, three assessments

Treatment	Baseline	3 months	6 moths
Control	μ_{11}	μ_{12}	μ_{13}
Intervention	μ_{21}	μ_{22}	μ_{23}

Utilising an additional subscript, h, to indicate treatment group the equation for the model is given by $Y_{hij} = \mu_{hj} + \varepsilon_{hij}$ where μ_{hj} is the average QoL score for the jth measurement of the hth group. Other covariates (e.g. age, gender) can be added to the model when needed and as such the model becomes $Y_{hij} = \mu_{hj} + X_i\beta + \varepsilon_{hij}$. With the addition of these covariates, the interpretation of the means μ_{hj} will change. To facilitate the interpretation it is advisable to centre continuous covariates at a meaningful point (Fairclough 2002). Addition of covariates with missing values can cause problems because of the deletion of cases from the analysis. In repeated measures models, the assessments for each subject are assumed to be correlated over time, such that $\mathrm{Var}[Y_i] = \Sigma_i$. To fit a repeated measures model, this covariance structure must be specified. There are two types: structured and unstructured. Unstructured covariance is the least restrictive and is often a good choice when the number of assessments is small (Fairclough 2002). With three assessments there are six covariance parameters – the variance at each time point $(\sigma_1^2, \sigma_2^2, \sigma_3^2)$ and the covariance between each pair of time points $(\sigma_{12}, \sigma_{13}, \sigma_{23})$.

Table 8.2: Covariance structure for three repeated measures

	No. of parameters	Structure
Unstructured	6	$\begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ & \sigma_2^2 & \sigma_{23} \\ & & \sigma_3^2 \end{pmatrix}$
Heterogeneous toeplitz	5	$\begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_1 & \sigma_1\sigma_3\rho_2 \\ & \sigma_2^2 & \sigma_2\sigma_3\rho_1 \\ & & \sigma_3^2 \end{pmatrix}$
Heterogeneous Compound symmetry	4	$\begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho \\ & \sigma_2^2 & \sigma_2\sigma_3\rho \\ & & \sigma_3^2 \end{pmatrix}$
Heterogeneous autoregressive	4	$\begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho^2 \\ & \sigma_2^2 & \sigma_2\sigma_3\rho \\ & & \sigma_3^2 \end{pmatrix}$
Toeplitz	3	$\begin{pmatrix} \sigma^2 & \sigma^2\rho_1 & \sigma^2\rho_2 \\ & \sigma^2 & \sigma^2\rho_1 \\ & & \sigma^2 \end{pmatrix}$
Compound symmetry	2	$\begin{pmatrix} \sigma^2 & \sigma^2\rho & \sigma^2\rho \\ & \sigma^2 & \sigma^2\rho \\ & & \sigma^2 \end{pmatrix}$
First order autoregressive AR(1)	2	$\begin{pmatrix} \sigma^2 & \sigma^2\rho & \sigma^2\rho^2 \\ & \sigma^2 & \sigma^2\rho \\ & & \sigma^2 \end{pmatrix}$

Structured covariance places a number of restrictions on these parameters. For example, if correlation between time points is similar regardless of how far apart they are, then compound symmetry may be suitable. The autoregressive structures allow observations further apart to be less strongly correlated. Another option is to allow the variance to be heterogeneous, but retain the covariance structure. Table 8.2 summarises some of the available structures for three repeated measures. The cell means model generates estimates of the means for each treatment. However, additional comparisons of these means may be of interest. Any hypothesis that can be expressed as a linear combination of the estimated parameters (means) can be tested.

8.2.3 Growth Curve models

The most common form of growth curve model is a *polynomial model*. This fits a curve to the data to describe change in QoL as a function of time (t), such as:

$$Y_{hij}(t) = \beta_{h0} + \beta_{h1} t_{hij} + \beta_{h2} t_{hij}^2 + \cdots + \varepsilon_{hij}$$

The purpose is to approximate a curve which fits the observed data, and can provide a good approximation of the average QoL. Higher order terms allow the curve to depart from linearity, however interpreting coefficients can be difficult. The maximum number of terms (including the intercept) is equal to the number of assessments. For example, if there are four QoL assessments, then the highest order term which can be considered in the model is t^3_{hij}.

An alternative is a *piecewise linear regression* model. The change in QoL is modelled as a linear function over short intervals of time. Although changes in QoL are not expected to be linear, it is reasonable to assume changes are approximately linear over short intervals. The model can be written as:

$$Y_{hij}(t) = \beta_{h0} + \beta_{h1} t_{hij} + \beta_{h2} t_{hij}^{(2)} + \beta_{h3} t_{hij}^{(3)} + \cdots + \varepsilon_{hij}$$
$$t_{hij}^{(c)} = \max\left(t_{hij} - T^{(c)}, 0\right).$$

Where $T^{(c)}$ represents the time at which there was a planned assessment or a time at which changes in QoL might occur. Ideally the study would have been designed such that the points of clinical interest match up with the time of the QoL assessment. Since there are several assessments on each subject, there is the expectation that assessments for each individual will be correlated with time. A typical approach for modelling the covariance structure is the use of a mixed-effects model, which is given by

$$Y_{hi} = X_{hi}\beta_h + Z_{hi} d_{hi} + e_{hi} \quad \text{(Diggle et al. 2002).}$$

The fixed effects ($X_i\beta$) model the average QoL and the random effects ($Z_i d_i$) model the variation among individuals relative to the average. The covariance of a mixed-effects model for Y_{hi} has the general structure:

$$\Sigma_{hi} = Var(Y_{hi}) = Var(Z_{hi}d_{hi} + e_{hi})$$
$$= Var(Z_{hi}d_{hi}) + Var(e_{hi})$$
$$= Z_{hi}D_hZ'_{hi} + R_{hi},$$

where D_h and R_{hi} can have a number of structures.

The simplest random effects model ($Z_{hi}d_{hi}$) has a single random effect (d_{hi1}). The model has a random intercept for each individual $Y_{hi} = X_{hi}\beta_h + d_{hi1} + e_{hi}$, where d_{hi} is the average difference between individual response and the mean response. This implies variation for individuals does not change over time. The individual curves are parallel.

More typically for longitudinal data there are two random effects. The second (d_{hi2}) allows variation in the rate of change over time (slope) among individuals. This model has a random intercept and random slope for each individual ($Y_{hi} = X_{hi}\beta_h + d_{hi1} + d_{hi2}t_{hi} + e_{hi}$). In most cases, the two random effects are correlated. The second part of the variance structure R_{hi} can be quite simple ($\sigma^2_h I$) or quite complex, such as autoregressive - equal spacing, autoregressive - unequal spacing and toeplitz - equal spacing. Compound symmetry cannot be used for modelling R_{hi}. It is possible to allow the covariance structure to vary among treatment arms.

8.2.4 Summary

The repeated measures model and growth curve model can be implemented in SAS using PROC MIXED (SAS Institute Inc. 2004). This type of modelling makes an assumption of MAR. Often in QoL studies, missingness is more likely to be MNAR with those displaying poorer QoL less likely to respond (Fielding et al. 2008a). Chapter five showed this to be the case in a number of the example trials. The missingness is in itself informative and non-ignorable. This should be taken into consideration in analysis. In the sections that follow, alternative model-based procedures, which can account for the MNAR data are discussed.

8.3 Pattern mixture models

One way to deal with the issue of informative missingness is to use a pattern-mixture model. Diggle *et al*. suggest that the rationale for a pattern mixture model is that each subject's dropout time is somehow predestined (Diggle et al. 2002). The classification of subjects in a longitudinal trial according to their dropout time provides an obvious way of dividing the patients into sub-groups after the event. It is then sensible to ask whether the response characteristics of interest do or do not vary between these sub-groups. The pattern-mixture model allows for different response models for each pattern of missing values (Fairclough 2002). The observed data is a mixture of these weighted by the probability of each missing value or dropout pattern. To undertake a pattern mixture model, the proportion of subjects with each pattern of missing data needs to be known, but how the missingness depends on the Y_i^{mis} need not be specified. However, the disadvantages are that there are a large number of potential patterns of missing data and that there may be difficulties in estimating all parameters in each pattern.

Little describes the basic concept of a pattern mixture model (Little 1994). There are P different missing data patterns ($R_i \in M^{\{p\}}$), where $M^{\{p\}}$ was previously defined (section 4.1.1) as a matrix of indicators (R_i) of the observed variables in pattern P. The distribution of responses, Y_i may differ across the P patterns with different parameters $\beta^{\{p\}}$ and variance $\Sigma^{\{p\}}$, such that

$$Y_i \mid M^{\{p\}} \sim N\left(X_i \beta^{\{p\}}, \Sigma_i^{\{p\}}\right), \quad p=1,\ldots,P.$$

This allows changes in QoL among subjects in each pattern to be described using a different intercept and slope. This enables patients who drop out earlier to have lower QoL scores initially and to decline more rapidly over time. The early dropouts may have more or less variability in their scores than patients who remain in the study.

The true distribution of the measures of QoL for the entire group of patients is a mixture of the distributions from each of the *P* groups of patients. The quantities of interest are the expected values of the parameters averaged over the missing data patterns:

$$\mathrm{E}[\beta] = \sum^{P} \pi^{\{p\}} \beta^{\{p\}}$$

where $\pi^{\{p\}} = n^{\{p\}}/N$ is the proportion of subjects observed with the *pth* pattern.

The general method is to stratify the patients by the missing data pattern. Then the parameters $\left(\beta^{\{p\}}, \Sigma^{\{p\}}\right)$ are estimated within each of the strata. The population estimate is the weighted average of the estimates from the *P* missing data patterns, given by:

$$\hat{\beta} = \sum \hat{\pi}^{\{p\}} \hat{\beta}^{\{p\}} .$$

Estimating these parameters is not as easy as it appears. There may be a large number of patterns, some of which occur for a small number of subjects. Secondly, for some patterns, the model is under identified. This means that all the parameters cannot be estimated without additional assumptions or restrictions placed on the model (Little 1994). The process of a pattern-mixture model, including some possible restrictions is now described using the simple case of two repeated measures.

8.3.1 Bivariate case

The simplest longitudinal study consists of two repeated measures. For example, a baseline score and one follow up score. There are potentially four possible response patterns: complete case, both missing, and one observed one missing (Table 8.3). In each of the four patterns, there are five parameters to be estimated: two means $\left(\hat{\mu}_1^{\{p\}}, \hat{\mu}_2^{\{p\}}\right)$ and three parameters for the covariance $\left(\hat{\sigma}_{11}^{\{p\}}, \hat{\sigma}_{12}^{\{p\}}, \hat{\sigma}_{22}^{\{p\}}\right)$. It is possible to estimate nine of the twenty parameters from the data. There is insufficient information to estimate the remaining parameters and the model is under-identified. An additional assumption is required to estimate these

parameters. There are a number of possible restrictions that can be placed on the data to estimate the missing parameters.

Table 8.3: Missing data patterns with two assessments

Pattern	Y_1	Y_2	R
1	$Y_1^{\{1\}}$	$Y_2^{\{1\}}$	(1,1)
2	$Y_1^{\{2\}}$	?	(1,0)
3	?	$Y_2^{\{3\}}$	(0,1)
4	?	?	(0,0)

(a) Complete case missing value (CCMV) restriction

This set of restrictions is based on the complete cases and the assumption is that the missing value distributions are equal to the complete case distributions (Fairclough 2002). In the example datasets, the baseline QoL measurement is usually present. Therefore, the CCMV restriction will be illustrated using the first two patterns only, where all subjects were observed initially, but some were missing at the second assessment. The means for the first assessment and that for the second (pattern one) can be estimated directly from the data. It is not possible to calculate the mean score at the second assessment for those with pattern two, as the data are missing. However, it can be estimated assuming the same relationship holds in pattern two as in pattern one.

A regression model of Y_2 on Y_1 for those with pattern one is evaluated. This model is then used to estimate Y_2 for those showing pattern two. The estimated mean for the second assessment in pattern two is given by

$$\hat{\mu}_2^{(2)} = \overline{Y}_2^{(1)} - \frac{\hat{\sigma}_{12}^{(1)}}{\hat{\sigma}_{11}^{(1)}}\left(\overline{Y}_1^{(1)} - \overline{Y}_1^{(2)}\right),$$

where $\hat{\sigma}_{12}^{(1)}$ is the covariance between the baseline and 3 month assessment and $\hat{\sigma}_{11}^{(1)}$ is the variance of the baseline assessment for those with pattern 1. Combining the estimates for the two patterns, the overall estimates of the two means are:

$$\hat{\mu}_1 = \hat{\pi}^{\{1\}}\hat{\mu}_1^{\{1\}} + \hat{\pi}^{\{2\}}\hat{\mu}_1^{\{2\}} = \hat{\pi}^{\{1\}}\overline{Y}_1^{\{1\}} + \hat{\pi}^{\{2\}}\overline{Y}_1^{\{2\}} = \overline{Y}_1$$

$$\hat{\mu}_2 = \hat{\pi}^{\{1\}}\hat{\mu}_2^{\{1\}} + \hat{\pi}^{\{2\}}\hat{\mu}_2^{\{2\}} = \overline{Y}_2^{\{1\}} + \frac{\sigma_{12}^{\{1\}}}{\sigma_{11}^{\{1\}}}\left(\hat{\mu}_1 - \overline{Y}_1^{\{1\}}\right)$$

Full details of the complete case missing value (CCMV) restriction, including procedures for estimating the covariance parameters are provided by Little (Little 1994). This restriction is essentially assuming MAR. If this was thought not to be the case then this procedure would not be appropriate. Note in this special case, these estimates are the same as those obtained using maximum likelihood estimates (MLE) on all available data.

(b) Brown's protective restriction

Brown's protective restriction, which assumes MNAR is an alternative to the CCMV restriction (Brown 1990). This restriction requires monotone dropout where only patterns one and two are observed. It also assumes that missingness of Y_2 depends on Y_2. In this case, a regression of Y_1 on Y_2 results in an estimated mean for the second assessment in pattern 2 of:

$$\hat{\mu}_2^{(2)} = \overline{Y}_2^{(1)} - \frac{\hat{\sigma}_{22}^{(1)}}{\hat{\sigma}_{12}^{(1)}}\left(\overline{Y}_1^{(1)} - \overline{Y}_1^{(2)}\right).$$

Combining the estimates for the two patterns, the overall estimates of the two means are:

$$\hat{\mu}_1 = \hat{\pi}^{\{1\}}\hat{\mu}_1^{\{1\}} + \hat{\pi}^{\{2\}}\hat{\mu}_1^{\{2\}} = \hat{\pi}^{\{1\}}\overline{Y}_1^{\{1\}} + \hat{\pi}^{\{2\}}\overline{Y}_1^{\{2\}} = \overline{Y}_1$$
$$\hat{\mu}_2 = \hat{\pi}^{\{1\}}\hat{\mu}_2^{\{1\}} + \hat{\pi}^{\{2\}}\hat{\mu}_2^{\{2\}} = \overline{Y}_2^{\{1\}} + \frac{\sigma_{22}^{\{1\}}}{\sigma_{12}^{\{1\}}}\left(\hat{\mu}_1 - \overline{Y}_1^{\{1\}}\right).$$

Algebraic details of this can be found in Brown (Brown 1990) or expressed more simply in Fairclough (Fairclough 2002).

The estimated QoL means, under Brown's protective restriction, will be less than those under the CCMV restriction. This is because CCMV uses a MAR assumption, while Brown uses an assumption of MNAR. The differences between the estimates will be greater if the proportion of missing data increases, or the magnitude of the difference between the first assessment means increases, or the correlation between baseline first and second assessment scores decreases (Fairclough 2002).

8.3.2 Extending to monotone dropout

The description above relates to a study with two assessments only and monotone dropout. Extending this procedure to a longitudinal study of three or more assessments gets tricky. There are three possible restrictions in this case: an extension of CCMV, available case missing value (ACMV) restriction and neighbouring case missing value (NCMV) restriction when the data are monotone missing. Under the CCMV restriction, the data from subjects in pattern one are used to impute the means for the missing observations in the remaining patterns. In the ACMV restriction, available data from subjects in all patterns are used to impute the means for the missing observations in the remaining patterns. The restrictions are a bit more feasible than the CCMV restriction as more observations are used to estimate some parameters. Finally, for the NCMV restriction, available data from subjects in the neighbouring pattern are used to impute the means for the missing observations in the remaining patterns. Each of these restrictions is described using four assessments (resulting in four patterns with monotone missingness). An example of these patterns is shown in Table 8.4.

Table 8.4: Monotone patterns with four assessments

	Assessment			
Pattern	1	2	3	4
1	X	X	X	X
2	X	X	X	
3	X	X		
4	X			

X represents observed data

(a) Complete case missing value restriction

As with the bivariate case in the CCMV restriction, the data from pattern one are used to impute the means for the missing observation in the remaining patterns.

$$\beta_{[4\cdot123]}^{\{2\}} = \beta_{[4\cdot123]}^{\{1\}}$$
$$\beta_{[34\cdot12]}^{\{3\}} = \beta_{[34\cdot12]}^{\{1\}}$$
$$\beta_{[234\cdot1]}^{\{4\}} = \beta_{[234\cdot1]}^{\{1\}}$$

where $\beta^{\{1\}}_{[34\cdot12]}$ denotes the parameters from the regression of Y_3 on Y_1 and Y_2 and the regression of Y_4 on Y_1 and Y_2 using the complete cases in pattern one. It is important to note that this restriction is only feasible when the number of cases in pattern one is sufficient to estimate these parameters reliably. These restrictions result in six equations that need to be solved to obtain the unknown means and variance parameters. This can be straightforward. However, deriving the appropriate variance of the pooled estimates is complex. Curran *et al*. suggest an analytic technique using multiple imputation to avoid this problem (Curran et al. 2004). The procedure is as follows:

1. Pattern two – impute missing values at $j = 1$, using cases in pattern one.
2. Pattern three – impute missing values at $j = 3$ and $j = 4$, using cases in pattern one.
3. Pattern four – impute missing values at $j = 2$, $j = 3$ and $j = 4$, using cases in pattern one.
4. Analyze each set of imputed data and combine estimates, as previously described in chapter six.

(b) Available case missing value (ACMV) restriction

In the ACMV restriction, available data from participants in all the patterns are used to impute the means for the missing observations in the remaining patterns. This restriction is equivalent to the MAR setting of other analyses. The restrictions are as follows:

$$\beta^{\{2\}}_{[4\cdot123]} = \beta^{\{1\}}_{[4\cdot123]}$$
$$\beta^{\{3\}}_{[3\cdot12]} = \beta^{\{1,2\}}_{[3\cdot12]} \qquad \beta^{\{3\}}_{[4\cdot123]} = \beta^{\{1\}}_{[4\cdot123]}$$
$$\beta^{\{4\}}_{[2\cdot1]} = \beta^{\{1,2,3\}}_{[2\cdot1]} \qquad \beta^{\{4\}}_{[3\cdot12]} = \beta^{\{1,2\}}_{[3\cdot12]} \qquad \beta^{\{4\}}_{[4\cdot123]} = \beta^{\{1\}}_{[4\cdot123]}$$

Where $\beta^{\{1,2\}}_{[3\cdot12]}$ denotes the parameters from the regression of Y_3 on Y_1 and Y_2 using the cases in patterns 1 and 2. This restriction is a bit more feasible than the CCMV restriction, as more observations are used to estimate some of these parameters. Using the same multiple imputation approach as in CCMV, the procedure for the ACMV is as follows:

1. Pattern two– impute missing values at $j = 4$, using cases in pattern one.

2. Pattern three – impute missing values at $j = 3$, using cases in patterns one and two. Impute missing values at $j = 4$, using cases in pattern one and the imputed values at $j = 3$, in pattern three.
3. Pattern four – impute missing values at $j = 4$, using cases in patterns one, two and three. Impute missing values at $j = 3$, using cases in patterns one and two and the imputed values at $j = 2$, in pattern four. Impute missing values at $j = 4$, using cases in pattern one and the imputed values at $j = 2$ and $j = 3$, in pattern four.
4. Analyze each set of imputed data and combine estimates, as previously described in chapter six.

(c) Neighbouring case missing value (NCMV) restriction

In the NCMV restriction, available data from subjects in the neighbouring pattern are used to impute the means for the missing observations.

$$\beta^{\{2\}}_{[4\cdot123]} = \beta^{\{1\}}_{[4\cdot123]}$$
$$\beta^{\{3\}}_{[3\cdot12]} = \beta^{\{2\}}_{[3\cdot12]} \qquad \beta^{\{4\}}_{[2\cdot1]} = \beta^{\{3\}}_{[2\cdot1]}$$
$$\beta^{\{4\}}_{[4\cdot123]} = \beta^{\{1\}}_{[4\cdot123]} \qquad \beta^{\{4\}}_{[3\cdot12]} = \beta^{\{2\}}_{[3\cdot12]} \qquad \beta^{\{3\}}_{[4\cdot123]} = \beta^{\{1\}}_{[4\cdot123]}$$

As before, using multiple imputation the procedure is as follows:

1. Pattern two – impute missing values at $j = 4$, using cases in pattern one.
2. Pattern three – impute missing values at $j = 3$, using cases in pattern two. Impute missing values at $j = 4$, using cases in pattern one and the imputed values at $j = 3$, in pattern three.
3. Pattern four – impute missing values at $j = 2$, using cases in pattern three. Impute missing values at $j = 3$, using cases in pattern two and the imputed values at $j = 2$, in pattern four. Impute missing values at $j = 4$, using cases in patterns one and the imputed values at $j = 2$ and $j = 3$, in pattern 4.
4. Analyze each set of imputed data and combine estimates, as previously described in chapter six.

8.3.3 Summary

Each of the restrictions in the previous section is described for monotone patterns with four assessments. The set of restrictions is easily adapted for a different

numbers of assessments. Chapter nine presents the results of these pattern mixture models for the trial datasets. Each particular dataset is restricted to those showing a monotone pattern and the pattern mixture model carried out accordingly. The final analysis model used is the ANCOVA reported in the original trial publication.

8.4 Modelling the dropout process

In addition to the pattern mixture model, a number of other approaches exist to model longitudinal data with potentially informative dropout. Each makes assumptions about the underlying dropout mechanism. The notation used to describe these methods below was introduced in chapter four. A common feature of these methods is that they require the analyst to make strong assumptions. These assumptions cannot formally be tested and often defence of the assumptions is from a clinical standpoint rather than a statistical one. Lack of evidence of non-ignorable missing data for any particular model does not mean that the missing data are ignorable. Finally, estimates are not robust, to model misspecification (Fairclough 2002).

Three possible models for non-ignorable data are the joint mixed effects model, the conditional linear model and the selection model. The conditional linear model was proposed by Wu and Bailey (Wu, Bailey 1989). In this model, each patient's rate of change of QoL is assumed to depend on covariates and the time to an event associated with dropout. Both this model and the joint mixed effects model with time to event (dropout) are appropriate in settings where simple growth curve models describe the changes in the outcome and where there is variation in the rates of change among the individual patients. The selection model proposed by Diggle and Kenward is appropriate to settings where the assessments are equally spaced repeated measures and the missing data patterns are strictly monotone (Diggle, Kenward 1994).

8.4.1 Conditional linear model

One option to model non-ignorable data is the conditional linear model (CLM). This is based on a random effects mixture model. The missing data process is defined by the matrix M_i. In this thesis, thus far, $M_i = R_i$ the vector of indicators of whether a QoL assessment was observed or not. In the context of random-effects mixture model, the random-effects model includes the dropout time and $M_i = \beta_i$, a random coefficient. The conditional linear model proposed by Wu and Bailey is an example of a random-effects mixture model (Wu, Bailey 1989). They proposed a CLM based on modelling the change of an outcome over time (slope) depending on covariates, the baseline value of the outcome and the time of dropout, defined as T_i^D. Fairclough provides the technical detail of the model set up (Fairclough 2002). Put simply, the model is such that each individuals QoL profile can be described as a linear function of time. Random variation for each individuals intercept and slope is allowed and the model can be fitted into the standard mixed effects model framework (Brown, Prescott 1999). Each individual's slope may depend on a set of covariates, the initial value of the outcome (Y_{i1}), as well as the polynomial function of the time of dropout (T_i^D). The form of the relationship is allowed to vary across the *h* treatment groups (usually h=2).

There are several practical consequences and assumptions of this model: changes over time are assumed to be roughly linear; there is enough variation in the rate of change among subjects to allow the modelling of the variation; all subjects must have baseline measurement and complete covariate data; the time of dropout is known for all subjects. All subjects are either followed up until dropout or have completed the final assessment. There is no way to distinguish between those who would have continued to have assessments and those who would have dropped out after the final assessment if follow up had continued. If the slopes were dependent only on covariates, then the data would be MCAR. If the slopes were dependent on the baseline measure of outcome, but not the time of dropout, the data would be considered MAR. When terms involving the time of dropout remain in the model, there is evidence that the data are MNAR. Further details of

this type of model can be found in the original publication by Wu and Bailey (Wu, Bailey 1989) and in the book by Fairclough (Fairclough 2002).

8.4.2 Joint mixed effects model

An alternative to the conditional linear model is the joint mixed effects model. The concept behind a joint model is such that the trajectory of a longitudinal outcome (QoL) will be correlated with time to event (or dropout). For example, those patients whose QoL decreases more rapidly are more likely to drop out (or experience death). The same basic model that was described in section 8.2.2 can be used. Each subject QoL profile can be described as a linear function of time, with random variation among subjects of the intercept, β_{i1} and the linear rate of change, β_{i2}. The time of an event associated with the dropout from QoL assessment (T_i^D) is incorporated into the model by allowing some function of time ($f(T_i^D)$) to be correlated with the random effects of the longitudinal model. Using the approach of Schluchter (Schluchter 1992), the time to dropout is given by

$$f(T_i) = \mu_T + r_i \quad \text{where } f(T_i) = \ln(T_i).$$

The longitudinal outcome is modelled using the standard mixed effects model

$$Y_i = X_i\beta_i + Z_i d_i + \varepsilon_i .$$

It is known that $\beta_i \sim N(\beta, D)$, $f(T) \sim N(\mu_T, \tau^2)$ and $\text{cov}(\beta_i, f(T_i)) = \sigma_{bt}$. Therefore, it follows that the joint model is given by

$$\begin{bmatrix} \beta_i \\ f(T_i^D) \end{bmatrix} \sim N\left(\begin{bmatrix} \beta \\ \mu_T \end{bmatrix}, \begin{bmatrix} D & \sigma_{bt} \\ (\sigma_{bt})' & \tau^2 \end{bmatrix} \right).$$

In this model, the random effects (d_i) are correlated with the dropout $f(T_i^D)$. One thing to note is that T_i^D need not be the time after which no QoL assessments are

available, but can be the time to an event associated with dropout. The expected changes in QoL from this model are a function of dropout time such that

$$E\left[\beta_i \mid T_i^D\right] = \beta + \sigma_{bt}\tau^{-2}\left(f\left(T_i^D\right) - \mu_t\right).$$

There are several differences between this model and the CLM described in the previous section. The joint model allows the intercept and slope to be related to the time of dropout. There is an added restriction that the relationships are linear functions of $f(T_i^D)$. In the joint model, the expected time of dropout is a function of the initial QoL scores, as well as rate of change over time.

Under this model, the missing data are non-ignorable if the random effects are correlated with the time of dropout $\left(\sigma_{bt} \neq 0\right)$. For example, if a patient has more rapid decline in QoL and fails earlier, the slope random effect is positively correlated with the failure time. Lack of significant correlation does not imply the missingness is ignorable unless the model for dropout is correct.

De Gruttloa and Xin Ming proposed an alternative parameterisation where $f(T_i) = \mu_T + \lambda d_i + r_i$ (De Gruttola, Xin Ming 1994). In this case, the joint distribution can be written as

$$\begin{bmatrix} Y_i \\ f\left(T_i^D\right) \end{bmatrix} \sim N\left(\begin{bmatrix} X_i\beta \\ \mu_T \end{bmatrix}, \begin{bmatrix} \Sigma_i & Z_i\lambda' D \\ \left(Z_i\lambda' D\right)' & \lambda' D\lambda + \tau_*^2 \end{bmatrix}\right).$$

The advantage of this parameterisation is that it is easier to implement in SAS using PROC NLMIXED (SAS Institute Inc. 2004). The results can then be converted to the Schluchter parameterisation as these are more-interpretable for reporting results (Schluchter 1992). By comparing the joint distributions under the two parameterisations, it can be seen that $Z_i\sigma_{bt} = Z_i\lambda' D$ and $\tau^2 = \lambda' D\lambda + \tau_*^2$.

The joint mixed effects model was carried out in SAS using a combination of PROC MIXED, PROC LIFEREG and PROC NLMIXED (SAS Institute Inc. 2004). Firstly, the most appropriate growth curve model for the longitudinal QoL outcome using PROC MIXED was identified (section 8.2.3). This assumes MAR and can be used to find initial estimates for β, D and σ^2. The distribution of the time to dropout is modelled using PROC LIFEREG and provides initial estimates of μ_T and τ^2. Using PROC NLMIXED, the likelihood function of the joint model is specified along with these initial estimates. PROC NLMIXED fits nonlinear mixed models by maximizing an approximation to the likelihood, integrated over the random effects. A variety of alternative optimization techniques are available to carry out the maximization (SAS Institute Inc. 2004). Final estimates of parameters from the mixed effects model are obtained. Estimates of change in QoL can then be obtained and the correlation of the random effects with the time of dropout can be calculated.

8.4.3 Selection model

An example of a selection model for monotone missingness was proposed by Diggle and Kenward (Diggle, Kenward 1994). The dropout process was defined in a similar manner to MCAR, MAR and MNAR. Completely random dropout (CRD) matches up with MCAR, where dropout is completely independent of observed measurements. Random dropout (RD) tallies with MAR and dropout is related to observed measurements. Finally, informative dropout (ID) corresponds to MNAR where dropout depends on unobserved measures.

The response model is the standard multivariate regression model that has been used earlier. However, for the dropout model Diggle and Kenward assume a logistic regression $f\left(R_i \mid Y_i, \Gamma\right)$, which may depend on covariates, previous observations or unobserved measurements at dropout (Diggle, Kenward 1994). The conditional probability of dropout at time *j*, given the previous measures through time *j-1* is:

$$P_j = \Pr\left(T_i^D = t_j \mid Y_{i1}, \ldots, Y_{ij-1}\right).$$

The linear logistic model for the dropout process takes the form

$$\text{logit}\left[P_j\right] = \gamma_0 \mathrm{X}_i + \gamma_1 y_{ij} + \gamma_2 y_{ij-1}.$$

It follows that if $\gamma_2 \neq 0$, then dropout depends on the previously observed responses. If $\gamma_1 \neq 0$, then dropout depends on the current unobserved response. Therefore, if $\gamma_1 = \gamma_2 = 0$ then the dropout process is CRD. If $\gamma_1 = 0$ and $\gamma_2 \neq 0$, then dropout depends on previously observed measures and therefore, missingness is RD. For $\gamma_1 \neq 0$, then dropout depends on current data and missingness is said to be informative (ID). In the first two cases (CRD and RD), the response model and dropout model can be estimated separately. However, in the third case (ID) the models must be estimated jointly.

8.4.4 Summary

Fairclough describes the requirements of each of the three models that were explained above (Fairclough 2002). Both the selection model and conditional linear model do not allow the baseline observation to be missing. The selection model requires a monotone missing data pattern, while the other two models allow an intermittent pattern if it is MAR. Unequally spaced QoL assessments are not allowed in the selection model. The growth curve model involved in the conditional linear model must be linear. For these reasons, in the example datasets the model which fits the framework best is the joint mixed effects model. This model is applied to the NPC trial data in chapter nine.

8.5 Overview

As discussed previously, a repeated measures model makes the assumption of MAR and uses available data to estimate means and covariance. Models which make the MNAR assumption include the pattern mixture model, conditional linear model, joint mixed effects model and selection model. Pattern mixture models have the advantage that a model for the dropout mechanism does not need to be specified. However, this is counteracted by the need for a set of restrictions to be able to estimate all parameters. The validity of the restrictions

cannot be tested and results may be sensitive to the restriction chosen. In situations with large numbers of patterns and small sample sizes, their applicability may be limited.

The conditional linear model, joint mixed effects model and selection model, all require strong assumptions. These assumptions cannot be formally tested. Lack of evidence of non-ignorable missing data for a particular model does not lead to a conclusion of ignorable data. SAS procedures PROC MIXED and PROC NLMIXED (SAS Institute Inc. 2004) were used to implement the model-based procedures described.

The different model-based procedures described above provide an alternative framework to analyse the QoL outcomes within the example datasets. The longitudinal nature of the data can be utilised rather than ignored and a complete-case strategy carried out. Chapter nine implements repeated measures models and pattern mixture models on the example datasets in two different data scenarios. The results are then compared to that obtained in the reported complete case ANCOVA. There were a large number of dropouts in the NPC due to death, as a result of the population under study (cancer population receiving palliative care). Therefore, the joint mixed effects model is an appropriate analysis strategy to consider in this instance. The aim of chapter nine is to find out whether one of these more sophisticated model-based techniques is more suitable than the current practice of a complete-case strategy or one that uses imputation.

Chapter 9 Investigating the use of model-based procedures for missing data

9.1 Application to the datasets

Chapter eight discussed the various model-based strategies for missing longitudinal data. This chapter deals with the application of these methods to the example trials. The aim of this chapter is to determine whether the model-based procedures were appropriate alternatives to the previously discussed complete-case analysis or use of imputation.

The first model-strategy presented is the repeated measures model. A number of different options can be employed within the repeated measures model framework. Chapter eight described that the baseline measure of QoL can be incorporated in the repeated assessments or can be included in the model as a covariate. The latter is more comparable to the original complete-case ANCOVA approach reported by the trial researchers and is, therefore, the most informative in this context. This will allow for a direct comparison of the results to the reported results and those under imputation (chapter seven). A number of covariance patterns were considered and have been described previously in Table 8.2. During the model building process, several covariance structures were considered and the most appropriate covariance pattern was identified as the model with the smallest Akaike information criterion (AIC) (Burnham, Anderson 2004). The results of this final model are presented here.

The second model-based approach to be applied is the pattern mixture model. Chapter eight discussed the different approaches and assumptions within a pattern mixture model. A monotone missing data structure was required and the datasets were restricted to those patients with a monotone pattern. The three restrictions, complete-case missing value (CCMV), available case missing value (ACMV) and nearest case missing value (NCMV) were applied to each of these datasets for each of the QoL scores. These three restrictions were applied making

use of the multiple imputation procedure (PROC MI) in SAS, as previously described in section 8.3.2 (Curran et al. 2004). The regression model for a monotone missing data pattern was implemented. Once the augmented datasets were created, the analysis model carried out was the original ANCOVA model for treatment difference at follow up, adjusting for baseline scores. PROC MIANALYZE was used to combine the results.

The third strategy considered is a joint mixed-effects model. This approach makes use of the growth curve model introduced in chapter eight since a mixed-effects model is needed to model QoL in the presence of dropout. The most appropriate growth curve model was identified. This was then combined with the time to event (dropout) in a joint mixed-effects model as described in section 8.4.2. The SAS procedure PROC NLMIXED was used to undertake the joint mixed-effects model (SAS Institute Inc. 2004). This process was undertaken for the NPC Trial data, as there were a large number of dropouts due to death.

Each type of model described above was applied to two data scenarios. The first consisted of the immediate-responders only (at the endpoint of interest) and all data collected by reminder was set to missing. The second dataset was the actual observed data (immediate and reminder-responders) in the trial. This situation reflects that which was seen by the trial researchers at the end of data collection. The findings from each trial are described in turn.

9.2 REFLUX

9.2.1 Model-based analysis strategies when reminder data were missing

The subset of data used here contained those patients who provided QoL scores at both baseline and 12 months. Any data collected by reminder at six or 12 months was removed. The repeated measures modelling and pattern mixture modelling process was then carried out. The results are presented along with the original reported results in Table 9.1. The pattern mixture models were carried out on those patients with a monotone pattern. As such there were fewer patients

included in the analysis. The direct comparison to the treatment difference estimates must be made with caution. It can be seen that for each of the four QoL scores, the repeated measures model was the least biased when compared to the original observed result. For three of the four QoL scores, the CCMV restriction was the least biased pattern mixture model.

Table 9.1: REFLUX - Results from model-based analysis when reminder data were missing

		12 month treatment difference					
Model	N	Estimate	SE	95% CI	p-value	Bias	Ratio of CI width
				EQ5D			
Observed data	**309**	**0.047**	**0.030**	**(-0.003, 0.097)**	**0.07**	-	-
Repeated measures (UN)	310	0.080	0.032	(0.016, 0.14)	0.015	0.033	1.56
Pattern mixture (CCMV)	253	0.13	0.033	(0.06, 0.20)	<0.001	0.083	2.60
Pattern mixture (ACMV)	253	0.13	0.032	(0.06, 0.19)	<0.001	0.083	2.50
Pattern mixture (NCMV)	253	0.14	0.03	(0.08, 0.20)	<0.001	0.093	2.80
				SF12 physical component score (PCS)			
Observed data	**299**	**3.51**	**0.88**	**(1.77, 5.25)**	**<0.001**	-	-
Repeated measures (CS)	305	4.05	1.26	(1.53, 6.57)	0.002	0.54	1.45
Pattern mixture (CCMV)	250	5.18	1.20	(2.80, 7.56)	<0.001	1.67	1.37
Pattern mixture (ACMV)	250	5.91	1.28	(3.37, 8.44)	<0.001	2.40	1.46
Pattern mixture (NCMV)	250	5.71	1.05	(3.65, 7.77)	<0.001	2.20	1.18
				SF12 mental component score (MCS)			
Observed data	**299**	**1.54**	**1.18**	**(-0.78, 3.86)**	**0.19**	-	-
Repeated measures (CS)	305	2.17	1.46	(-0.74, 5.07)	0.14	0.63	1.25
Pattern mixture (CCMV)	250	2.95	0.84	(1.53, 4.85)	<0.001	1.41	0.72
Pattern mixture (ACMV)	250	4.04	0.91	(2.25, 5.84)	<0.001	2.50	0.77
Pattern mixture (NCMV)	250	4.43	0.86	(2.74, 6.13)	<0.001	2.89	0.73
				RQLS			
Observed data (N=276)	**276**	**14.1**	**2.29**	**(9.53, 18.6)**	**<0.001**	-	-
Repeated measures (CS)	294	12.8	2.93	(7.00, 18.7)	<0.001	1.3	1.29
Pattern mixture (CCMV)	241	20.6	2.58	(15.5, 25.6)	<0.001	6.5	1.11
Pattern mixture (ACMV)	241	18.5	2.62	(13.3, 23.6)	<0.001	4.4	1.14
Pattern mixture (NCMV)	241	16.0	2.37	(11.3, 20.7)	<0.001	1.9	1.04

For the RQLS score, the least biased pattern mixture model was under the NCMV missing value restriction. Taking into account bias and precision (ratio of CI width close to one), the repeated measures model seemed to provide more accurate treatment difference estimates. Part of this will be due to the fact that the repeated measures models are carried out on the same number of patients as the observed data. However, the pattern mixture models were on a reduced dataset containing only those patients with a monotone missing data pattern. For the primary outcome, the RQLS under the pattern mixture model (NCMV restriction)

appeared to be the best method. This ties in with the conclusion found in chapter five, that the missing data mechanism was MAR or possibly MNAR.

For the EQ5D data, the use of a repeated measures model found a significant treatment difference at 12 months (p=0.015) compared to borderline evidence in the observed data (p=0.07). In the case of the PCS the same conclusion of a significant treatment difference was reached. However, under the modelling process the magnitude of this difference was greater. The repeated measures model for the MCS was consistent with the observed ANCOVA and found no evidence of a treatment difference, yet the pattern mixture models did find evidence of a difference. Lastly, for the RQLS, although all methods found a significant treatment difference, the magnitude of the estimate was greater with the three pattern mixture models and particularly large in the case of CCMV restriction.

9.2.2 Model-based analysis strategies on the observed data

The previous section compared the trial result under several model-based strategies when the reminder data was removed. This section applies these techniques on the actual observed data. Therefore, any missing data was that which was actually missing in the trial. The observed data includes both the immediate and reminder responses.

It is not possible to make a direct comparison between the results, as the number of patients, N, involved in each method differs by nature of the method being carried out. The covariance pattern for the repeated measures model was not necessarily the same one as was found in section 9.2.1. The process of identifying the covariance structure was repeated here. The pattern mixture model shown to be most appropriate in the previous section is presented here.

The estimates of 12 month treatment difference were similar between the observed ANCOVA and the repeated measures model, which incorporated all available data for each of the four QoL scores (Table 9.2). The estimates from the pattern

mixture models were not as consistent. In the case of the EQ5D and mental component scores, they provided a different conclusion with regard to treatment difference.

Table 9.2: REFLUX - Results from model-based analysis on the observed data

		12 month difference		
Model	N	Estimate	95% CI	p-value
EQ5D				
Observed data	**309**	**0.047**	**(-0.003, 0.097)**	**0.07**
Repeated measures (CS)	344	0.049	(-0.001, 0.099)	0.05
Pattern mixture (ACMV)	316	0.13	(0.06, 0.20)	0.001
SF12 physical component score (PCS)				
Observed data	**299**	**3.51**	**(1.77, 5.25)**	**<0.001**
Repeated measures (CS)	336	3.57	(1.83, 5.32)	<0.001
Pattern mixture (CCMV)	305	6.02	(2.89, 9.16)	<0.001
SF12 mental component score (MCS)				
Observed data	**299**	**1.54**	**(-0.78, 3.86)**	**0.19**
Repeated measures (CS)	336	1.71	(-0.53, 3.95)	0.13
Pattern mixture (CCMV)	305	4.13	(1.79, 6.48)	0.002
RQLS				
Observed data	**276**	**14.1**	**(9.53, 18.6)**	**<0.001**
Repeated measures (UN)	327	14.1	(9.7, 16.5)	<0.001
Pattern mixture (NCMV)	290	15.5	(11.1, 20.0)	<0.001

CS – compound symmetry; UN - unstructured

9.2.3 Summary

Using a repeated measures model based on only the immediate responses (reminder responses missing) was found to be better than a pattern mixture model. When applied to all the observed data, the conclusion did not change, but the number of patients used in the analysis increased. The number of participants used by a repeated measures model should include everyone with at least one QoL assessment, but the limiting factor here was that there was some missing covariate data (including baseline QoL). One way round this, would be to include the baseline QoL as part of the repeated assessments, rather than as a covariate. Doing this, provided an estimate of treatment difference in the RQLS at 12 months of 11.5 with 95% CI (6.55, 16.5) and $p<0.001$. This was based on all 357 patients in the trial. This estimate is lower than that which was reported on the 276 patients, but there was still a significant treatment difference. The advantage of this over the ANCOVA is that every patient was involved in the analysis even if they did not provide the final assessment.

9.3 MAVIS

9.3.1 Model-based analysis strategies when reminder data were missing

Table 9.3 shows the calculated 12 month treatment difference from each of the models and for each of the three QoL scores in MAVIS. The observed data showed borderline evidence of a treatment difference at 12 months (p=0.08), while all the models showed no significant treatment difference (p>0.1) for the EQ5D scores. The repeated measures model which adjusted for baseline scores was the least biased model (b=0.004) and also the most precise (ratio = 1.12). For the two SF12 component scores under the different models, a treatment difference estimate was in the opposite direction to that which was observed, but the conclusion of no treatment difference remained the same. Each of the three pattern mixture models performed similarly for the three QoL scores. For both the PCS and MCS, the repeated measures model was the least biased, despite the direction of the treatment difference estimate. The precision values were close to the desired value of one.

Table 9.3: MAVIS – Results from model-based analysis when reminder data were missing

		12 month treatment difference				
Model	N	Estimate	95% CI	p-value	Bias	Ratio
EQ5D						
Observed data	**830**	**-0.019**	**(-0.04, 0.002)**	**0.08**	**-**	**-**
Repeated measures (CS)	829	-0.015	(-0.04, 0.007)	0.17	0.004	1.12
Pattern mixture (CCMV)	825	-0.014	(-0.04, 0.007)	0.20	0.005	1.12
Pattern mixture (ACMV)	825	-0.013	(-0.04, 0.01)	0.30	0.006	1.19
Pattern mixture (NCMV)	825	-0.013	(-0.04, 0.01)	0.29	0.006	1.19
SF12 physical component score (PCS)						
Observed data	**827**	**0.07**	**(-0.90, 1.03)**	**0.89**	**-**	**-**
Repeated measures (CS)	823	-0.13	(-1.13, 0.87)	0.80	0.20	1.04
Pattern mixture (CCMV)	815	-0.04	(-0.94, 0.85)	0.93	0.11	0.93
Pattern mixture (ACMV)	815	-0.03	(-0.92, 0.86)	0.95	0.10	0.92
Pattern mixture (NCMV)	815	-0.04	(-1.09, 1.01)	0.94	0.11	1.09
SF12 mental component score (MCS)						
Observed data	**827**	**-0.03**	**(-1.11, 1.05)**	**0.96**	**-**	**-**
Repeated measures (UN)	823	0.17	(-0.94, 1.27)	0.77	0.20	1.02
Pattern mixture (CCMV)	815	0.18	(-0.82, 1.18)	0.72	0.21	0.93
Pattern mixture (ACMV)	815	0.19	(-0.80, 1.19)	0.70	0.22	0.92
Pattern mixture (NCMV)	815	0.20	(-0.97, 1.36)	0.74	0.23	1.08

9.3.2 Model-based analysis strategies on the observed data

Table 9.4 displays the results of the modelling process on the observed data. The results for the EQ5D score were consistent among the different model-based methods and that observed in the ANCOVA analysis. The previous section did not favour one particular pattern mixture model restriction and thus, all three are presented. Using the CCMV restriction in the pattern mixture model, the estimate and 95% CI for the difference in EQ5D score was identical to that which was calculated from the original ANCOVA analysis. This was not surprising given the conclusion from chapter four, that there was no evidence against the MCAR assumption for EQ5D scores in MAVIS. For the PCS, no significant treatment difference was found in the observed data and this was mirrored by the model-based procedures. The magnitude of the estimate did vary between 0.02 and 0.18 compared to the observed of 0.07. For the MCS, using the CCMV and ACMV restrictions in the pattern mixture model, the calculated treatment estimate was negative, compared to the positive observed value. These values were, however, still non-significant.

Table 9.4: MAVIS - Results from model-based analysis on the observed data

		12 month difference		
Model	N	Estimate	95% CI	p-value
EQ5D				
Observed data	**830**	**-0.019**	**(-0.04, 0.002)**	**0.08**
Repeated measures (CS)	908	-0.021	(-0.042, 0.0002)	0.05
Pattern mixture (CCMV)	904	-0.019	(-0.04, 0.002)	0.08
Pattern mixture (ACMV)	904	-0.018	(-0.04, 0.004)	0.11
Pattern mixture (NCMV)	904	-0.020	(-0.04, 0.001)	0.07
SF12 physical component score (PCS)				
Observed data	**827**	**0.07**	**(-0.90, 1.03)**	**0.89**
Repeated measures (CS)	906	0.02	(-0.93, 0.97)	0.97
Pattern mixture (CCMV)	897	0.16	(-0.78, 1.10)	0.74
Pattern mixture (ACMV)	897	0.18	(-0.78, 1.13)	0.71
Pattern mixture (NCMV)	897	0.10	(-0.91, 1.11)	0.84
SF12 mental component score (MCS)				
Observed data	**827**	**-0.03**	**(-1.11, 1.05)**	**0.96**
Repeated measures (CS)	906	-0.18	(-1.24,0.88)	0.74
Pattern mixture (CCMV)	897	0.08	(-0.99, 1.15)	0.88
Pattern mixture (ACMV)	897	0.01	(-0.98, 1.17)	0.86
Pattern mixture (NCMV)	897	-0.10	(-1.27, 1.06)	0.86

9.3.3 Summary

Overall carrying out a repeated measures model on the response data with reminder data missing was better than a pattern mixture model. There was no apparent 'best' pattern mixture model restriction. When these procedures were applied to all the observed data again, the three restrictions were consistent with each other. The repeated measures approach provided a similar answer to the reported ANCOVA but included more patients. In the case of the EQ5D score, the difference between treatment groups was more apparent with p=0.05 rather than p=0.08, which was reported in the ANCOVA.

9.4 RECORD

9.4.1 Model-based analysis strategies when reminder data were missing

The results from the modelling process on the responder data with reminder-data missing are shown in Table 9.5. Differences in the EQ5D score and the two SF12 component scores were provided for both the two year calcium supplementation and two year vitamin D supplementation comparisons. To determine the best method, both treatment comparisons need to be considered together as the model fits the parameters simultaneously. Firstly, with the EQ5D scores, both the repeated measures model and the pattern mixture model with NCMV restriction showed the least bias. The repeated measures model was preferred, as the measure of precision was closer to the desired value of one.

Secondly, for the PCS the pattern mixture model with ACMV restriction was the best option. Although, it was not the least biased or most precise for each treatment comparison, the overall combination was better with this strategy. The same was seen for the MCS, suggesting that the pattern mixture model with ACMV restriction was the preferred strategy for the SF12 component scores.

9.4.2 Model-based analysis strategies on the observed data

The various model-based strategies were implemented on the observed data. The missing data was that which was truly missing. Table 9.6 shows the results of the process for each QoL score and treatment comparison. The three pattern mixture models provided similar estimates of treatment difference for both the calcium and vitamin D comparisons of EQ5D score. In particular, the CCMV restriction provided results very similar to the reported ANCOVA result for the calcium comparison. The use of a repeated measures model did not alter the conclusion for the vitamin D comparison. It provided more significant evidence of a difference in EQ5D scores for the calcium comparison and was based on a large number of patients. For the vitamin D comparison, the pattern mixture models did provide estimates of a different direction, yet they were still non-significant. For the PCS, the ACMV restriction was the most comparable model to the observed data. Finally, with the MCS, the repeated measures model gave results similar to the observed data. It made use of all patients who provided at least one assessment and was not restricted to only those with the two year assessment.

9.4.3 Summary

Undertaking the model-based procedures on the RECORD data did not identify one best method. In different situations, the repeated measures model or a pattern mixture model was appropriate. Using these model-based procedures on the observed data provided varying estimates of treatment difference when compared to the published ANCOVA analysis. The conclusion of the significance test was unchanged except for the comparison of the EQ5D scores between those who did and did not receive calcium supplementation. The direction of the difference was unchanged (those without calcium supplementation showed better EQ5D scores). However, the resultant p-value did differ. The reported result showed borderline significance ($p=0.05$), but the repeated measures model and two of the pattern mixture models (ACMV and NCMV restrictions) gave p-values less than 0.05. These models also allowed more patients to be included in the analysis, thus increasing the power to detect the difference.

Table 9.5: RECORD - Results from model-based analysis when reminder data were missing

		2 year comparison: Calcium					2 year comparison: Vitamin D				
Model	N	Estimate	95% CI	p-value	Bias	Ratio	Estimate	95% CI	p-value	Bias	Ratio
EQ5D											
Observed data	**3204**	**0.015**	**(0.00, 0.03)**	**0.05**	**-**	**-**	**-0.002**	**(-0.017, 0.013)**	**0.81**	**-**	**-**
Repeated measures (UN)	2265	0.016	(-0.001, 0.032)	0.07	0.001	1.10	0.004	(-0.012, 0.021)	0.62	0.006	1.10
Pattern mixture (CCMV)	1913	0.011	(-0.007, 0.029)	0.23	0.004	1.20	0.004	(-0.014, 0.023)	0.63	0.006	1.23
Pattern mixture (ACMV)	1913	0.012	(-0.007, 0.030)	0.21	0.003	1.23	0.004	(-0.015, 0.022)	0.68	0.006	1.23
Pattern mixture (NCMV)	1913	0.016	(-0.003, 0.034)	0.09	0.001	1.23	0.003	(-0.015, 0.021)	0.78	0.005	1.20
SF12 physical component score (PCS)											
Observed data	**3149**	**0.44**	**(-0.17, 1.05)**	**0.16**	**-**	**-**	**0.16**	**(-0.45, 0.77)**	**0.61**	**-**	**-**
Repeated measures (UN)	2140	0.59	(-0.11, 1.30)	0.10	0.15	1.15	0.48	(-0.22, 1.19)	0.18	0.32	1.16
Pattern mixture (CCMV)	1786	0.61	(-0.15, 1.36)	0.11	0.17	1.24	0.24	(-0.50, 0.99)	0.52	0.08	1.22
Pattern mixture (ACMV)	1786	0.63	(-0.13, 1.38)	0.10	0.19	1.24	0.22	(-0.53, 0.97)	0.57	0.06	1.23
Pattern mixture (NCMV)	1786	0.71	(-0.035, 1.45)	0.06	0.27	1.22	0.22	(-0.51, 0.96)	0.55	0.06	1.20
SF12 mental component score (MCS)											
Observed data	**3149**	**0.030**	**(-0.63, 0.68)**	**0.94**	**-**	**-**	**-0.02**	**(-0.68, 0.68)**	**0.94**	**-**	**-**
Repeated measures (UN)	2140	-0.023	(-0.78, 0.73)	0.95	0.053	1.15	-0.10	(-0.85, 0.66)	0.80	0.08	1.11
Pattern mixture (CCMV)	1786	-0.020	(-0.80, 0.77)	0.97	0.05	1.20	<0.001	(-0.78, 0.78)	>0.99	0.020	1.15
Pattern mixture (ACMV)	1786	0.013	(-0.77, 0.80)	0.97	0.017	1.20	-0.017	(-0.79, 0.76)	0.97	0.003	1.14
Pattern mixture (NCMV)	1786	0.150	(-0.63, 0.92)	0.71	0.12	1.18	-0.004	(-0.77, 0.77)	0.99	0.016	1.13

Table 9.6: RECORD - Results from model-based analysis on the observed data

		2 year comparison: Calcium			2 year comparison: Vitamin D		
Model	N	Estimate	95% CI	p-value	Estimate	95% CI	p-value
EQ5D							
Observed data (N=3204)	**3204**	**0.015**	**(0.00, 0.03)**	**0.05**	**-0.002**	**(-0.017, 0.013)**	**0.81**
Repeated measures (UN)	3906	0.017	(0.003, 0.032)	0.02	0.006	(-0.009, 0.021)	0.46
Pattern mixture (CCMV)	3634	0.015	(0.0001, 0.030)	0.05	0.003	(-0.011, 0.018)	0.66
Pattern mixture (ACMV)	3634	0.016	(0.0006, 0.031)	0.04	0.004	(-0.010, 0.018)	0.62
Pattern mixture (NCMV)	3634	0.018	(0.0002, 0.033)	0.03	0.010	(-0.009, 0.021)	0.45
SF12 physical component score (PCS)							
Observed data (N=3149)	**3149**	**0.44**	**(-0.17, 1.05)**	**0.16**	**0.16**	**(-0.45, 0.77)**	**0.61**
Repeated measures (UN)	3643	0.11	(-0.49, 0.71)	0.72	0.42	(-0.18, 1.02)	0.17
Pattern mixture (CCMV)	3355	0.38	(-0.25, 1.02)	0.24	0.13	(-0.47, 0.73)	0.67
Pattern mixture (ACMV)	3355	0.42	(-0.19, 1.03)	0.18	0.10	(-0.50, 0.69)	0.75
Pattern mixture (NCMV)	3355	0.37	(-0.27, 1.004)	0.25	0.04	(-0.55, 0.63)	0.90
SF12 mental component score							
Observed data (N=3149)	**3149**	**0.03**	**(-0.63, 0.68)**	**0.94**	**-0.02**	**(-0.68, 0.68)**	**0.94**
Repeated measures (UN)	3643	0.03	(-0.62, 0.69)	0.92	-0.01	(-0.66, 0.65)	0.98
Pattern mixture (CCMV)	3355	-0.10	(-0.77, 0.57)	0.76	-0.14	(-0.78, 0.49)	0.66
Pattern mixture (ACMV)	3355	-0.07	(-0.70, 0.59)	0.86	-0.15	(-0.78, 0.48)	0.64
Pattern mixture (NCMV)	3355	-0.07	(-0.75, 0.61)	0.83	-0.12	(-0.75, 0.52)	0.72

9.5 KAT

9.5.1 Model-based analysis strategies when reminder data were missing

The results of the model-based analysis strategies for the two year treatment comparison (patella resurfacing versus no patella resurfacing) are shown in Table 9.7. These estimates were based on the dataset of responders at two years, with reminder-responses set to missing. For the EQ5D score, the result under the repeated measures model was close to that which was observed. It was very close to being unbiased (b=0.004) with reasonable precision (ratio = 0.92). The estimate of treatment difference under the pattern mixture model showed the opposite direction, but did still show no significant treatment difference. The precision was not adequate, as the width of the confidence interval was larger.

Table 9.7: KAT – Results from model-based analysis when reminder data were missing

		2 year treatment comparison				
Model	N	Difference	95% CI	p-value	Bias	Ratio
EQ5D						
Observed data	**1378**	**0.013**	**(-0.01, 0.04)**	**0.35**	-	-
Repeated measures (UN)	1378	0.009	(-0.019, 0.036)	0.53	0.004	0.92
Pattern mixture (CCMV)	1100	-0.005	(-0.04, 0.03)	0.76	0.018	1.40
Pattern mixture (ACMV)	1100	-0.006	(-0.04, 0.03)	0.75	0.019	1.40
Pattern mixture (NCMV)	1100	-0.006	(-0.04, 0.03)	0.75	0.019	1.40
SF12 physical component score						
Observed data	**1361**	**-0.10**	**(-1.19, 1.00)**	**0.86**	-	-
Repeated measures (UN)	1361	-0.04	(-1.21, 1.12)	0.95	0.06	1.06
Pattern mixture (CCMV)	1063	0.10	(-1.27, 1.48)	0.88	0.20	1.23
Pattern mixture (ACMV)	1063	0.10	(-1.28, 1.47)	0.89	0.20	1.26
Pattern mixture (NCMV)	1063	-0.003	(-1.42, 1.42)	0.99	0.10	1.30
SF12 mental component score						
Observed data	**1361**	**0.29**	**(-0.74, 1.31)**	**0.58**	-	-
Repeated measures (UN)	1361	0.15	(-0.94, 1.24)	0.48	0.14	1.06
Pattern mixture (CCMV)	1063	-0.12	(-1.36, 1.13)	0.85	0.41	1.21
Pattern mixture (ACMV)	1063	-0.13	(-1.38, 1.11)	0.83	0.42	1.21
Pattern mixture (NCMV)	1063	-0.22	(-1.51, 1.06)	0.73	0.51	1.25
Oxford Knee Score (OKS)						
Observed data	**1372**	**0.27**	**(-0.86, 1.39)**	**0.64**	-	-
Repeated measures (UN)	1372	0.40	(-0.71, 1.51)	0.79	0.13	0.99
Pattern mixture (CCMV)	1008	-0.58	(-2.07, 0.92)	0.45	0.85	1.33
Pattern mixture (ACMV)	1008	-0.46	(-2.44, 1.52)	0.63	0.73	1.76
Pattern mixture (NCMV)	1008	-0.57	(-2.14, 0.99)	0.47	0.84	1.39

For the PCS, both the repeated measures model and the NCMV pattern mixture model provide treatment estimates in the direction of that which as observed. The

less biased of the two methods was the repeated measures model, which also showed the better precision. For both the MCS and OKS, the repeated measures model was least biased and most precise. The pattern mixture models, although consistent with each other, provided estimates of treatment difference in the opposite direction to the actual observed estimate.

9.5.2 Model-based analysis strategies on the observed data

Table 9.8 displays the results of the modelling process on all the actual observed data. For the EQ5D score, each of the model-based methods provided a result very close to that seen with the observed complete-case analysis.

Table 9.8: KAT – Results from model-based analysis on the observed data

		2 year treatment comparison		
Model	N	Estimate	95% CI	p-value
EQ5D				
Observed data	**1378**	**0.013**	**(-0.01, 0.04)**	**0.35**
Repeated measures (UN)	1598	0.014	(-0.01, 0.04)	0.29
Pattern mixture (CCMV)	1460	0.014	(-0.01, 0.04)	0.31
Pattern mixture (ACMV)	1460	0.014	(-0.01, 0.04)	0.34
Pattern mixture (NCMV)	1460	0.010	(-0.02, 0.04)	0.43
SF12 physical component score (PCS)				
Observed data	**1361**	**-0.10**	**(-1.19, 1.00)**	**0.86**
Repeated measures (UN)	1572	0.033	(-1.03, 1.10)	0.95
Pattern mixture (CCMV)	1419	0.17	(-0.97, 1.31)	0.77
Pattern mixture (ACMV)	1419	0.15	(-1.10, 1.39)	0.82
Pattern mixture (NCMV)	1419	0.16	(-0.99, 1.32)	0.79
SF12 mental component score (MCS)				
Observed data	**1361**	**0.29**	**(-0.74, 1.31)**	**0.58**
Repeated measures (UN)	1572	0.23	(-0.78, 1.23)	0.66
Pattern mixture (CCMV)	1383	-0.15	(-1.21, 0.91)	0.78
Pattern mixture (ACMV)	1383	-0.18	(-1.33, 0.97)	0.76
Pattern mixture (NCMV)	1383	-0.28	(-1.36, 0.80)	0.61
Oxford Knee Score (OKS)				
Observed data	**1372**	**0.27**	**(-0.86, 1.39)**	**0.64**
Repeated measures (UN)	1591	0.57	(-0.47, 1.60)	0.28
Pattern mixture (CCMV)	1460	-0.24	(-1.48, 1.01)	0.71
Pattern mixture (ACMV)	1460	-0.23	(-1.48, 1.01)	0.71
Pattern mixture (NCMV)	1460	-0.11	(-1.39, 1.15)	0.86

For the PCS, the modelling methods provided an estimate of treatment difference in the opposite direction to the observed estimate, but the non-significant result was maintained. The result of the repeated measures model for the MCS was

similar to the observed data, but the pattern mixture models produced a negative estimate. Similarly, for the OKS score, the pattern mixture models provided negative estimates, while the observed difference was positive, yet non-significant.

9.5.3 Summary

The repeated measures model appeared to be the better model-based strategy of those considered. Undertaking this method on the observed data allowed a greater number of patients to be included in the analysis. The primary trial outcome was the OKS and in the published analysis 1372 (80%) of recruited patients were included. Using a repeated measures approach, where baseline QoL was as a covariate, allowed 1591 (93%) patients to be included. Using baseline QoL as part of the repeated measures, rather than as a covariate would allow all 1715 patients to be included in the analysis. This is because those who did not have the baseline assessment did provide assessment at one or more of the follow-up times. In this analysis approach, the estimate of treatment difference and 95% CI was 0.53 (-0.52, 1.59) with p=0.32. Using an analysis which includes all patients did not change the result of the original ANCOVA strategy. However, it provides a more reliable estimate as all patients are included. The mechanism of missing data in KAT was found not to be MCAR. Thus, the repeated measures approach has greater validity when this is known to be the case.

9.6 PRISM

9.6.1 Model-based analysis strategies when the reminder data were missing

Table 9.9 displays the estimate of treatment difference at two years for the different model-based strategies. All the models and the original observed analysis did not show any evidence of a treatment difference for the four QoL scores. For the EQ5D score, each of the model-based strategies showed a small bias. However, the repeated measures model and the pattern mixture model with the ACMV restriction showed good precision (ratio = 1.03). The ACMV restriction was also a good method for the PCS, despite the negative estimate of treatment

difference. In the case of the MCS, both the repeated measures model and the NCMV model provided estimates close to the observed result, with the repeated measures model showing a better precision value. The CCMV restriction appeared to be the best method for the ASHI. Both the repeated measures model and the NCMV restriction pattern mixture model provided positive treatment difference estimates compared to the negative difference observed.

Table 9.9: PRISM –Results from model-based analysis when reminder data was missing

		Two year treatment comparison				
Model	N	Estimate	95% CI	p-value	Bias	Ratio
EQ5D						
Observed data	**889**	**0.015**	**(-0.019, 0.049)**	**0.38**	-	-
Repeated measures (toeplitz)	973	0.006	(-0.03, 0.04)	0.73	0.009	1.03
Pattern mixture (CCMV)	837	0.009	(-0.03, 0.05)	0.62	0.006	1.18
Pattern mixture (ACMV)	837	0.008	(-0.03, 0.04)	0.69	0.007	1.03
Pattern mixture (NCMV)	837	0.009	(-0.03, 0.05)	0.64	0.006	1.18
SF36 physical component score (PCS)						
Observed data	**866**	**0.14**	**(-0.90, 1.17)**	**0.80**	-	-
Repeated measures (UN)	939	-0.26	(-1.34,0.81)	0.63	0.40	1.04
Pattern mixture (CCMV)	772	0.24	(-1.03, 1.51)	0.71	0.10	1.23
Pattern mixture (ACMV)	772	-0.01	(-1.18, 1.16)	0.99	0.015	1.13
Pattern mixture (NCMV)	772	-0.16	(-1.32, 1.00)	0.79	0.30	1.12
SF36 physical component score (PCS)						
Observed data	**866**	**0.69**	**(-0.65, 2.04)**	**0.31**	-	-
Repeated measures (UN)	939	0.77	(-0.65, 2.20)	0.29	0.08	1.06
Pattern mixture (CCMV)	772	0.08	(-1.41, 1.57)	0.92	0.61	1.11
Pattern mixture (ACMV)	772	0.59	(-0.89, 2.08)	0.43	0.10	1.10
Pattern mixture (NCMV)	772	0.75	(-0.75, 2.25)	0.33	0.06	1.11
Arthritis Index (ASHI)						
Observed data	**866**	**-0.04**	**(-1.22, 1.14)**	**0.95**	-	-
Repeated measures (UN)	939	0.21	(-1.18, 1.60)	0.76	0.25	1.18
Pattern mixture (CCMV)	772	-0.14	(-1.47, 1.20)	0.84	0.18	1.13
Pattern mixture (ACMV)	772	-0.24	(-1.58, 1.09)	0.72	0.28	1.13
Pattern mixture (NCMV)	772	0.77	(-0.65, 2.20)	0.29	0.81	1.21

9.6.2 Model-based analysis strategies on the observed data

Table 9.10 displays the estimate of treatment difference after implementing the model-based strategies on the observed data. The NCMV restriction provided an estimate of treatment difference equal to that which was observed for the EQ5D score. It also provided a similar confidence interval and resultant p-value, but utilised a greater number of patients (N=1118). In the case of the PCS, the repeated measures model was closest to the observed result. For the MCS, the repeated measures model and the ACMV restriction provided results close to the

observed value. Finally, for the ASHI, the repeated measures model on all missing data provided a result close to the observed value for the complete-case analysis.

Table 9.10: PRISM - Results from model-based analysis on observed data

		Two year treatment comparison		
Model	N	Estimate	95% CI	p-value
EQ5D				
Observed data	**889**	**0.015**	**(-0.019, 0.049)**	**0.38**
Repeated measures (UN)	1250	0.011	(-0.022, 0.044)	0.52
Pattern mixture (CCMV)	1118	0.019	(-0.014, 0.052)	0.26
Pattern mixture (ACMV)	1118	0.013	(-0.019, 0.046)	0.42
Pattern mixture (NCMV)	1118	0.015	(-0.018, 0.048)	0.37
SF36 physical component score (PCS)				
Observed data	**866**	**0.14**	**(-0.90, 1.17)**	**0.80**
Repeated measures (UN)	1198	0.09	(-0.89, 1.08)	0.85
Pattern mixture (CCMV)	1023	0.80	(-0.29, 1.89)	0.15
Pattern mixture (ACMV)	1023	0.34	(-0.65, 1.33)	0.50
Pattern mixture (NCMV)	1023	0.25	(-0.75, 1.24)	0.63
SF36 mental component score (MCS)				
Observed data	**866**	**0.69**	**(-0.65, 2.04)**	**0.31**
Repeated measures (UN)	1198	0.75	(-0.56, 2.05)	0.26
Pattern mixture (CCMV)	1023	-0.57	(-1.92, 0.77)	0.40
Pattern mixture (ACMV)	1023	0.71	(-0.59, 2.01)	0.28
Pattern mixture (NCMV)	1023	1.12	(-0.23, 2.46)	0.10
Arthritis Index (ASHI)				
Observed data	**866**	**-0.04**	**(-1.22, 1.14)**	**0.95**
Repeated measures (UN)	1198	-0.09	(-1.22, 1.04)	0.87
Pattern mixture (CCMV)	1023	0.22	(-1.01, 1.46)	0.72
Pattern mixture (ACMV)	1023	0.20	(-0.95, 1.35)	0.73
Pattern mixture (NCMV)	1023	0.23	(-0.94, 1.39)	0.70

9.6.3 Summary

Firstly, the model-based strategies were undertaken on the dataset containing only the responders at two years and the reminder-responses were set to missing. Overall, the repeated measures model seemed to be the most appropriate providing small bias and good precision. The only exception was in the case of the PCS, where the pattern mixture model showed less bias, yet this was counteracted by a poorer precision value. Implementing the model-based strategies on all the observed data found some inconsistencies between results. The estimates were fairly consistent between methods for the EQ5D score, with the repeated measures model utilising a greater number of patients. However, for the PCS, MCS and ASHI, results were less consistent. The magnitude of treatment difference differed for the PCS, while both magnitude and direction differed for

the MCS and ASHI. Despite the differences between the different analyses, all were consistent in finding no evidence of a treatment difference in QoL scores at two years.

9.7 TOMBOLA

9.7.1 Model-based analysis strategies when the reminder data were missing

Table 9.11 displays the results of the repeated measures models and pattern mixture models for TOMBOLA when the reminder data are missing. Estimates of treatment difference in EQ5D at 12 and 30 months are given for the R1 (cytology vs. colposcopy) and R2 (biopsy and selected recall vs. immediate treatment) comparisons. For the R1 comparison at 12 months, the pattern mixture model using the NCMV restriction was the least biased. Although, all the model-based methods provided a non-significant treatment difference estimate, as was observed in the actual data.

No significant difference was found between the R1 treatment groups at 30 months by any of the model-based methods. The repeated measures model and the ACMV restriction were least biased when compared to the actual observed result, but the repeated measures model showed the better value of precision. A significant difference was observed in the R2 comparison at 12 months. This was maintained by all the model-based methods, but the repeated measures model was the least biased. In fact, it was unbiased with an estimate equal to that seen in the observed data. The confidence interval was increased and the resultant p-value borderline significant (p=0.05) compared to the observed result of p=0.03. The least biased model-based method for the R2 comparison was the repeated measures model, which also showed good precision. Two of the pattern mixture models (CCMV and ACMV) showed a significant result rather than the non-significant estimate that was observed. The discrepancies between the pattern mixture models and the other two strategies will to some extent be due to the large reduction in sample size.

Table 9.11: TOMBOLA – Results from model-based analysis when reminder data were missing

		Treatment comparison			
Model	Estimate	95% CI	p-value	Bias	Ratio
R1 comparison of EQ5D at 12 months					
Observed data (N=2294)	**-0.001**	**(-0.011, 0.008)**	**0.81**	**-**	-
Repeated measures (UN)[1]	0.003	(-0.008, 0.014)	0.64	0.004	1.16
Pattern mixture (CCMV)[2]	0.007	(-0.013, 0.027)	0.51	0.008	2.11
Pattern mixture (ACMV)[2]	0.003	(-0.013, 0.018)	0.75	0.004	1.63
Pattern mixture (NCMV)[2]	-0.0002	(-0.019, 0.018)	0.99	0.0008	1.95
R1 comparison of EQ5D at 30 months					
Observed data (N=1825)	**0.004**	**(-0.007, 0.015)**	**0.49**	-	**-**
Repeated measures (UN)[3]	0.005	(-0.008, 0.017)	0.46	0.001	1.14
Pattern mixture (CCMV)[4]	0.010	(-0.013, 0.033)	0.41	0.006	2.09
Pattern mixture (ACMV)[4]	0.005	(-0.014, 0.025)	0.58	0.001	1.77
Pattern mixture (NCMV)[4]	0.010	(-0.022, 0.037)	0.60	0.006	2.68
R2 comparison of EQ5D at 12 months					
Observed data (N=709)	**0.019**	**(0.002, 0.035)**	**0.03**	-	-
Repeated measures (UN)[5]	0.019	(0.0001, 0.037)	0.05	0.000	1.12
Pattern mixture (CCMV)[6]	0.034	(0.008, 0.059)	0.012	0.015	1.55
Pattern mixture (ACMV)[6]	0.037	(0.013, 0.060)	0.015	0.018	1.42
Pattern mixture (NCMV)[6]	0.030	(0.004, 0.056)	0.022	0.011	1.58
R2 comparison of EQ5D at 30 months					
Observed data (N=584)	**0.006**	**(-0.014, 0.026)**	**0.56**	-	-
Repeated measures (UN)[7]	0.002	(-0.017, 0.021)	0.83	0.004	0.95
Pattern mixture (CCMV)[8]	0.030	(0.003, 0.058)	0.03	0.024	1.38
Pattern mixture (ACMV)[8]	0.027	(0.005, 0.049)	0.02	0.021	1.10
Pattern mixture (NCMV)[8]	0.029	(-0.006, 0.065)	0.11	0.023	1.78

R1 – cytology vs. colposcopy; R2 - biopsy and selected recall vs. immediate treatment; [1] N=2074; [2] N= 776; [3] N=1660 ; [4] N=540; [5] N=644; [6] N=326; [7] N=533; [8] N=239;

9.7.2 Model-based analysis strategies on the observed data

Table 9.12 shows the results of the model-based strategies carried out on all the observed data. The results under the different strategies for the R1 comparison at 12 months were fairly consistent and in line with that found in the observed ANCOVA analysis. The main difference between the approaches is that the repeated measures model includes a greater number of the participants. Therefore, the results are potentially more reliable (assuming the data are MAR). In the R1 comparison, there were 3300 (97%) patients included in analysis using a repeated measures model, compared to only 2294 (67%) using the ANCOVA. The results at 30 months were consistent between methods and each showed a non-significant treatment difference. The ANCOVA utilised 54% of participants and whilst the repeated measures model increased this to 97%, the estimate of

treatment difference at 30 months differed by only 0.001 units. Therefore, the results of the analysis seemed reliable.

Table 9.12: TOMBOLA – Results from model-based analysis on observed data

		Treatment difference		
Model	N	Estimate	95% CI	p-value
R1 comparison of EQ5D at 12 months				
Observed data	**2294**	**-0.001**	**(-0.011, 0.008)**	**0.81**
Repeated measures (UN)	3300	-0.0003	(-0.010, 0.009)	0.94
Pattern mixture (CCMV)	1895	-0.004	(-0.015, 0.007)	0.45
Pattern mixture (ACMV)	1895	-0.005	(-0.016, 0.005)	0.35
Pattern mixture (NCMV)	1895	-0.003	(-0.020, 0.014)	0.72
R1 comparison of EQ5D at 30 months				
Observed data	**1825**	**0.004**	**(-0.007, 0.015)**	**0.49**
Repeated measures (UN)	3300	0.003	(-0.007, 0.014)	0.55
Pattern mixture (CCMV)	1895	0.004	(-0.009, 0.017)	0.55
Pattern mixture (ACMV)	1895	0.002	(-0.10, 0.013)	0.79
Pattern mixture (NCMV)	1895	0.000	(-0.023, 0.020)	0.90
R2 comparison of EQ5D at 12 months				
Observed data	**709**	**0.019**	**(0.002, 0.035)**	**0.03**
Repeated measures (UN)	966	0.017	(0.0003, 0.033)	0.05
Pattern mixture (CCMV)	684	0.019	(0.0003, 0.038)	0.05
Pattern mixture (ACMV)	684	0.015	(-0.006, 0.031)	0.08
Pattern mixture (NCMV)	684	0.017	(-0.006, 0.031)	0.20
R2 comparison of EQ5D at 30 months				
Observed data	**584**	**0.006**	**(-0.014, 0.026)**	**0.56**
Repeated measures (UN)	966	0.002	(-0.017, 0.021)	0.83
Pattern mixture (CCMV)	684	0.012	(-0.010, 0.033)	0.28
Pattern mixture (ACMV)	684	0.006	(-0.011, 0.023)	0.46
Pattern mixture (NCMV)	684	0.002	(-0.021, 0.025)	0.87

There were potentially 986 participants to be included in the R2 comparison. However, the observed ANCOVA used only 709 (72%) at 12 months and 584 (59%) at 30 months. A repeated measures approach allowed 966 (98%) of participants to be included. For the R2 comparison at 12 months, the pattern mixture model with CCMV restriction showed a similar result to that which was observed. Using ACMV provided a borderline significant result and NCMV provided a non-significant result. The repeated measures model found an estimate of treatment difference of 0.002 units different to the ANCOVA and p=0.05 rather than 0.03. The different strategies would have provided different conclusions for the R2 comparison at 12 months. For the 30 month comparison of R2 treatment groups, the ACMV restriction result was similar to the observed ANCOVA, but utilised 100 patients more. Using a repeated measures model

reduced the estimate of treatment difference. Thus, the conclusion of no significant difference was maintained. The repeated measures model and NCMV restriction provided similar results.

9.7.3 Summary

Implementing the model-based strategies on the datasets with reminder data missing showed that the repeated measures models were perhaps most appropriate and that they allowed more patients to be included than the ANCOVA. Undertaking the difference strategies on the observed data did not alter the result for the R1 comparison, but did have an effect on the R2 comparison at 12 months. The ANCOVA provided evidence of a treatment difference ($p=0.03$), while the repeated measures model and the CCMV restriction provided borderline evidence ($p=0.05$). The other two pattern mixture model restrictions did not find any evidence of a treatment difference ($p>0.05$). The repeated measures model allowed 98% of participants to be included compared to 72% in the ANCOVA. Therefore, the results are perhaps more reliable.

9.8 Norwegian Palliative Care Study

9.8.1 Model-based analysis strategies when the reminder data were missing

Table 9.13 shows the results of the repeated measures model and the pattern mixture models implemented on the dataset, containing responders only and where reminder-responses were missing. No significant treatment difference was observed at four months for any of the three QoL dimensions of the QLQ-C30 in the original ANCOVA. This result was also found by each of the model-based strategies. There was a difference in the magnitude and direction of the estimate under the various strategies. The direction of the estimate for the emotional functioning score from each of the different model-based strategies was in the opposite direction to the ANCOVA. The closest estimate was seen in the repeated measures model, which also had the better value for precision. The same was true for the physical functioning score at four months. In the case of the pain score at four months, it was the pattern mixture model with the ACMV restriction which showed least bias and reasonable precision.

Table 9.13: NPC Trial – Results from model-based analysis when reminder data was missing

		Treatment comparison at 4 months				
Model	N	Estimate	95% CI	p-value	Bias	Ratio
Emotional Functioning						
Observed Data	**132**	**1.35**	**(-4.97, 7.66)**	**0.67**	**-**	**-**
Repeated measures (CS)	152	-0.95	(-8.11, 6.20)	0.79	2.30	1.13
Pattern mixture (CCMV)	92	-5.44	(-21.9, 11.1)	0.52	6.79	2.61
Pattern mixture (ACMV)	92	-3.97	(-20.2, 12.3)	0.63	5.32	2.57
Pattern mixture (NCMV)	92	-2.33	(-21.8, 17.1)	0.81	3.68	3.08
Physical Functioning						
Observed data	**135**	**-5.68**	**(-15.1, 3.73)**	**0.23**	**-**	**-**
Repeated measures (toeplitz)	152	-3.48	(-12.5, 5.57)	0.45	2.20	0.96
Pattern mixture (CCMV)	91	2.51	(-13.7, 18.7)	0.76	8.19	1.72
Pattern mixture (ACMV)	91	2.94	(-12.5, 18.4)	0.71	8.62	1.64
Pattern mixture (NCMV)	91	2.52	(-14.7, 19.7)	0.77	8.20	1.88
Pain						
Observed data	**136**	**3.26**	**(-6.69, 13.2)**	**0.52**	**-**	**-**
Repeated measures (AR(1))	153	1.56	(-8.90, 12.0)	0.77	1.70	1.05
Pattern mixture (CCMV)	90	2.56	(-6.14, 11.3)	0.56	0.70	0.88
Pattern mixture (ACMV)	90	2.90	(-5.91, 11.8)	0.51	0.36	0.89
Pattern mixture (NCMV)	90	5.29	(-3.82, 14.4)	0.25	2.03	0.92

9.8.2 Model-based analysis strategies on the observed data

Table 9.14 displays the results of the modelling process carried out on the actual observed dataset.

Table 9.14: NPC Trial – Results from model-based analysis on observed data

		4 month treatment comparison		
Model	N	Estimate	95% CI	p-value
Emotional functioning				
Observed data	**132**	**1.35**	**(-4.97, 7.66)**	**0.67**
Repeated measures (UN)	433	-3.73	(-9.83, 2.37)	0.23
Pattern mixture (CCMV)	402	4.79	(-3.54, 13.1)	0.25
Pattern mixture (ACMV)	402	5.83	(-2.15, 13.8)	0.15
Pattern mixture (NCMV)	402	0.88	(2.07, 15.7)	0.011
Physical functioning				
Observed data	**135**	**-5.68**	**(-15.1, 3.73)**	**0.23**
Repeated measures (UN)	413	-7.19	(-16.0, 1.66)	0.11
Pattern mixture (CCMV)	403	-3.81	(-11.2, 3.53)	0.30
Pattern mixture (ACMV)	403	-2.02	(-9.93, 5.88)	0.61
Pattern mixture (NCMV)	403	-4.55	(-11.9, 2.84)	0.23
Pain				
Observed data	**136**	**3.26**	**(-6.69, 13.2)**	**0.52**
Repeated measures (toeplitz)	433	3.45	(-5.52, 12.4)	0.45
Pattern mixture (CCMV)	405	1.70	(-3.64, 7.03)	0.52
Pattern mixture (ACMV)	405	0.91	(-3.56, 5.38)	0.69
Pattern mixture (NCMV)	405	0.49	(-4.83, 5.80)	0.86

All but one of the strategies showed no significant difference in QoL scores at four months between the treatment groups. There was one exception for the emotional functioning score, where the NCMV restriction showed a significant treatment difference. The estimates of treatment difference did vary in magnitude and direction with the various model-based strategies. Utilising a repeated measures model allowed over 95% of participants to be included in the analysis. The direction of effect was consistent for the physical functioning and pain scores, but did differ for the emotional functioning score when compared to the ANCOVA.

9.8.3 Joint mixed effects model

The joint mixed effects model was proposed in section 8.4.2 to model longitudinal QoL data, in conjunction with time to event (e.g. time of dropout or time of death). Previous chapters have shown that in the NPC trial, there was a high proportion of dropout due to death and therefore, this type of model-based procedure can be considered here. The first stage of this process uses a growth curve model (piecewise linear model) within the mixed effects model framework, where changes in QoL are assumed to be linear over short intervals of time. This approach was undertaken on the pain, physical functioning and emotional functioning dimensions of the QLQ-C30. The results are shown in Table 9.15.

Table 9.15: NPC Trial – growth curve model results

Treatment comparison	Difference	SE	95%CI	p-value
Pain (N=431)				
Difference at 1m	-2.97	3.58	(-9.99,4.06)	0.41
Difference at 2m	2.66	6.08	(-9.27,14.6)	0.66
Difference in slope (baseline to 1m)	-2.04	3.31	(-8.54,4.47)	0.54
Difference in slope (1m to 2m)	3.60	4.23	(-4.71,11.9)	0.40
Physical functioning (N=434)				
Difference at 1m	-2.53	3.22	(-8.86,3.79)	0.43
Difference at 2m	-1.94	4.88	(-11.5,7.65)	0.69
Difference in slope (baseline to 1m)	-0.91	2.60	(-6.02,4.20)	0.73
Difference in change (1m to 2m)	-0.31	3.25	(-6.70,6.07)	0.92
Emotional functioning (N=432)				
Difference at 1m	1.61	2.66	(-3.61,6.82)	0.55
Difference at 2m	1.87	4.22	(-6.42,10.2)	0.66
Difference in slope (baseline to 1m)	-0.33	2.26	(-4.76,4.10)	0.88
Difference in change (1m to 2m)	-0.06	2.87	(-5.71,5.58)	0.98

The final slope change was at two months. Therefore, the estimate of treatment difference at four months was the same as at two months. No significant differences were found between treatment groups at either one or two months. The change in QoL from baseline to one month and from one month to two months was not significantly different between treatment groups for each of the three dimension scores.

The next stage was to incorporate the time to event (time to death) in order to create a joint model which would model the longitudinal outcome of QoL alongside the time to death. The initial estimates for the joint model were provided from the growth curve model and the survival model. These were implemented using PROC NLMIXED (SAS Institute Inc. 2004). This allowed estimates of the difference in slopes between groups to be assessed along with the correlation of dropout with the random effects (intercept and slope). The results of this are shown in Table 9.16.

Table 9.16: NPC Trial – joint mixed effect model results

				Correlation*	
Treatment comparison	Difference	95% CI	p-value	Intercept	Slope
Pain (N=431)					
Difference in slope (baseline to 1m)	-2.09	(-8.61, 4.44)	0.53	-0.03	0.05
Difference in slope (1m to 2m)	3.57	(-4.76, 11.9)	0.40		
Physical functioning (N=434)					
Difference in slope (baseline to 1m)	-0.87	(-6.00, 4.25)	0.74	0.06	-0.03
Difference in slope (1m to 2m)	-0.38	(-6.77, 6.01)	0.91		
Emotional functioning (N=432)					
Difference in slope (baseline to 1m)	-0.29	(-4.72, 4.14)	0.90	0.05	0.23
Difference in slope (1m to 2m)	0.01	(-5.64, 5.66)	0.99		

*Dropout event is time to death

The estimates of the treatment difference from the joint mixed effects model did not differ greatly from the results of the basic mixed effects model. For each of the three QoL scores, weak correlation existed with the intercept and the slope. The strongest correlation occurred for the slope and emotional functioning ($\rho = 0.23$), yet this is still regarded as weak. Since there does not appear to be much correlation between dropout and QoL, the standard mixed-effects model is likely to be sufficient in this instance.

9.8.4 Summary

Several different modelling strategies were carried out for the NPC trial. In each case, no significant treatment difference was found between pain, physical functioning and emotional functioning scores. The high dropout rate in the trial (largely due to death) meant that the ANCOVA did not utilise that many patients. The repeated measures approach, pattern mixture model and joint mixed effects model incorporated the majority of patients and made different assumptions about the missing data mechanism. There was some evidence that the missing data was not MCAR (chapter four). The use of Fairclough's logistic regression on the reminder responses did not find any evidence of MNAR. This suggested that perhaps the pattern mixture model and joint mixed-effects model were less appropriate than the standard mixed-effects model. The results of the joint mixed effects model did not differ greatly from the standard growth curve model. Weak correlation of dropout to QoL was found. This implied that the joint mixed effects model was not appropriate in this particular dataset.

9.9 Discussion

On the whole, the use of a repeated measures model appeared to provide a more valid approach than the ANCOVA. Despite the fact that these two methods did not provide estimates of treatment difference vastly different from each other, the repeated measures model did include a greater number of patients. For example, in the REFLUX trial, only 77% of patients were included in ANCOVA for the primary outcome (RQLS), compared to 92% in the repeated measures analysis. Table 9.17 shows what was considered as the best model-based strategy for each of the trials. However, there is a caveat to consider when using a repeated measures approach. The advantage of the repeated measures model is that more patients are included, yet they are done so assuming MAR. If this is not the case, then repeated measures may not be suitable.

The pattern mixture models performed less well. This would partly be due to the fact that they were limited to only those patients with a monotone missing data

pattern. The example datasets contained both monotone and intermittent missingness making this type of approach of limited applicability. A more robust analysis would use as many patients as possible. An alternative to the pattern mixture models described, that is applicable for non-monotone missing data, was by Troxel *et al*. (Troxel, Lipsitz & Harrington 1998). This is an extension of the model by Diggle and Kenward (Diggle, Kenward 1994) and describes a likelihood method with Markovian correlation structure, in which the missingness follows a logistic model. Numerical integration is used to solve the problem. This model has not been implemented here, but would be of interest to consider in future work.

Table 9.17: Summary of 'best' model-based strategy for each trial

Trial	QoL Measure	Model-based strategy
REFLUX	RQLS	Repeated measures
	EQ5D	Repeated measures
	SF12	Repeated measures
MAVIS	EQ5D	Pattern mixture model
	SF12	Pattern mixture model
RECORD	EQ5D	Repeated measures
	SF12	Repeated measures
KAT	OKS	Repeated measures
	EQ5D	Repeated measures
	SF12	Repeated measures
PRISM	EQ5D	Pattern mixture model
	SF36	Repeated measures
TOMBOLA	EQ5D	Repeated measures
NPC Trial	QLQ-C30	Repeated measures

Each of the different model-based strategies makes assumptions about the missing data mechanism. The ANCOVA assumes that the complete cases are representative of all patients in the sample and that the data are MCAR. A repeated measures approach assumes MAR, while a pattern mixture model and joint mixed effects model can be set up such that MNAR is assumed. Given that the missing data within the example trials were found to be more likely MAR or MNAR, it seems sensible to think that a repeated measures model or one that takes account of the non-ignorable dropout might be more appropriate than the ANCOVA. In particular, in the situation of a large proportion of dropouts (for example due to death) a joint mixed effects model is useful. Despite the fact that this type of model proved not to be appropriate here, this type of modelling may

be appropriate in other situations. In particular, this is the case when there is some evidence of a treatment difference alongside a large number of dropouts.

9.9.1 Conclusion

The analysis of longitudinal data with potentially informative dropout involves assumptions which are difficult to check from observed data. It is unwise to rely on precise conclusions of an analysis based on a particular informative dropout model (Diggle et al. 2002). The usefulness of some of the different model-based strategies could be assessed in a sensitivity analysis. This is consistent with the conclusion set out in chapter seven, that there was no single imputation method found to be applicable in all situations. Knowing the missing data mechanism will ultimately help in determining which, if any, imputation or model-based strategy is appropriate.

The work presented in this chapter shows that repeated measures approaches have an important role to play in the analysis of longitudinal QoL outcomes. They make the assumption of MAR which is more plausible for QoL. In the context of the example trials, they allowed a greater number of patients to be included in the analysis. Thus, this improves the power of the study and reduces bias. Pattern mixture models or joint models could be considered if the data are thought to be MNAR. However, the model for dropout needs to be considered carefully.

Chapter 10 Evaluating the economic benefit of different data collection strategies

10.1 Introduction

Within health care, decisions must be made on where to allocate scarce resources in order to result in the most benefit. Economic evaluation aims to aid priority setting and allocation of resources. As part of a clinical trial, cost-effectiveness analyses are often carried out, as it is no longer sufficient to determine if a new intervention is effective, but also to determine if the new treatment is cost-effective. For example, in the REFLUX trial, the remit was not only to determine the effectiveness of surgery compared to medical management, but also to determine the cost-effectiveness (Grant et al. 2008; Grant et al. 2009). Drummond *et al.* define economic evaluation as

> '…the comparative analysis of alternative courses of action in terms of both their costs and consequences' (Drummond et al. 1997).

The basic idea behind an economic evaluation is to identify, measure and value both costs and benefits of competing alternatives. These alternatives may be two different treatments of a particular disease or alternative methods of collecting data for use in an evaluation.

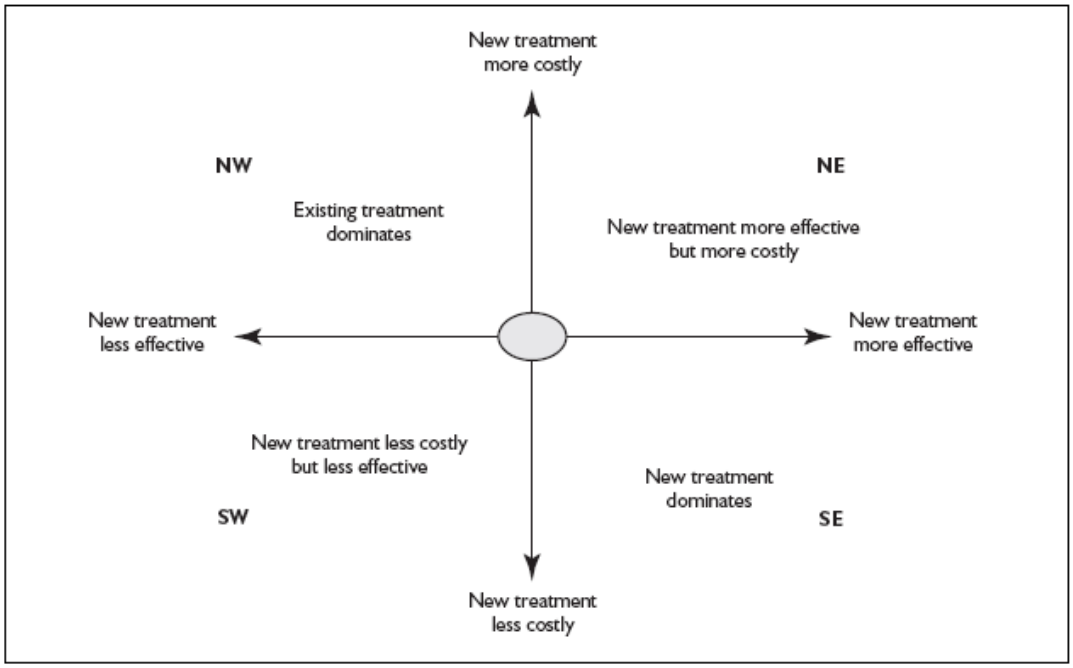
In this thesis, the strategies under investigation are the use of reminders to collect data and employing imputation for missing data. The reminder system incurs the costs of printing additional questionnaires, phone calls, mailing and additional staff time to identify, administer reminders and input subsequent data. The main additional cost of imputation is the staff time to process the data, identify missing values and program the imputation methods into statistical software. The benefit of interest is the ability to draw a correct conclusion about the effectiveness or efficiency of the interventions under consideration; quantity and quality (how accurate and reliable the data are) of additional data act as a proxy for this. These issues are discussed further throughout this chapter.

The aim of this chapter is to determine if the reminder system and/or imputation are cost-effective in improving the precision of the eventual measure of treatment difference. Within this framework, there is the initial assumption that all types of data are of equal value: i.e. the original responses, reminder responses and those under imputation are of equal value. Clearly, data received after reminder are likely to reflect a time frame shifted on two (or four) weeks from the original time of assessment. This may not truly represent the QoL at the time when they were originally due to be assessed. As discussed in chapters six and seven, data obtained through imputation are not as reliable as truly observed data and may be subject to a certain amount of bias. This chapter will take these issues into account, while also considering the economic implications of different data collection strategies. This chapter begins with an overview of some principles of economic evaluation that will be utilised later in the chapter.

10.1.1 Cost-effectiveness analysis

Cost-effectiveness analysis (CEA) is one form of economic evaluation, where both the costs and consequences of health programmes or treatments are examined. The cost-effectiveness plane (Figure 10.1) describes this basic concept visually (Drummond et al. 1997).

Figure 10.1: The cost-effectiveness plane



The horizontal axis represents the difference in effectiveness of a possible new treatment and the existing standard treatment. The two treatments are deemed equally efficient if the estimate of effectiveness and cost of the new treatment fall within the circle. That is to say, there is no evidence that there are any important differences in the costs and benefits. If the estimates for the new treatment fall in the north-west (NW) quadrant, then the new treatment is dominated by the standard treatment as it is more costly, but less effective. Conversely, if the new treatment lies in the south-east (SE) quadrant, then it dominates the standard treatment, as it is less costly and more effective.

In the case of the final two quadrants, south-west (SW) and north-east (NE) a judgement must be made as to whether the extra cost of the treatment is worth the additional effectiveness. In the SW quadrant, the new treatment is less costly, but also less effective. In the NE quadrant, the new treatment is more effective, but also more costly. In these latter two quadrants, the incremental cost-effectiveness ratio (ICER) can be calculated in order to assist with decision making. The ICER is defined as the difference in costs between the new treatment and standard treatment divided by the difference in benefits (effectiveness) observed between the new and standard treatments. The higher the ICER of the new intervention in comparison with current standard treatment, the less likely the new treatment will be considered cost-effective.

As stated above, rather than comparing treatment options, alternative data collection strategies are compared. However, the theory behind decision making remains the same. The idea is to determine the cost of an additional unit of (quality) data. This value is then used in a decision-making framework in order to identify the most cost-effective data collection strategy.

10.1.2 Quality weights

In the calculation of the ICER, the different strategies are implicitly given equal weighting in terms of the "quality" of data provided. Clearly, the best scenario is

that all data are collected through the initial wave of questionnaires and that there are no missing data. In reality, this will not be the case. A proportion of data are collected through the reminder system. Given the issues surrounding imputation, which have been discussed in earlier chapters, data obtained in this way are considered of less quality than actual observed data. A cost-effectiveness analysis will ignore this issue. Hence, it will give a distorted view of relative efficiency. Therefore, quality weights for the different data collection strategies should be incorporated into an economic analysis. These "quality" weights will be discussed further in section 10.2.5. This is analogous to cost utility analysis (CUA), which is used in economic evaluation. In a CUA, length of life is weighted by the quality of that life, in order to derive estimates of quality adjusted life years (QALYs). The additional costs of a treatment are compared with the utility gained as a result of the treatment (i.e. the incremental cost per QALY). Cost-utility analysis can be used to compare alternative health treatments with different outcomes. Therefore, it is a useful tool in allocating health resources.

10.1.3 Calculation of the incremental cost-effectiveness ratio

The ICER is calculated as the change in costs divided by the change in effectiveness. It indicates the cost required to obtain an additional unit of effectiveness. Let the number of data units collected through immediate responses be denoted N_T. Then, let the additional number of responses collected through reminders be denoted N_R at a cost of C_R. Similarly, the number of additional responses obtained by imputation is N_I at a cost of C_I. Then assume the quality weight associated with reminder data is W_R and with imputed data is W_I. The ICER is then defined as

$$\text{ICER} = \frac{C_R - C_I}{N_R W_R - N_I W_I}.$$

In the case that additional data (by reminder or through imputation) is regarded as perfect quality ($W_R=W_I=1$), this equation reverts to the standard formula for the ICER. A negative value of ($N_R W_R$ - $N_I W_I$) implies imputation dominates the use of reminders. The point at which imputation ceases to dominate occurs when

$N_R W_R - N_I W_I = 0$, or alternatively when $\hat{W}_I = \frac{N_R W_R}{N_I}$. A positive value of the ICER occurs when $W_I < \hat{W}_I$. This indicates the cost of obtaining an additional unit of top quality data.

10.1.4 Cost-benefit analysis

Cost-benefit analysis (CBA) requires programme consequences to be valued in monetary terms. This enables the analyst to make a direct comparison of the programmes incremental costs with its incremental consequences. In simple terms, the goal of analysis is to determine whether a programme's benefits exceed the costs, indicating the programme is worthwhile. CBA has a broader scope than CEA, due to the fact that CBA converts all costs and benefits to monetary terms. Thus, it is better suited to compare programmes both within and out with the healthcare facets (Drummond et al. 1997).

The willingness to pay (WTP) is an evaluation method used to determine the maximum amount of money an individual is willing to pay to gain a particular benefit. The method is often used in CBA in order to quantify a benefit in monetary terms. The technique of WTP is often criticised for attempting to assign monetary value to things which cannot be valued e.g. the saving of a human life. A second criticism of WTP is that it is inevitably a function of ability to pay (McIntosh, Donaldson & Ryan 1999).

10.1.5 Net benefit statistic

An alternative decision making tool to the ICER or the elicitation of WTP is the net benefit (NB) framework. Stinnett and Mullahy provide a comprehensive account of the net benefit framework (Stinnett, Mullahy 1998). The cost-effectiveness decision rule uses the value λ to represent that which you are willing to pay for the extra information. The net-benefit is then plotted as a function of λ. In the context of this thesis, a net benefit statistic can be calculated for each of the two data collection strategies: imputation (NB_I) or reminders (NB_R). A decision rule as

to which strategy is most cost-effective can then be formed. The net-benefit statistics in the context of this work are as follows:

NB_I = [(Additional data from imputation* λ) - Cost of imputation = N_I* λ - C_I

NB_R = [(Additional data from reminders* λ) - Cost of reminders = N_R* λ – C_R.

The decision rule is such that if

$$NB_I - NB_R > 0 \text{ choose imputation,}$$
$$NB_I - NB_R < 0 \text{ choose reminders.}$$

The calculation of N_I has taken into account the quality of the imputed data. Therefore, this does not need to be considered in the direct calculation of the NB statistic.

10.1.6 Sensitivity analysis

Sensitivity analysis is a technique used in economic analysis or decision-making to allow uncertainty. It tests whether plausible changes in the values of the main variables affect the results of the analysis. In any model, a specific cost and consequence is used. Sensitivity analysis allows changes in these values to determine how they affect the results. If the study results are not greatly affected by these changes, then the study results are considered to be robust.

10.2 Methods

10.2.1 Strategies for comparison

The aim of this chapter is to consider the economic consequences of the different data collection strategies. There were several strategies considered:

1) Data collected without reminder (immediate data)
2) Data collected by any means (reminder or additional telephone interviews)
3) Data collected without reminder and imputation applied.

The choice of imputation strategy will have some bearing on the complexity of computer programming and, therefore, cost. The most suitable imputation method (one simple and one multiple) identified in chapter seven are used here for illustration. Strategy one above will be assumed to be the baseline. Hence, the additional costs of the other strategies over and above this strategy will be estimated. The cost attached to strategies two and three will be the additional cost required to obtain the data, e.g. the additional cost of issuing a reminder or additional cost of computer/staff time to carry out imputation.

10.2.2 Costs of the reminder process

The cost of the reminder process is always considered when putting together a proposal for trial funding. The assumption on which this costing is based is that at each follow up, 40% of the number recruited would require a first reminder and 25% of the total recruited would require a second reminder. This equates to the number of reminders being 65% of the total number of patients recruited. The example trials are completed and it is known exactly how many reminders were required at each assessment. It is this number on which the cost of the reminder process was based, rather than the anticipated cost of reminders. Data on the cost of reminders were obtained from the Centre for Healthcare Randomised Trials (CHaRT) at the University of Aberdeen. The cost of printing, other stationery and postage per reminder questionnaire is estimated to be £3.50 (Table 10.1). The cost of a member of staff to put together this reminder is estimated as £0.58, giving a total cost of a single reminder costs £4.08. The figures presented in Table 10.1 are based on the cost as at December 2008.

Table 10.1: Cost of issuing a reminder in CHaRT trials

Item of paperwork	Cost (£)
Printing of questionnaire	1.00
Other stationary (envelopes, letterhead etc)	1.00
Postage (outward and return)	1.50
Preparation time / secretarial cost (5 minutes)	0.58
Total cost	**4.08**

Five of the example trial datasets used in this thesis, were administered by CHaRT/HSRU staff. The remaining two were not co-ordinated by this group. Therefore, the work presented in this chapter relates to the five trial datasets of REFLUX, MAVIS, RECORD, KAT and PRISM. Using this information the cost of the reminder process was obtained for each of the five HSRU trials (Table 10.2). This cost was based on the number of reminders required at the main endpoint of interest. The cost of the reminder process for the whole trial would be larger, as more assessments would need to be considered.

Table 10.2: Cost of reminder process for each of the trial datasets

Trial	N recruited	Main endpoint	Number of reminders sent*	Total Cost
REFLUX	357	1 year	203	£ 828.24
MAVIS	910	1 year	123	£ 501.84
RECORD	5292	2 years	1340	£ 5467.20
KAT	2356	2 years	855	£ 3488.40
PRISM	1324	2 years	302	£ 1232.16

* Total includes both 1st and 2nd reminders at the main endpoint

10.2.3 Cost of imputation

The cost of imputation is essentially the cost of staff time to write the necessary computer program needed to identify those patients with missing data, those requiring imputation and to carry out the chosen method. Costs of the computers and necessary software are assumed to be zero. These would have already been considered in the costs of running the trial, due to the fact that they would be needed for analysis, regardless of whether imputation was carried out. The only additional cost of imputation is the extra time needed to write and implement the imputation programs. The cost of imputation was not dependent on the sample size, as the difference in computer processing time to impute 10 or 100 data points would be minimal.

Carrying out imputation is not a simple process. The chosen method can impact on the time required to write suitable computer programs. Multiple imputation procedures exist within statistical software, but time may be needed for the researcher to understand which approach is most suitable and understand the

programming syntax. Depending on the experience of the trial researcher, a qualified statistician may be required to carry out the imputation. The Medical Statistics Team at the University of Aberdeen currently (at January 2008) charge £60 per hour for a consultation fee. For the purposes of what follows, the cost of carrying out imputation will be based on this £60 per hour rate. This cost of imputation does not vary with the amount of missing data. However, it will vary according to the time taken to carry out imputation.

It is difficult to accurately estimate the time for imputation, as it would depend on which method was being used and the knowledge and software experience of the person carrying out the imputation. A range of values for the cost of imputation were considered throughout to act as a sensitivity analysis. These were one day, 2.5 days and five days of a researcher's time costing £420, £1050 and £2100 respectively. An upper limit of £5000 was also included.

10.2.4 Additional data

The previous sections have discussed the cost of obtaining additional data through the use of reminder questionnaires or through imputation. This amount of additional data needs to be quantified for use in the cost-effectiveness calculations. For these purposes, the calculations will be based on the additional data obtained at the main endpoint of interest. Both the reminder process and imputation can be used to obtain data at intervening time points. However, ultimately it is the main endpoint which is of most interest. The amount of data available through reminder questionnaires and through imputation at the main endpoint is shown in Table 10.3.

Table 10.3: Amount of data available under different data collection strategies

	Recruited	Immediate	Reminder data		Imputation data	
	N	Data	Additional (N_R)	Total	Additional (N_I)	Total (N)
REFLUX (1 year)	357	134	182	316	213	347
MAVIS (1 year)	910	730	100	830	180	910
RECORD (2 years)	5292	2511	693	3204	1211	3722
KAT (2 years)	2356	1560	325	1885	712	2272
PRISM (2 years)	1324	740	149	889	565	1305

10.2.5 Quality of additional data

Section 10.1.2 introduced the idea of quality weights for the additional data. Current trial practice assumes that data collected through reminders, are as good as that obtained through the initial mailing. This assumption will be maintained throughout the remainder of this chapter. The implications of this assumption are discussed further at the end. The quality of the imputed data is of much greater concern, as it is known that imputation can never be a substitute for real data. It is merely a tool to aid analysis (Fayers, Machin 2007). The information on the accuracy of different imputation methods presented in chapter seven can provide one way of generating a weighting mechanism which provides quality adjusted data. The bias in the calculated treatment estimate under imputation (compared to that which was observed) can be converted into a measure of quality on a zero to one scale where zero represents the worst situation and one represents perfect quality. Data collected through reminders are assumed to be perfect quality. Thus, they have a quality rating of one.

For example, in REFLUX, the observed difference in EQ5D scores between the two treatment groups was 0.047 units (Table 7.1); while using simple mean imputation it was 0.016. Therefore, simple mean imputation under-estimated treatment difference in this instance. A proxy measure of 'quality' can be calculated as the ratio of these two quantities, such that quality = 0.047/0.016 = 0.34. Hence, the estimate for quality of imputed data via simple mean imputation is 0.34 (where 0 represents worst quality and 1 is perfect quality). If the treatment difference is over-estimated (compared to the observed estimate), the calculation of quality is slightly different. In this case, the difference between the observed estimate and the bias in the imputed estimate is found. A measure of quality is the ratio of this new calculated quantity and the observed treatment estimate. For example, using last value carried forward (LVCF) imputation the estimate of treatment difference in EQ5D scores was found to be 0.068. Therefore, a bias of 0.068-0.047 = 0.021 was calculated for this method. The estimate of quality is then (0.047-0.021)/0.047 = 0.55. This value of 0.55 is greater than the 0.34 found for mean imputation, which reflects the fact that LVCF was less biased than mean imputation. The final

scenario to consider is the case in which the direction of treatment difference is altered under imputation. It can be argued that this leads to the wrong conclusion with regard to treatment difference. This implies that imputation should be assigned a zero weight.

Table 10.4: Estimates of imputation quality

Trial	Imputation Method	Treatment estimate Observed	Under imputation	Quality Estimate of imputed data
REFLUX	BCF	0.047	0.033	0.70
	PMM*	0.047	0.055	0.83
MAVIS	BCF	-0.019	-0.017	0.89
	PMM*	-0.019	-0.02	0.95
RECORD	LVCF	0.015	0.013	0.87
	MCMC	0.015	0.013	0.87
KAT	LVCF	0.013	0.009	0.69
	Regression**	0.013	0.014	0.92
PRISM	Max	0.015	0.019	0.73
	PMM*	0.015	0.012	0.80

PMM - predictive mean match multiple imputation; * covariates in MI imputation model; ** covariates and QoL in MI imputation model

In the calculations that follow for the ICER and net-benefit statistic, two imputation methods (one simple and one multiple) will be compared to the reminder strategy. The choice of these imputation methods arises from those which showed least bias in chapter seven. Table 10.4 details each of these methods and the estimates of imputed data quality weights. The number of additional units of data obtained from imputation (N_I) is then multiplied by this estimate of quality weight (W_I). This gives an adjusted N_I for use in the calculation of the ICER and net-benefit statistics. For example, if 200 additional units of data were obtained through imputation and the estimate of quality was 0.5, this would equate to the number of quality adjusted additional data units obtained through imputation (200*0.5=100 units). It is this value of N_I which is then used in the calculations of the ICER and net-benefit statistic. In addition to this, a sensitivity analysis will be performed for the value of W_I, in order to see how the decision may change if the quality of imputation changes.

10.3 Calculation of the incremental cost effectiveness ratio (ICER)

The concept of the incremental cost-effectiveness ratio (ICER) was introduced in section 10.1.3. The numbers of additional pieces of data under reminders and imputation presented in Table 10.3 are utilised. It has previously been discussed that data obtained through imputation cannot be regarded as the same as actual observed data. The idea of a quality weighting for imputation to account for this was introduced in section 10.1.2. Data obtained by reminder are assumed to be of equal quality to that obtained immediately. Therefore, for the purposes of calculation $W_R = 1$. The numbers of additional units of data available after imputation does to some extent depend on the method of imputation and what data are required to carry out the chosen method. However, for the purposes of what follows, it will be assumed that it is possible to obtain an imputed value for each piece of missing data. For each of the five trials, the ICER was calculated in order to compare the reminder system with imputation for a range of values for the cost of imputation (C_I) and the different values of the quality weight for imputation. The cost of the reminder process was fixed (C_R), as was the quality of the reminder data (W_R=1). These calculations are now presented for each of the five trials in turn. The formulae for the ICER were given in section 10.1.3. ICER values are presented to the nearest whole pound sterling (£).

10.3.1 REFLUX

In REFLUX, 134 (38%) responses were received through the initial mailing, with an additional 183 returned after reminder. As shown in Table 10.2 the cost of the reminder process was £828.54. Using imputation data could be available for all 357 participants. This involved an additional 223 units of data to that which was obtained immediately and an additional 40 units compared to that which was obtained though reminders.

Given the situation where imputed data were of equal quality to that obtained by reminder, then imputation would dominate, if $C_I < C_R$. This was due to the fact that, more data were obtained at less cost. Where imputation costs more than the

reminder process, a judgement would be needed as to whether the additional data obtained from imputation compared with reminders was worth the additional cost. This decision should also take into account the reduced quality of the data.

Table 10.5: REFLUX - ICER for different imputation costs and quality weight (C_R = 828.54, W_R=1)

ICER	Cost of imputation (C_I) in £			
W_I	420	1050	2100	5000
1	Imputation dominates	5	31	102
0.9		12	68	223
0.83#		72	412	1350
0.8	113	Reminder strategy dominates		
0.7*	16			
0.6	8			
0.5	6			
0.01	2			

*estimate from simple imputation; # estimate from multiple imputation

In this trial, the reminder data was assumed to be of the highest quality (W_R = 1). The cost of this process was £828.54. The calculated ICER for different costs of imputation (C_I) and quality weight for imputation (W_I) are shown in Table 10.5. When W_I = 1 and $C_I < C_R$, imputation dominates as it is just as effective and less costly. When W_I = 0.9 and $C_I < C_R$, imputation dominates, as despite the fact that $W_I < W_R$, imputation has the possibility to provide more data.

In the situation that W_I = 0.9 and $C_I > C_R$, a cost is then attached to obtaining more, better quality data (through reminders). For example, if W_I = 0.9, W_R=1, C_R=828.54 and C_I = £2100, then the cost per additional unit of high quality data is £68 (Table 10.5). As the imputation quality decreases, with cost of imputation remaining fixed, the additional cost attached to obtaining perfect quality information is reduced. In REFLUX, the reminder system was the preferred strategy, when $W_I < 0.816$ and the cost of imputation was greater than the reminders ($C_I > C_R$).

10.3.2 MAVIS

There were 730 of 910 (80%) initial responses at 12 months in MAVIS, with an additional 100 after reminder. The cost of obtaining these reminder questionnaires was C_R = £501.84. Imputation had the ability to provide values for the additional 180 participants to make it a complete sample. Table 10.6 shows the calculated ICER for a range of values of W_I and C_I. When $W_I \geq 0.6$ and $C_I \leq C_R$, imputation provided more data of highest quality at less cost, such that imputation was dominant.

Only when W_I fell below the threshold weight ($W_I = 0.556$) and imputation cost £420 was there a trade-off between costs and quantity/quality of data. At this point, there is a cost attached to obtaining each additional unit of top quality data. For example, if imputation costs £420 but the quality weight is $W_I = 0.5$ and if reminders are judged to be worthwhile, then this implies we are willing to pay £8 for an additional unit of top quality data. Where the cost of imputation was greater than the cost of the reminder process and the quality of imputation was less than or equal to 0.5, the reminder strategy dominated imputation.

Table 10.6: MAVIS - ICER for different imputation costs and quality weight (C_R = £501.84, W_R=1)

ICER	Cost of imputation (C_I) in £			
W_I	420	1050	2100	5000
1		7	20	56
0.95#		8	23	63
0.89*	Imputation	9	27	75
0.80	Dominates	12	36	102
0.70		21	61	173
0.60		69	200	563
0.50	8	Reminders dominate		
0.01	1			

*estimate from simple imputation; #estimate from multiple imputation

Using simple imputation provided an estimate of quality equal to 0.89. Making this assumption concluded that imputation dominated, until the cost of imputation rose above that of the reminder strategy. Similarly, under multiple imputation, $W_I = 0.95$ and $C_I \geq$ £1050, imputation did not dominate. There was a cost attached to obtaining the top quality data.

10.3.3 RECORD

RECORD was the largest of the example trials and had a poor initial response rate. This meant that the cost of the reminder process (C_R= £5467.20) was much larger than in any of the other trials. The reminder process generated an additional 693 pieces of data to the 2511 responses which were obtained immediately. Imputation had the potential to provide responses for all 5292 participants who were recruited to the trial. Imputation was by far the less costly of the two strategies, but the quality of the data was in doubt. The reminder strategy was dominated by imputation, until the threshold weight of $\hat{W}_I = 0.249$ (Table 10.7).

Table 10.7: RECORD - ICER for different imputation costs and quality weight (C_R = £5467.20, W_R=1)

ICER	Cost of imputation (C_I) in £			
W_I	420	1050	2100	5000
1.00				
0.90				
0.87*#				
0.80		Imputation dominates		
0.70				
0.60				
0.50				
0.01	8	7	5	1

*estimate from simple imputation; #estimate from multiple imputation

The cost of imputation was less and despite its poorer quality provided more data than the reminder process. Therefore should be adopted, if the willingness to pay for top quality data is that shown in Table 10.7. For example, if the C_I = 420 and W_I = 0.01, the cost per additional unit of high quality data is £8. Under imputation (both simple and multiple), the estimate of quality was W_I=0.87. In this situation, imputation dominated for range of C_I considered. The domination of imputation in this trial is down to the fact that because of the large sample size and poor initial response rate, the cost of the reminder process was significantly larger than the estimated cost of imputation.

10.3.4 KAT

As in the RECORD trial, the cost of the reminder process in KAT was substantial (C_R = £3488.40). In the KAT trial, at two years, 1560 pieces of data were obtained through the initial wave of questionnaires. An additional 325 responses were acquired through reminders, producing a total of 1885 compared to the 2356 (additional 746) pieces of data obtained using imputation.

Table 10.8 shows that the reminder strategy is more costly and less effective than imputation except when imputation costs £5000. The quality of imputation was estimated as 0.69 by simple imputation and 0.92 under multiple imputation. Imputation was found to dominate the reminder strategy in both cases, as long as $C_I \neq$ £5000. Only when the quality of imputation falls below a quality weight of 0.408 did imputation cease to dominate. In this situation, there will be a trade-off between cost and quality. For example, if C_I = 420 and W_I = 0.01 then the cost per addition piece of top quality data would be £10. The reminder strategy dominated when the W_I = 0.01 and C_I = £5000. As was described for the RECORD trial, it is unlikely that this low level of quality of imputation would ever be accepted.

Table 10.8: KAT-ICER for different imputation costs and quality weight (C_R = £3488.40, W_R=1)

ICER	**Cost of imputation (C_I) in £**			
W_I	420	1050	2100	5000
1.00	Imputation dominates			3
0.92#				4
0.80				6
0.69*				7
0.60				10
0.50				11
0.01	10	8	4	Reminders dominate

*estimate from simple imputation; # estimate from multiple imputation

10.3.5 PRISM

The reminder strategy cost £1232.16 in PRISM. At two years, there were 740 immediate responses, with an additional 149 responses obtained by reminder (totalling 889). Imputation could potentially provide data for the remaining 584, giving the total of 1324 patients.

Table 10.9: PRISM - ICER for different imputation costs and quality weight (C_R = £1232.16, W_R=1)

ICER	Cost of imputation (C_I) in £			
W_I	420	1050	2100	5000
1.00			2	9
0.90			2	10
0.80#	Imputation		3	11
0.73*	dominates		3	14
0.60			4	19
0.50			6	26
0.01	6	1	Reminders dominate	

*estimate from simple imputation; # estimate from multiple imputation

Table 10.9 shows the results of the ICER calculation for the PRISM trial. Imputation dominated when quality weight was above 0.255 and the cost of imputation was less than £2100. When the quality of imputation drops below this a decision is required as to whether the additional cost of reminders is worth the additional quality. This decision can be informed by estimating the incremental costs per unit of top quality data. For example, when C_I = £420 and W_I = 1, reminders would be considered worthwhile, if the incremental cost per unit of top quality data was £6 or greater. The estimate of quality under imputation was 0.73 for simple imputation and 0.80 for multiple imputation. In both circumstances, imputation dominated when imputation cost less than the reminder process. The reminder strategy dominated when imputation was of poor quality and the cost of imputation was greater than that of the reminder process.

10.3.6 Threshold weight

The primary assumption in any of these trials is that reminder data is as good as that obtained through immediate responses (W_R = 1). The threshold weight of imputation $\hat{W}_I$, is the point at which if the quality weight falls below, this value for a given W_R and imputation no longer dominates.

Table 10.10: Threshold weight for imputation ($W_R = 1$)

Dataset	Threshold weight ($\hat{W}_I$)
REFLUX (12m)	0.816
MAVIS (12m)	0.556
RECORD (2 years)	0.249
KAT (2 years)	0.408
PRISM (2 years)	0.255

For each of the trial datasets, these threshold weights (assuming $W_R = 1$) are shown in Table 10.10. These are based on the number of additional units of data obtained with reminder and under imputation using the formula $\hat{W}_I = \frac{N_R W_R}{N_I}$.

The less disparity there is between the proportion of extra data obtained through reminders and that obtained through imputation, then the higher the threshold weight value for imputation. This is also linked to the proportion of immediate responses.

10.4 Calculation of the net-benefit statistic

Section 10.1.5 described the concept of the net-benefit statistic. This cost-effectiveness decision rule uses the value λ that one is willing to pay for the extra information. The net benefit statistic can be calculated for both the reminder process and that of imputation. Whichever is larger is the most cost-effective strategy. The idea of willingness to pay was introduced in section 10.1.4. In this context, this is the amount the researcher would be willing to pay for the piece of additional information. A range of values for the WTP (λ) were used. The estimate of imputation quality was found in section 10.2.5. The costs for each strategy and the additional data collected were shown earlier in section 10.2. A negative net-benefit statistic implies that there was a net-benefit of imputation over and above the reminder strategy. A positive value indicates there was a net benefit of using reminders over and above imputation. The estimate of net-benefit is given to the nearest whole pound.

The net-benefit statistic was calculated for a range of costs of imputation, a range of values for the WTP and three estimates of quality weight for imputation. These are, where imputation is assumed to be of perfect quality (W_I=1), the estimate obtained from the bias of simple imputation and lastly, that for multiple imputation. The number of additional units of data obtained through imputation (N_I) was then multiplied by the weight (W_I) in order to obtain an adjusted number of additional units of data. This adjusted N_I was then used in the calculation of the net-benefit statistic.

10.4.1 REFLUX

The net-benefit statistics are presented in Table 10.11. The quality weight from simple imputation was $W_I = 0.70$. From multiple imputation this was $W_I = 0.83$.

Table 10.11: REFLUX – Net benefit statistic

	Cost of Imputation (£)			
WTP (£)	420	1050	2100	5000
Net benefit statistic when W_I = 1				
5	-613	17	1,067	3,967
10	-818	-188	862	3,762
50	-2,458	-1,828	-778	2,122
100	-4,508	-3,878	-2828	72
500	-20,908	-20,278	-19,228	-16,328
Net benefit statistic when W_I = 0.83				
5	-424	206	1,256	4,156
10	-439	191	1,241	4,141
50	-563	67	1,117	4,017
100	-717	-87	963	3,863
500	-1,953	-1,323	-273	2,627
Net benefit statistic when W_I = 0.70				
5	-279	351	1,401	4,301
10	-149	481	1,531	4,431
50	887	1,517	2,567	5,467
100	2,182	2,812	3,862	6,762
500	12,542	13,172	14,222	17,122

NB: Shading implies reminder strategy has a greater net-benefit than imputation

The interpretation is such that when $W_I = 1$ and C_I = £420 the net benefit of imputation over and above reminders is £613, if WTP for additional information equals £5. When imputation costs £5000, the net benefit of using reminders over and above imputation is £3967, if WTP=£5. If the net-benefit statistic is negative, then imputation is more efficient than reminders (no shading in Table 10.11).

When the net benefit statistic is positive, then reminders are more efficient than imputation (grey shading). In the situation that imputation is of equal quality to the reminder strategy ($W_R=W_I=1$), imputation has a net-benefit over reminders irrespective of the WTP, if the cost of imputation = £420 (i.e. less than the reminder strategy). Once the cost of imputation rises above the cost of the reminder strategy, the WTP for a unit of top quality data comes into play into the decision over which data collection strategy is of most benefit. If WTP=£5, then reminders have a net-benefit over imputation.

A reduction in quality of imputation had a greater impact on which strategy was of most benefit. Under simple imputation using BCF, the estimate of quality was 0.70. Under multiple imputation using a predictive mean match model, $W_I = 0.83$. No matter what the WTP was, if $C_I \geq$ £1050, then the reminder strategy was most efficient. Reminders were of most benefit for WTP≥£50, irrespective of the cost of imputation.

Increasing the quality of imputation through the use of a multiple imputation procedure, meant imputation was of greater benefit than reminders, in a larger number of circumstances. The net-benefit of imputation over the reminder strategy occurred for all values of the WTP, if the cost of imputation was equal to £420. When imputation cost £2100 and WTP = £500, imputation still provided a greater net-benefit. However, if imputation cost £5000, the reminder strategy was more efficient than imputation, for any value of WTP for an additional unit of top quality data.

This analysis showed that the cost of imputation, quality of imputation and WTP for a unit of top-quality data all play a role in determining which data collection strategy is most efficient. For a definitive answer on which strategy is best, accurate values for each of these quantities would need to be obtained. However, the analysis above has shown that how the decision on which data collection strategy is best, is affected by the value of these parameters.

10.4.2 MAVIS

The cost of the reminder process in MAVIS was £432.48. This was less than the cost of imputation except for the lower value considered (C_I = £420). Section 10.2.5 showed that baseline carried forwards and a predictive mean match multiple imputation model were the best imputation strategies for the MAVIS data. The imputation quality weights for these two methods were 0.89 and 0.95 respectively.

Table 10.12: MAVIS – Net benefit statistic

	Cost of Imputation (£)			
WTP (£)	420	1050	2100	5000
Net benefit statistic when W_I = 1				
5	-482	148	1,198	4,098
10	-882	-252	798	3,698
50	-4,082	-3,452	-2,402	498
100	-8,082	-7,452	-6,402	-3,502
500	-40,082	-39,452	-38,402	-35,502
Net benefit statistic when W_I = 0.95				
5	-437	193	1,243	4,143
10	-792	-162	888	3,788
50	-3,632	-3,002	-1,952	948
100	-7,182	-6,552	-5,502	-2,602
500	-35,582	-34,952	-33,902	-31,002
Net benefit statistic when W_I = 0.89				
5	-383	247	1,297	4,197
10	-684	-54	996	3,896
50	-3,092	-2,462	-1,412	1,488
100	-6,102	-5,472	-4,422	-1522
500	-30,182	-29,552	-28,502	-25,602

NB: Shading implies reminder strategy has a greater net-benefit than imputation

Table 10.12 shows the net-benefit statistic for each of these two scenarios along with that where imputation was considered of perfect quality (W_I=1). When imputation cost £420, it had a greater net-benefit over reminders for each value of WTP and W_I. Reminders had a net-benefit over imputation in some situations when $C_I \geq$ £1050. For example, if imputation cost £2100, WTP for additional information was £10 and imputation quality was W_I = 1, then the net-benefit of reminders over and above imputation was £798.

10.4.3 RECORD

The reminder process for RECORD was estimated to cost £5467.20. This was greater than the previous two trials (REFLUX and MAVIS) and was also larger than the upper value placed on the cost of imputation. The reason for this was that the sample size in RECORD was over 5000 patients compared to 357 in REFLUX and 910 in MAVIS. It is, therefore, not surprising that Table 10.13 shows that imputation has an extra net-benefit over the reminder strategy for each value of WTP and when the cost of imputation C_I was less than or equal to £5000.

Table 10.13: RECORD – Net benefit statistic

	Cost of Imputation (£)			
WTP (£)	420	1050	2100	5000
Net benefit statistic when $W_I = 1$				
5	-15,487	-14,857	-13,807	-10,907
10	-25,927	-25,297	-24,247	-21,347
50	-109,447	-108,817	-107,767	-104,867
100	-213,847	-213,217	-212,167	-209,267
500	-1,049,047	-1,048,417	-1,047,367	-1,044,467
Net benefit statistic when $W_I = 0.87$				
5	-13,680	-13,050	-12,000	-9,100
10	-22,312	-21,682	-20,632	-17,732
50	-91,371	-90,741	-89,691	-86,791
100	-177,694	-177,064	-176,014	-173,114
500	-868,282	-867,652	-866,602	-863,702

Both simple (LVCF) and MCMC multiple imputation procedures were found to have a quality weight of $W_I = 0.87$. Reducing the quality of imputation did not alter the conclusion that imputation provided extra net-benefit compared with the reminder strategy. For example, if imputation were to cost £2100, WTP = £5 and $W_I = 1$, then the net-benefit of imputation over the reminder strategy was £13,807. When the quality of imputation reduced to $W_I = 0.87$, then the net-benefit of imputation over reminders reduced to £12,000.

10.4.4 KAT

Administering reminders in the KAT trial cost £3488.40. Chapter seven showed that the best simple imputation method for the KAT data was LVCF. The quality weight for this method was calculated to be 0.69 (section 10.2.5). A regression

method for monotone missingness following MCMC imputation was the best multiple imputation method. This resulted in a quality weight of 0.92.

Table 10.14: KAT – Net benefit statistic

	Cost of Imputation			
WTP (£)	420	1050	2100	5000
Net benefit statistic when W_I = 1				
5	-5,423	-4,793	-3,743	-843
10	-7,778	-7,148	-6,098	-3,198
50	-26,618	-25,988	-24,938	-22,038
100	-50,168	-49,538	-48,488	-45,588
500	-238,568	-237,938	-23,6888	-233,988
Net benefit statistic when W_I = 0.92				
5	-5,105	-4,475	-3,425	-525
10	-7,142	-6,512	-5,462	-2,562
50	-23,434	-22,804	-21,754	-18,854
100	-43,800	-43,170	-42,120	-39,220
500	-206,728	-206,098	-205,048	-202,148
Net benefit statistic when W_I = 0.69				
5	-4,190	-3,560	-2,510	390
10	-5,311	-4,681	-3,631	-731
50	-14,280	-13,650	-12,600	-9,700
100	-25,492	-24,862	-23,812	-20,912
500	-115,188	-114,558	-113,508	-110,608

NB: Shading implies reminder strategy has a greater net-benefit than imputation

Table 10.14 displays the net-benefit statistic calculated for each of these two values of W_I and $W_I = 1$. Since NB > 0 in all but one case, it can be concluded that imputation provided an extra net-benefit compared with reminders. For example, if $W_I = 0.92$, C_I = £2,100 and WTP = £10, then imputation provided an extra net benefit of £5,462 over the reminder strategy. The only scenario where the reminder strategy was most efficient was when imputation cost £5000, the quality was $W_I = 0.69$ and WTP = £5.

10.4.5 PRISM

The PRISM trial reminder process cost £1232.16. Chapter seven showed that the maximum value imputation method was the best simple imputation procedure. A predictive mean match model was the best multiple imputation procedure. The bias in the calculated treatment difference was used to calculate quality weights of 0.73 and 0.80 respectively (section 10.2.5).

Table 10.15: PRISM – Net benefit statistic

	Cost of Imputation (£)			
WTP (£)	420	1050	2100	5000
Net benefit statistic when $W_I = 1$				
5	-2,987	-2,357	-1,307	1,593
10	-5,162	-4,532	-3,482	-582
50	-22,562	-21,932	-20,882	-17,982
100	-44,312	-43,682	-42,632	-39,732
500	-218,312	-217,682	-216,632	-213,732
Net benefit statistic when $W_I = 0.80$				
5	-2,403	-1,773	-723	2,177
10	-3,994	-3,364	-2,314	586
50	-16,722	-16,092	-15,042	-12,142
100	-32,632	-32,002	-30,952	-28,052
500	-159,912	-159,282	-158,232	-155,332
Net benefit statistic when $W_I = 0.73$				
5	-2,199	-1,569	-519	2,381
10	-3,585	-2,955	-1,905	995
50	-14,678	-14,048	-12,998	-10,098
100	-28,544	-27,914	-26,864	-23,964
500	-139,472	-138,842	-137,792	-134,892

NB: Shading implies reminder strategy has a greater net-benefit than imputation

Table 10.15 shows the net-benefit statistic calculated for these two values of W_I and for $W_I = 1$. As was shown with RECORD and KAT, in most situations, the net benefit was greater than zero for the values of WTP and C_I considered. This implies that imputation provided an extra net-benefit over reminders. For example, if C_I = £2100, W_I = 0.73 and WTP = £5, then the net-benefit of imputation over reminders was nearly £519. When C_I = £5000, the reminder strategy was shown to be more efficient than imputation in several cases for WTP ≤ £10. For example, if C_I = £5000, W_I = 1 and WTP = £5, then the net-benefit of reminders over imputation was £1,593.

10.4.6 Summary

A decision on whether imputation or the reminder strategy is the most cost-effective method of data collection depends on a number of factors. These are the cost of the reminder process, cost of imputation, quality of imputation and the WTP for a unit of perfect information. The process above has considered various options for these in the context of the five case-studies.

Using the net-benefit approach has shown that in three of the five trials imputation provided greater net-benefit than the reminder strategy for most combinations of the parameters (cost and quality of imputation, WTP). The quality of imputation took a value between 0.69 and 0.95, depending on the method of imputation and trial involved. This analysis estimated this value of imputation quality (W_I) using the estimates of bias from chapter seven.

The upper bound on the cost of imputation was estimated to be £5000 which was greater than the cost of administering reminders in four of the trials. The trial for which the net-benefit statistic of reminders was not always less than imputation was REFLUX. The cost of the reminder process in REFLUX was £828.24, which compared to the potential cost of imputation was cheaper. This is due to the smaller recruited sample size. However, for the other trials, this was not the case and imputation provided a net-benefit over reminders in most cases. The work conducted in this chapter used estimated values of WTP, cost of imputation and quality of imputation. Further research is needed to elicit more accurate values for each of these quantities. This is discussed further in section 10.5.

10.5 Discussion

This chapter aimed to determine whether the reminder strategy or the use of imputation was a cost-effective method of additional data collection. The calculation of the ICER in section 10.3 showed that imputation dominated the reminder strategy in two of the five trials. These were the largest trials in terms of sample size, which also meant the cost of the reminder process was greater. In the smallest trial (REFLUX, N=357), imputation only dominated if the quality of imputation was greater than 0.82 and the cost of imputation was £420. The reminder strategy dominated when $W_I < 0.82$ and the cost of imputation was greater than £1050. This was because the potentially poorer quality data (through imputation), was being obtained at a higher cost.

Section 10.4 considered the net-benefit statistic as a way of determining which data collection strategy was the most cost-effective. This approach used a decision rule that involved the amount (λ) that you are willing to pay for extra information. It was shown that in REFLUX and MAVIS, there was a net-benefit of reminders over and above imputation under certain combinations of WTP, C_I and W_I. As imputation quality reduced and cost of imputation (or WTP) increased, then the reminder strategy provided a net-benefit over imputation. In the three remaining trials for the values of these quantities considered, imputation had a net-benefit over the reminder system in nearly all cases. This is mainly due to the fact that these trials were larger (>1000 participants). Thus, the cost of the reminder strategy was increased. The reduction in quality of imputed data was not enough to counteract the relative difference in cost of the two data collection strategies. This method backed up the findings from the ICER calculations, but used a different decision rule.

Throughout all the work presented in this chapter there is a degree of uncertainty over the true value of W_I, C_I and WTP. Calculating the value of the ICER and net-benefit statistic for a range of these values has highlighted the different decisions that may be made about which data collection strategy was most cost-effective. Further research is needed about the most appropriate values for these quantities. A number of other economic approaches might be considered helpful. These are described in brief here.

Time trade-off is a method of deriving utilities values that can be used in cost-utility analysis in the context of economic evaluations of healthcare alternatives (Drummond et al. 1997). They assess the value of health states by asking a hypothetical question of willingness to trade that health state and its duration for the state of perfect health for a lesser number of years. The subject is given a choice between an illness for a specific period or the alternative of perfect health for a shorter life-span. Perfect heath has a value of one, while the state of illness has a value between 0 and < 1. The time period in the second alternative (perfect health) is altered during the survey until the subject judges both alternatives the

same. Therefore, the more debilitating the health states being valued, the more years of life, one would be willing to give up to attain perfect health for those years. This approach can be adapted to suit the purposes of determining the weight of imputation and how low researchers are willing to allow the quality of data to drop in order to obtain more data. Rather than being a time-trade off approach it could be framed as a quality trade off.

An alternative to the quality-trade off approach might be the discrete choice experiment (DCE). A DCE is a technique for eliciting preferences. Ryan *et al.* describe the rationale behind a DCE as:

> 'any good or service can be described by its characteristics (or attributes) and, the extent to which an individual values a good service depends upon the nature and levels of these characteristics.' (Ryan et al. 2001)

The technique involves presenting participants with a choice of scenarios based on the characteristics and levels. For each choice, they are asked to choose their preferred scenario. The responses are then modelled using a benefit (or satisfaction) function which provides information on whether or not the characteristics are important, the relative importance of characteristics, the rate at which participants are willing to trade between characteristics and overall benefit scores for alternative scenarios. In the context of this work, scenarios could be provided which allow for different amounts of data collected through the first wave of mailing, the proportion of data collected through reminder and the proportion of imputed data. This could then be used to obtain estimates of WTP for the top quality data. Participants of the DCE would need to have sufficient knowledge of the clinical trial setting, understand the principle of imputation and be able to consider the concept of bias.

One assumption that has not been discussed so far is that data collected by reminder were assumed to be of equal quality to that which was obtained from immediate responders. It can be argued that this is not the case. QoL questionnaires are usually administered at a clinical point of interest (e.g. one

month post surgery). If a responder does not fill this questionnaire in until four weeks after it was sent out, then the time frame is shifted. For example, for questions such as 'how is your health state today?', if you were to respond after the second reminder (around four weeks), the QoL being reported is for two months after surgery and not one. Current trial practice is to treat responses collected after a reminder as equivalent to that which was obtained from the initial questionnaire mailing. The current practice would only be a problem if the proportion of reminder-responders was different across treatment groups. In these trials, this was not found to be the case. Therefore, the assumption of reminder data being as good quality as immediate responses was valid.

10.6 Implications for research practice

The work carried out in this chapter indicates that for trials of smaller sample sizes, where the anticipated cost of the reminder system is less than £5000, the reminder strategy is likely to be cost-effective. Once the reminder cost increases beyond this then imputation could have a role to play. However, a decision on whether it should be used is dependent on the perceived quality of the data. The question could be answered with further research as discussed above. On a trial by trial basis the quality of imputation could be determined using the process described in chapter seven. This provides an estimate of bias which can be translated to an estimate of quality.

The amount of missing data within a trial will also have an impact on the use of reminders. In a large trial with a large amount of missing data, the cost of the reminder process will be significant and likely to be greater than the cost of imputation. In this case, there then needs to be a trade off between this additional cost and the improved quality the reminder responses will provide, over those obtained through imputation.

The recommendation from this work is that in smaller trials, where less than 1000 questionnaires are to be issued, reminders should be used. They provide much

more reliable data and the results of subsequent analysis will be more credible than if imputation has been used. In larger trials, the reminder process is still extremely important. However, it has an obvious impact on the resources. Therefore, sufficient funds should be sought ahead of time to allow for the reminder process.

Chapter 11 Conclusions and Recommendations

11.1 Introduction

The randomised controlled trial (RCT) is an important way of evaluating healthcare interventions, forming the basis of evidence-based medicine. The publication of a wrong conclusion from a trial for a particular therapy could have disastrous consequences. Therefore, it is essential that the results of RCTs can be relied upon. A significant amount of missing data within a trial dataset has the potential to affect trial conclusions. Ensuring that a dataset is as complete as possible is a continuing battle that researchers face. It also entails significant resources. The trial outcomes considered in this work were those of quality of life (QoL), which are perhaps more susceptible to the problems of missing data. A missing QoL assessment is likely to be informative. Thus, this must be considered in any statistical analysis. The aim of this thesis was to investigate the different strategies available to deal with QoL missing data, with particular reference to the role of the reminder-responses.

Throughout this work seven example trial datasets were utilised. Each of these collected repeated assessments of one or more QoL instruments. The follow-up questionnaire process involved a reminder system, whereby if two weeks after mailing the participant had not responded, then a reminder questionnaire was sent. In some cases, a second reminder was sent a further two weeks later. This process generated an additional proportion of data within each trial, which would have otherwise been missing. It is this data which forms the basis of this thesis. To analyse the QoL outcomes, the trial researchers had carried out a complete-case analysis on the assessment of interest, adjusting for baseline QoL and other patient characteristics. Clearly, this ignored any patients for whom the assessment of interest or the baseline assessment was missing. This approach has the potential to bias the estimate of treatment difference. In this thesis, alternative model-based strategies have been considered, along with investigating the use of imputation.

The investigation into the missing data mechanism informed which of the imputation or model-based methods would be most appropriate.

The current literature on missing data is extensive, but it relies heavily on simulated data or data removed in a certain way from a complete dataset. The work presented in this thesis used a novel approach to investigate the ways of dealing with missing data. It has the advantage that actual real 'known' data were utilised. The short review carried out in chapter two highlighted that in practice, clinical trial researchers do not always take account of missing data correctly. Often the missing data mechanism is ignored. Methods of imputation are used without any discussion of the assumptions they have made. By utilising the data collected by reminder, the mechanism of missingness was investigated and appropriate methods of imputation (if any) identified.

11.2 Discussion of findings

11.2.1 Investigating the mechanism of missing data

Using a number of methods, the missing data mechanism inherent in each trial dataset was investigated. Two hypothesis tests were used, with the results of Little's test being the most reliable (Little 1988). This was due to the fact that the missing data pattern was a combination of intermittent and monotone missingness. The Listing and Schlittgen test is only applicable with a monotone missing data pattern (Listing, Schlittgen 1998). Two logistic regression procedures were used to identify the missing data mechanism at a particular assessment. These differed with respect to the binary outcome of dropout. Ridout's procedure models dropout after an observed assessment (Ridout 1991). Fairclough's method uses the indicator of response or not, at a particular assessment (Fairclough 2002).

Different data scenarios were considered when investigating the missingness mechanism. Scenario one investigated the mechanism of non-response. The response data included only the data collected immediately (and not through reminders, as these data were missing). The observed responses in scenario two included the data collected by reminder. Scenario two represents the situation of

most clinical trials, and the platform on which researchers would base their investigation. On the whole, in the trial datasets presented here, missingness was found to be MAR (associated with covariates and observed QoL). Participants, who had shown lower observed QoL, were more likely to provide missing responses. Ignoring this finding in analysis could potentially bias the results. Patients displaying lower QoL at other assessments are liable to display lower QoL at the missing assessment. If the missing data are ignored, then it is possible that calculated means on the observed sample are inflated and not reflective of the true mean.

The mechanism of missingness identified was not always the same in scenario one and two. This suggested that the reminder data had an important role to play. In a trial which does not employ a reminder system, only the data collected immediately would be available. If the investigation into the missingness mechanism was based only on this data, then one could potentially get a distorted view of the mechanism behind the missing data. Obtaining as much data as possible in any setting is always going to give a more informed decision and ultimately, reduce any potential bias in analysis results.

Scenario three considered the subset of data which only included responders; the reminder data was set to missing. The advantage of this approach was that the current (observed) QoL was known. This allowed an investigation as to whether current scores were different between immediate and reminder responders. In several situations, this was shown to be MNAR, suggesting that the missingness was informative. The use of reminder data allowed this conclusion. In a more usual setting, this would not have been possible, as the data required are missing.

In any study that contains missing data, the missingness mechanism should be identified in advance of any analysis, in order that the most appropriate methods can be identified (Curran et al. 1998a). Chapter two showed that in the majority of reported clinical trials there was no formal discussion about reasons for missingness and no investigation into the mechanism of missing data.

11.2.2 Methods of imputation

Imputation was proposed as one way of dealing with missing data, whereby an alternative value is substituted for one that is missing. The review of published clinical trials found that simple imputation (usually LVCF) was widely used by researchers, but there was no discussion of the rationale for the choice of method or the assumptions it makes. Simple imputation often produces inappropriate standard errors, which lead to inappropriate confidence intervals and p-values. This can clearly have an impact on the trial result. Multiple imputation can to some extent overcome this as it models the uncertainty in the imputed values. A variety of these imputation procedures were carried out on the example datasets. As explained previously, this process utilised the reminder data. It was the data collected by reminder which was imputed. This allowed the accuracy (in terms of bias and precision) of imputation to be assessed. This differed to previous authors' work, as in their studies the missing data was simulated or removed subject to a pre-specified pattern. This resulted in the mechanism to be known and the accuracy of imputation to be predicted in advance.

Simple imputation procedures were not expected to perform that well, as they often make the assumption of MCAR. This was shown not to be likely in the example datasets. This proved to be the case, with multiple imputation out-performing simple imputation methods in most situations. No obvious difference between the two types of imputation was consistent with the situations where there was no evidence against the MCAR assumption. The most appropriate simple imputation method seemed to be LVCF or BCF. This was surprising as current literature recommends against this type of procedure, due to the fact that they make strong assumptions about the stability of QoL (Carpenter, Kenward 2007; Gadbury, Coffey & Allison 2003). Since these methods were found to be reasonable in some of the example datasets, it may be that the populations under study did have stable QoL.

Multiple imputation was anticipated to perform much better than simple imputation, as the methods make the assumption of MAR. This is more plausible

than MCAR in the QoL setting. The reminder responses were found to be MAR and MNAR in some cases, suggesting that MI would be more appropriate. Molenberghs and Kenward discussed the different MI strategies, including the relative merits of each (Molenberghs, Kenward 2007). They recommend that for longitudinal data, a regression or predicted mean match model is appropriate. This tied in with the findings of this thesis, as it was one of these two MI procedures which were usually the most accurate. The main advantage of a predictive mean match model over regression is that the imputed values are always within the range of the data (Fairclough 2002). It is important to obtain a suitable imputation model, which includes variables that are predictive of both missingness and outcome. The model used for imputation should be at least as complex as the model used for analysis (Carpenter, Kenward 2007).

11.2.3 Model-based strategies

An analysis of covariance (ANCOVA) to determine if there was a treatment difference in QoL outcomes was the reported analysis for five of the trials. This method was undertaken as a complete-case analysis, with those patients providing QoL scores at baseline and at the assessment of interest. This approach assumed that the data were MCAR, which has been shown not to be the case in most situations. An alternative is to use a longitudinal analysis method. A repeated measures model was considered along with a number of pattern mixture models. The repeated measures model makes the assumption of MAR, while the pattern mixture model assumed MNAR. As there was evidence against the MCAR assumption, these two mechanisms were shown to be more plausible in the example trials. A pattern mixture model has the advantage that the model for dropout need not be specified. However, this was negated with the need for restrictions to enable estimation of the parameters. The validity of these restrictions cannot be tested and the results may be sensitive to the choice of restriction. One drawback of the pattern mixture models was the need for a restriction to those patients with a monotone missing data pattern. This caused a substantially reduced number of patients to be included in the analysis, compared to the repeated measures model. Although not implemented here, an alternative

which is applicable to all missing data patterns was proposed by Troxel *et al*. (Troxel, Lipsitz & Harrington 1998).

The use of a joint mixed effects model was illustrated with the NPC trial data. In this trial, this type of model-based procedure was found not to be any more useful than the standard mixed effects model, but did act as an illustration of when it might be appropriate. This type of modelling should be considered if the missing data are thought to be MNAR and if there are a large number of dropouts, for example, due to death.

11.2.4 Economic benefit of different data collections strategies

Throughout this thesis, alternative ways of dealing with missing data were investigated. Current practice in CHaRT, is to utilise a reminder system at follow-up, with the aim of recovering data initially missing. This process involves significant resources. This can be a limiting factor when resources are scarce. An alternative proposed was the use of imputation for the missing data, which is potentially cheaper but has the disadvantage of greater bias. Two economic techniques were utilised to investigate the cost-effectiveness of these two data collection strategies. The cost of the reminder process was calculated directly, but the cost of imputation was estimated. The resultant bias in estimates of treatment difference under imputation was used to estimate a quality weight for imputation.

This process showed that depending on the values of the cost of imputation, quality of imputation and willingness to pay for information, different decisions about which strategy was more cost-effective were made. In REFLUX, the reminder strategy seemed to be cost-effective. The cost of reminders was relatively small compared to the other trials and not that different to imputation. In the remaining trials, imputation dominated, despite the reduction in quality. Better estimates of the quality of imputation are needed before this approach could be used to make a definitive argument over which data collection strategy is most cost-effective. The early indications are that if a trial is of 1000 participants or less, a reminder system is likely to be worthwhile.

In larger trials, the amount of missing data would impact on whether or not the reminder strategy and/or imputation are useful. It is well known that in studies with more than 20% of missing data, the credibility is reduced and the use of imputation is regarded as questionable (Schulz, Grimes 2002). They describe the five-and-20 rule of thumb, where fewer than 5% of missing data will probably lead to little bias. However, if this amount is greater than 20% has the potential to pose a serious threat to validity. The reminders would prove particularly beneficial if the initial response rate was poor, but the use of them increases the overall response rate to over 80%. However, the cost of the process may prove to be a limiting factor.

11.3 Impact of different approaches on trial results

Throughout this thesis, a number of different strategies have been presented to deal with the missing data. This section aims to highlight the impact different analysis strategies have on the trial results. The scenarios considered are: ANCOVA on immediate responses only (no reminders); ANCOVA on all observed responses (the published trial analysis); a repeated measures approach on the immediate responses; repeated measures approach on all observed responses; simple imputation following immediate responses; simple imputation following observed responses; multiple imputation following immediate responses; multiple imputation following observed responses. For illustration, the analyses have been undertaken for one QoL measure within each trial. The aim of this section is to illustrate the impact alternative analysis strategies may have on the result. A recommendation on which would be considered the most valid approach, and what the trial result should have been, is made.

11.3.1 REFLUX

The primary trial outcome in the REFLUX trial was the RQLS. Table 11.1 shows the estimate of treatment difference using the different strategies for the RQLS at

12 months. Figure 11.1 presents the estimate of treatment difference and 95% CI in a visual manner. All estimates were positive suggesting that the surgical group had better reflux-specific QoL at 12 months. Each result was statistically significant but the number of patients involved differed, as did the magnitude of the effect. Carrying out the ANCOVA analysis on the immediate responses only used 121 of the 357 (34%) recruited patients. Although statistically significant, the magnitude of treatment difference was less than the estimate that was based on the data that included the reminder-responses.

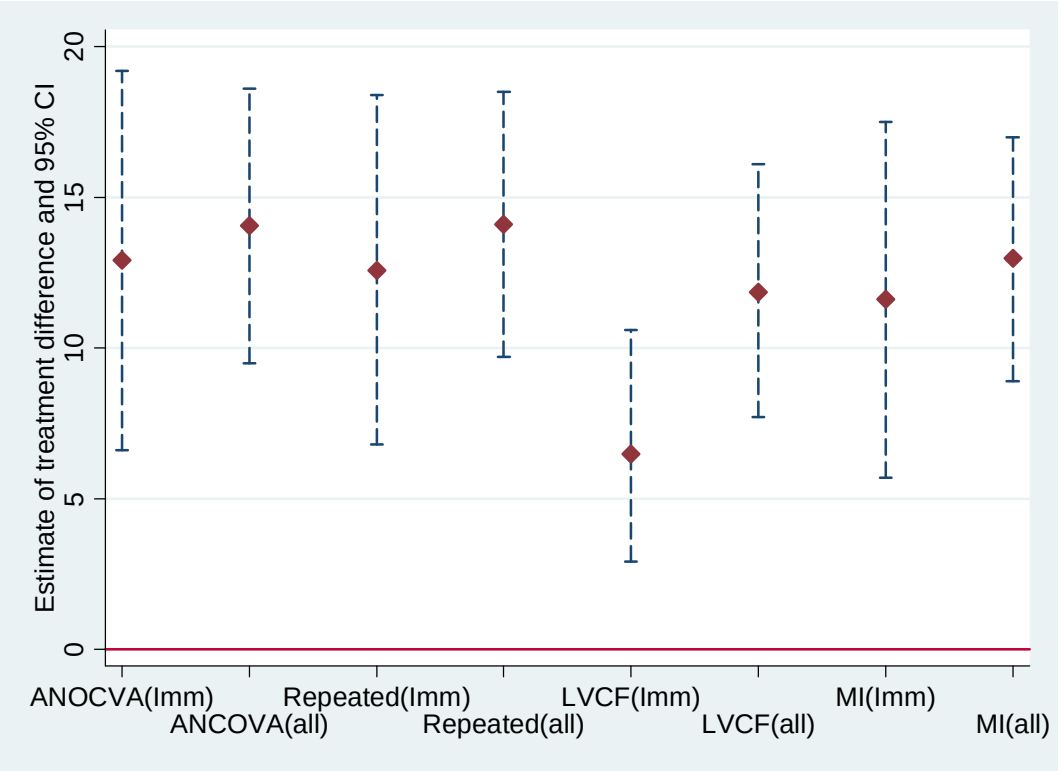
Table 11.1: REFLUX – Estimates of treatment difference in RQLS under different analysis strategies

		Treatment difference at 12m			
Strategy	N	Estimate	SE	95% CI	p-value
Immediate responses (ANCOVA)	121	12.91	3.18	(6.6, 19.2)	<0.001
All observed responses (ANCOVA)	276	14.05	2.09	(9.5, 18.6)	<0.001
Immediate responses (repeated measures)	327	12.58	2.95	(6.8, 18.4)	<0.001
All observed responses (repeated measures)	327	14.11	2.23	(9.7, 18.5)	<0.001
Immediate responses plus LVCF	327	6.48	1.81	(2.9, 10.6)	<0.001
All observed responses plus LVCF	327	11.86	2.11	(7.7, 16.1)	<0.001
Immediate responses plus MI*	342	11.61	2.73	(5.7, 17.5)	0.001
All observed responses plus MI*	353	12.97	2.08	(8.9, 17.0)	<0.001

* Predictive mean match model including covariates and previous QoL

Using a repeated measures approach allowed 327 (92%) patients to be included. In this case, the estimates of treatment difference were similar to the ANCOVA approaches. Carrying out simple imputation using LVCF provided treatment difference estimates of less magnitude than the other methods. The number of patients included was the same as for the repeated measures approach. Multiple imputation included a greater number of patients in the analysis and provided estimates slightly less than the ANCOVA estimates. The different analysis strategies reached the same conclusion, whereby there was a significant difference in the 12 month RQLS. Those in the surgical group displayed better scores. The choice of method did, however, impact on the magnitude of the treatment difference.

Figure 11.1: REFLUX – Estimates of treatment difference (95% CI) in RQLS under different analysis strategies



The missing data in REFLUX was found to be MCAR. The ANCOVA analysis (including reminders) was based on 77% of participants. A sensitivity analysis on this result showed that there was a treatment difference in all cases. Therefore, in REFLUX, the original published analysis strategy is likely to be the most valid. The patients receiving surgery reported significantly better (mean difference =14.1 units with 95% CI = (9.5, 18.6)) 12 month reflux specific QoL scores than those on medical management.

11.3.2 MAVIS

The MAVIS trial contained both the EQ5D and SF12 instruments. The estimate of treatment difference under different analysis strategies for the two SF12 component scores are shown in Table 11.2 and in Figure 11.2 (PCS) and Figure 11.3 (MCS). In both cases, all methods provided p>0.05 and concluded that there was no evidence of a treatment difference at 12 months. Multiple imputation allowed all patients to be involved in analysis, rather than the 823 (90%) in the

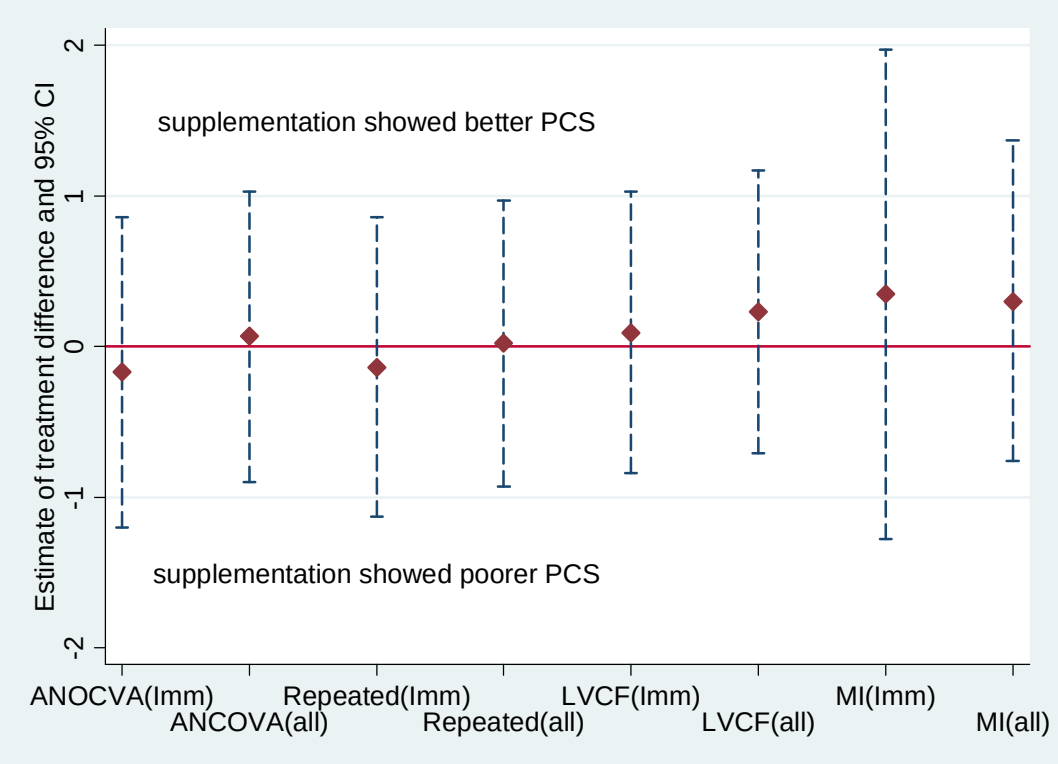
published trial analysis. An ANCOVA based on immediate responses only used 79% of participants.

Table 11.2: MAVIS– Estimates of treatment difference in SF12 component scores for different analysis strategies

		Treatment difference in physical score at 12 months			
Strategy	N	Estimate	SE	95% CI	p-value
SF12 Physical component score (PCS)					
Immediate responses (ANCOVA)	723	-0.17	0.53	(-1.20, 0.86)	0.75
All observed responses (ANCOVA)	823	0.07	0.49	(-0.90, 1.03)	0.89
Immediate responses (repeated measures)	906	-0.14	0.51	(-1.13, 0.86)	0.79
All observed responses (repeated measures)	906	0.02	0.48	(-0.93, 0.97)	0.97
Immediate responses plus mean	906	0.09	0.48	(-0.84, 1.03)	0.85
All observed responses plus mean imputation	906	0.23	0.48	(-0.71, 1.17)	0.63
Immediate responses plus MI*	910	0.35	0.77	(-1.28, 1.97)	0.66
All observed responses plus MI*	910	0.30	0.54	(-0.76, 1.37)	0.58
SF12 Mental component score (MCS)					
Immediate responses (ANCOVA)	723	0.05	0.58	(-1.09, 1.20)	0.93
All observed responses (ANCOVA)	823	-0.03	0.55	(-1.11, 1.05)	0.96
Immediate responses (repeated measures)	906	0.02	0.56	(-1.10, 1.13)	0.98
All observed responses (repeated measures)	906	-0.18	0.54	(-1.24, 0.88)	0.74
Immediate responses plus BCF	906	0.03	0.48	(-0.92, 0.97)	0.96
All observed responses plus BCF	906	-0.06	0.51	(-1.06, 0.94)	0.90
Immediate responses plus MI*	910	0.30	0.55	(-0.79, 1.39)	0.58
All observed responses plus MI*	910	-0.04	0.54	(-1.09, 1.02)	0.95

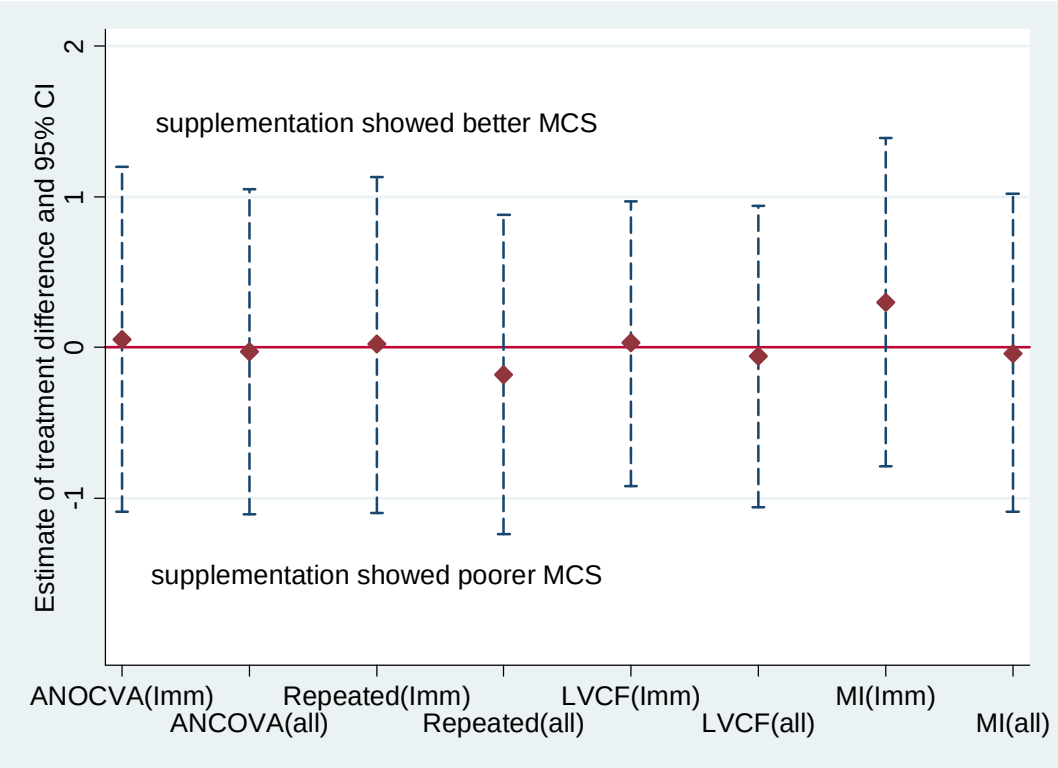
* Predictive mean match model including covariates

Figure 11.2: MAVIS– Estimates of treatment difference (95% CI) in the SF12 physical component scores for different analysis strategies



The estimates of treatment difference did vary in both magnitude and direction for the different analysis strategies. For the PCS, using only the immediate data in both the ANCOVA and repeated measures model provided a negative treatment difference. A positive treatment difference was shown by the other methods. For the MCS, using immediate data provided positive estimates of treatment difference. Whereas, a negative estimate was found for each of the methods carried out on all observed responses.

Figure 11.3: MAVIS– Estimates of treatment difference (95% CI) in the SF12 mental component scores for different analysis strategies



The initial response rate at 12 months for MAVIS was high. This increased further to 90% once reminders had been implemented. Using imputation on the remaining missing data gave consistent conclusions, suggesting that the original ANCOVA on all observed responses (including reminders) was reasonable. The trial concluded that there was no significant effect of multivitamin and multimineral supplementation on the QoL of the participants.

11.3.3 RECORD

RECORD contained two treatment comparisons: calcium versus no calcium and vitamin D versus no vitamin D supplementation. The results for the comparison in two year EQ5D scores are presented here. The published estimate of treatment difference in the EQ5D scores for the calcium comparison was found to be borderline significant (p=0.05). There was no evidence that Vitamin D supplementation improved quality of life (p=0.81). Table 11.3 shows the estimates of treatment difference for each of the analysis strategies. Figures 11.4 (calcium comparison) and 11.5 (vitamin D comparison) show this information graphically. The number of patients involved in the comparison differed, depending on which method was being used. The effect of vitamin D supplementation on QoL was not found to be significantly different by any of the analysis methods considered. The magnitude of the estimate was very similar across the methods, although the direction did change.

Table 11.3: RECORD– Estimates of treatment difference in EQ5D scores for different analysis strategies

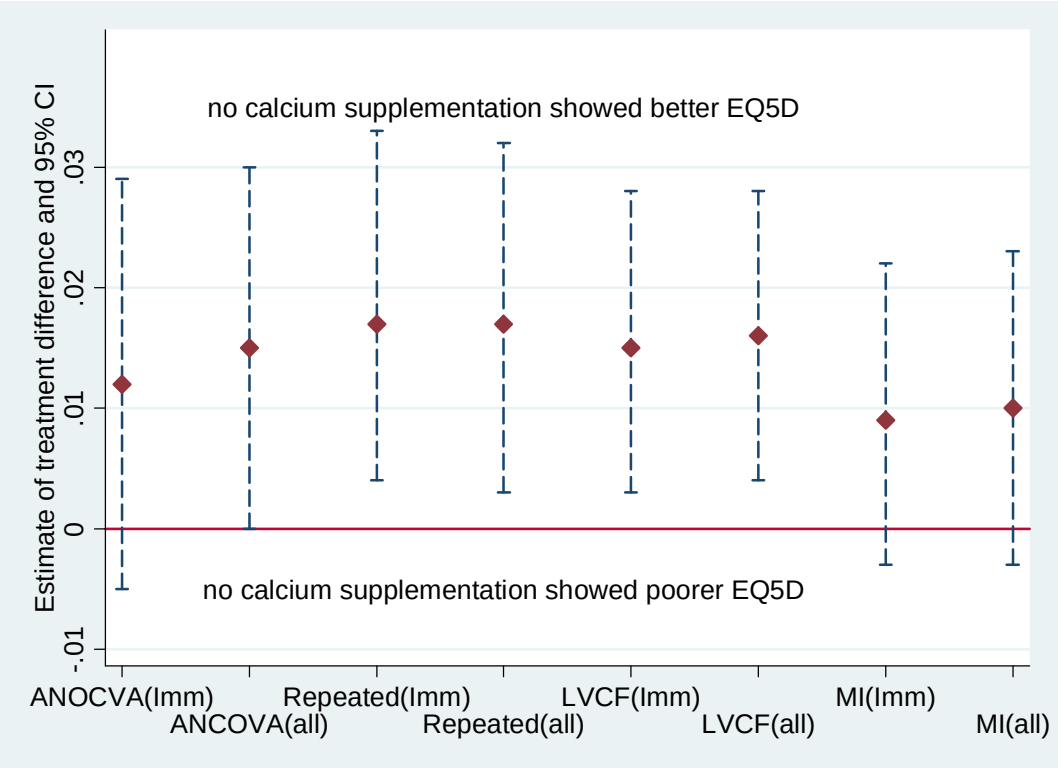
		Treatment difference at 2 years			
Strategy	N	Estimate	SE	95% CI	p-value
Calcium versus no Calcium					
Immediate responses (ANCOVA)	1919	0.012	0.009	(-0.005, 0.029)	0.16
All observed responses (ANCOVA)	2879	0.015	0.008	(0.000, 0.030)	0.05
Immediate responses (repeated measures)	2907	0.017	0.008	(0.004, 0.033)	0.04
All observed responses (repeated measures)	3906	0.017	0.008	(0.003, 0.032)	0.02
Immediate responses plus LVCF	2907	0.015	0.007	(0.003, 0.028)	0.02
All observed responses plus LVCF	3906	0.016	0.006	(0.004, 0.028)	0.01
Immediate responses plus MI*	5291	0.009	0.006	(-0.003, 0.022)	0.15
All observed responses plus MI*	5291	0.010	0.010	(-0.003, 0.023)	0.14
Vitamin D versus no Vitamin D					
Immediate responses (ANCOVA)	1919	0.009	0.009	(-0.008, 0.026)	0.30
All observed responses (ANCOVA)	2879	-0.002	0.008	(-0.017, 0.013)	0.81
Immediate responses (repeated measures)	2907	0.006	0.008	(-0.011, 0.022)	0.49
All observed responses (repeated measures)	3906	0.006	0.008	(-0.009, 0.021)	0.46
Immediate responses plus LVCF	2907	0.003	0.006	(-0.010, 0.015)	0.70
All observed responses plus LVCF	3906	0.003	0.006	(-0.009, 0.015)	0.62
Immediate responses plus MI*	5291	0.002	0.006	(-0.010, 0.022)	0.15
All observed responses plus MI*	5291	-0.0002	0.010	(-0.014, 0.013)	0.97

* MCMC imputation for all the missing data

The conclusion for the calcium comparison was affected by the choice of analysis method. The published analysis suggested borderline evidence of a difference

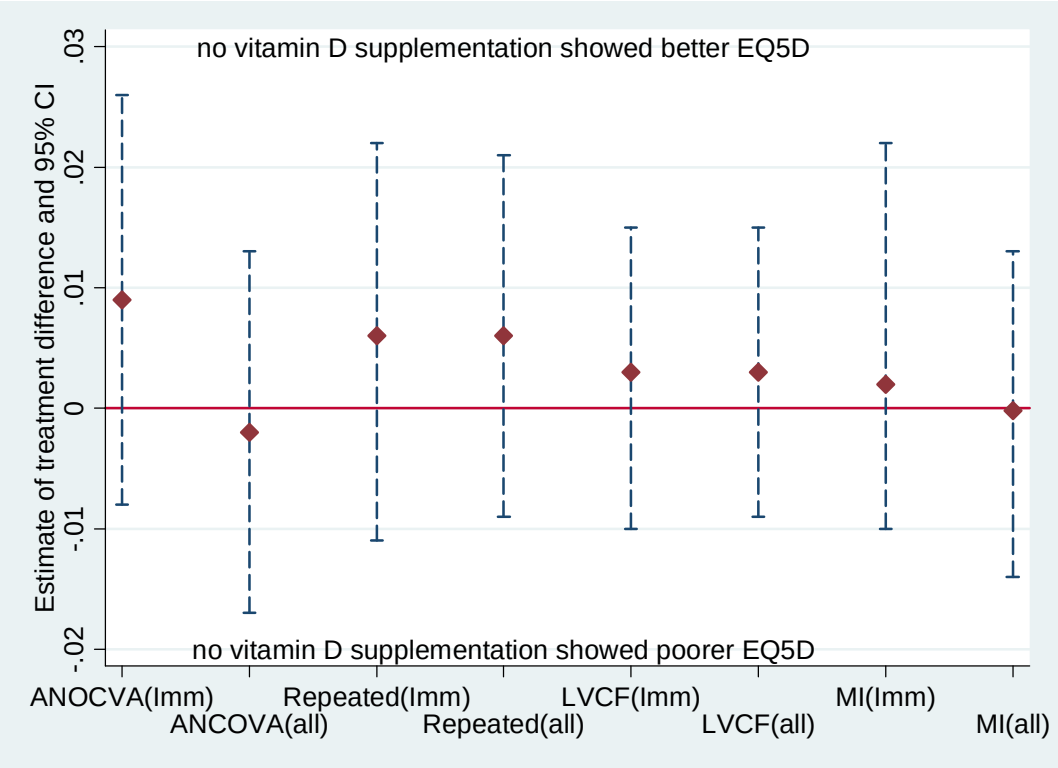
(p=0.05). If this analysis had been based on only immediate responses, no significant difference would have been found (p=0.16). Using a repeated measures or simple imputation approach for both immediate and all responses, found a significant effect of calcium supplementation (p<0.05). This strategy allowed more patients to be included in the analysis, therefore increasing the power to find a difference. Multiple imputation found no effect (p>0.05). This example highlights that a different choice of analysis method could have resulted in a different conclusion.

Figure 11.4: RECORD– Estimates of treatment difference (95% CI) in the EQ5D for the calcium treatment comparison for different analysis strategies



The response rate at the main endpoint (24 months) was poor in RECORD. For the EQ5D outcome this was only 54%. Making conclusions from this should be done so with caution, as only just over half the patients were included. The mechanism of missing data was found to be MAR, suggesting that a repeated measures approach would be more appropriate than the reported ANCOVA. This increased the proportion of included participants to 74%. Although multiple imputation had the potential to include all participants, given the extremely poor response rate, in this situation imputation would be considered unreliable.

Figure 11.5: RECORD– Estimates of treatment difference (95% CI) in the EQ5D for the vitamin D treatment comparison for different analysis strategies



The most appropriate analysis would be the repeated measures approach, which assumed MAR. In this case, the result of the calcium treatment comparison would have been 0.017 (0.003, 0.032) with p = 0.02. Those patients who did not receive calcium supplementation displayed significantly better QoL (measure by EQ5D). For the vitamin D comparison, using the same analysis approach, no significant difference (p=0.46) in QoL was found between treatment groups (mean difference = 0.006 with 95% CI (-0.009, 0.021)).

11.3.4 KAT

The primary outcome in the KAT trial was the Oxford Knee Score (OKS). No significant difference was reported between patella resurfacing and no patella resurfacing (p=0.64). Using the different analysis strategies did not change this conclusion. However, the number of patients involved in the analysis and the magnitude of the difference did vary (Table 11.4 and Figure 11.6). Using only immediate responses in an ANCOVA underestimated the treatment difference compared to the ANCOVA on all responses. Simple and multiple imputation

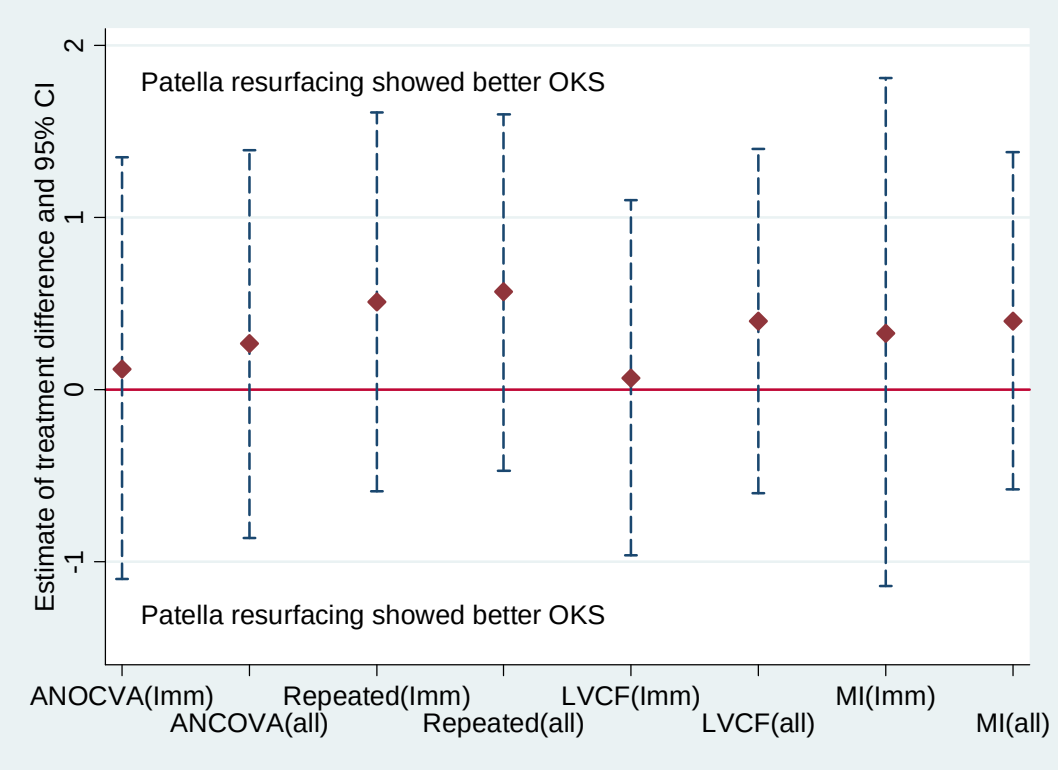
following all observed responses provided very similar results. The magnitude of the difference was less for the methods based on immediate responses compared to the equivalent method based on all observed responses.

Table 11.4: KAT – Estimates of treatment difference in OKS scores for different analysis strategies

		Treatment difference at 12m			
Strategy	N	Estimate	SE	95% CI	p-value
Immediate responses only (ANCOVA)	900	0.12	0.62	(-1.10, 1.35)	0.84
All observed responses (ANCOVA)	1091	0.27	0.57	(-0.86, 1.39)	0.64
Immediate responses (repeated measures)	1591	0.51	0.56	(-0.59, 1.61)	0.36
All observed responses (repeated measures)	1591	0.57	0.53	(-0.47, 1.60)	0.28
Immediate responses plus LVCF	1591	0.07	0.53	(-0.96, 1.10)	0.90
All observed responses plus LVCF	1591	0.40	0.51	(-0.60, 1.40)	0.43
Immediate responses plus MI*	1715	0.33	0.68	(-1.14, 1.81)	0.64
All observed responses plus MI*	1715	0.40	0.50	(-0.58, 1.38)	0.42

* MCMC imputation for all the missing data

Figure 11.6: KAT– Estimates of treatment difference (95% CI) in the OKS for different analysis strategies



The comparison of patella resurfacing versus no patella resurfacing in KAT recruited 1715 participants. The published ANCOVA analysis included 64% of these participants. Therefore, since the proportion of missing data is large, imputation would not be perceived as credible by most researchers. The repeated

measures approach allowed 92% of participants to be included. Since the mechanism was found to be MAR, this seems a sensible option. In this case, the result of the patella resurfacing comparison in KAT would have shown a mean difference of 0.57 with 95% CI (-0.47, 0.60). This provided no evidence of a difference in the Oxford Knee Score between treatment groups (p=0.46). Despite the change in analysis approach from that which was reported, the conclusion of the trial remained the same.

11.3.5 PRISM

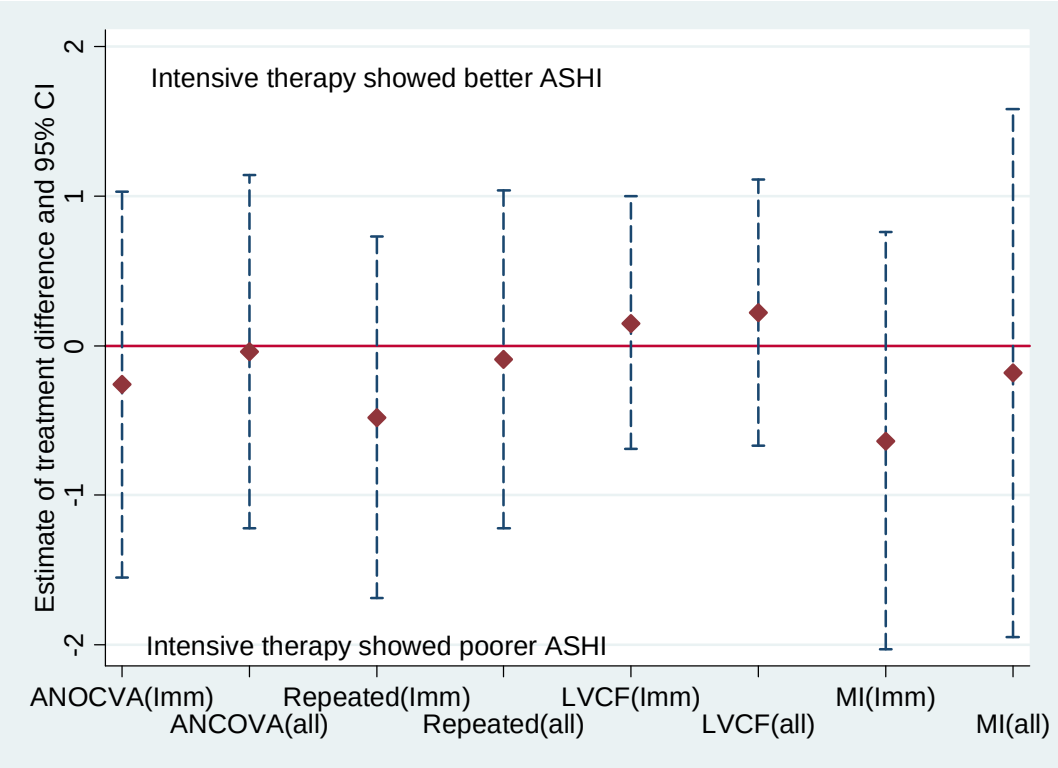
The PRISM trial calculated the Arthritis Index (ASHI) which was an alternative weighting of the SF36 question items. The results under the different analysis strategies are presented in Table 11.5 and Figure 11.7. The published analysis did not find a difference in the ASHI between the two treatment groups (p=0.95). The other analysis strategies also found no significant treatment difference (p>0.05). As with the other example trials, the magnitude and direction of the estimate differed (under simple imputation). All recruited patients were included when multiple imputation was carried out, compared to the 798 (60%) in the reported analysis. Using only immediate responses increased the magnitude of the difference, when compared to the same method on all response data. Although the conclusion of the hypothesis test is unchanged for the different analyses, it is easily seen that a different choice of method can affect the magnitude and direction of any treatment estimate.

Table 11.5: PRISM – Estimates of treatment difference in Arthritis index for different analysis strategies

		Treatment difference at 24m			
Strategy	N	Estimate	SE	95% CI	p-value
Immediate responses (ANCOVA)	622	-0.26	0.66	(-1.55, 1.03)	0.69
All observed responses (ANCOVA)	798	-0.04	0.60	(-1.22, 1.14)	0.95
Immediate responses (repeated measures)	1198	-0.48	0.62	(-1.69, 0.73)	0.43
All observed responses (repeated measures)	1198	-0.09	0.57	(-1.22, 1.04)	0.87
Immediate responses plus mean	1324	0.15	0.43	(-0.69, 1.00)	0.72
All observed responses plus mean	1324	0.22	0.45	(-0.67, 1.11)	0.64
Immediate responses plus MI	1324	-0.64	0.69	(-2.03, 0.76)	0.37
All observed responses plus MI	1324	-0.18	0.82	(-1.95, 1.58)	0.83

The published analysis for PRISM was ANCOVA at 24 months and included 798 (60%) participants. No evidence of a treatment difference in the Arthritis Index was found (p=0.95). Given the poor response rate, imputation should be used with caution. The mechanism of missing data was found to be MCAR and possibly MAR. Thus, a repeated measures approach could be considered. This approach utilised 90% of participants and is likely to be the most suitable analysis. The estimated treatment difference was -0.09 (-1.22, 1.04) and this was not significantly different (p=0.87). This matched the original trial conclusion.

Figure 11.7: PRISM– Estimates of treatment difference (95% CI) in the ASHI for different analysis strategies



11.3.6 Summary

The sections above presented the estimate of treatment difference in QoL outcomes for each of the five HSRU trials. Data for the other two trials are not presented. This is due to the fact that the comparison analysis (ANCOVA on all observed responses) was not actually the true analysis for the trial, but was used for illustrative purposes in this thesis. In REFLUX, the different strategies concluded that there was evidence of a treatment difference in reflux specific QoL.

For MAVIS, KAT and PRISM, no significant treatment differences were found. The different analysis strategies agreed on this. However, the magnitude and in some cases the direction of the effect differed.

The interesting finding was the calcium comparison in RECORD. The published analysis provided borderline evidence of a difference in QoL (EQ5D). However, different analysis strategies provided different conclusions. Multiple imputation and the ANCOVA on immediate responses found no evidence of a difference ($p>0.05$), while the remaining methods did ($p<0.05$). This highlighted that the impact the choice of analysis has on the trial. Earlier in this thesis, it was shown that there was evidence against the MCAR assumption for the missing EQ5D scores in RECORD. There was evidence that these were more likely MAR or even MNAR. This suggested that the ANCOVA or simple imputation strategies were likely to be inappropriate. The repeated measures and MI strategies make the MAR assumption, yet these methods differed with respect to the result of the hypothesis test. Repeated measures found evidence of treatment difference ($p<0.05$), while multiple imputation did not ($p>0.05$).

The response rate at 24 months in RECORD was 54%. The use of imputation is questionable here, as nearly half the data would have been imputed. Most readers would not believe the eventual result. The more appropriate strategy would be the use of repeated measures, as 74% of participants could be included and there was evidence that the MAR assumption was satisfied. In this case, the estimate of treatment difference was 0.017 (0.003, 0.023) with $p=0.02$. Therefore, despite the fact that it was reported that there was borderline evidence of a difference in QoL between treatment groups (calcium versus no calcium), it is perhaps more likely that there was a significant difference.

11.4 Limitations and Future Work

There were a number of limitations of this work. Firstly, the issue of missing items was ignored. Missing forms are only one type of missing data and the second type should be considered when analysing the trial data. An assumption was made throughout that if a QoL score was not calculated, then the participant had not returned the questionnaire. This was not always the case. It was possible that the participant had not completed sufficient items for the score to be calculated. The imputation of these items could have been undertaken in order to enable the calculation of the dimension score. In some situations this did happen, as the half-item rule came into play. For example, in the QLQ-C30 and SF36, if at least half the items within a particular scale are present, the mean of these items is used to impute the missing items. Future work could consider the issue of missing items, in addition to that of missing forms.

A second problem encountered was the death of participants during follow-up. Undertaking imputation for these people is in some sense not sensible as they were not alive to have QoL. In some QoL measures, such as the EQ5D, the value zero represents death. This value could be imputed where appropriate. This issue was ignored when undergoing imputation, which has some obvious implications as patients who were dead were being given a QoL value. This would not have been representative of their actual health state. This was only really a major problem in the NPC trial, where there were a number of deaths. In the remaining trials, the number of deaths was minimal. Some further consideration of this issue is needed, in particular, how best to handle missing data due to death. To some extent a joint model could be useful. The implementation of this model needs further investigation.

Six of the seven example trials (not NPC trial) administered both first and second reminders. The work carried out here did not distinguish between whether the reminder response was a consequence of the first or second reminder, only that it was after reminder. It would be of interest to investigate the impact of the second reminder in order to determine whether it is necessary.

In each of the trials, there was a small amount of missing baseline information. This was particularly a problem in RECORD, as the four-month assessment was regarded as baseline. This had been administered through a postal questionnaire and was susceptible to missing data. In the remaining trials, the baseline questionnaire had been issued at a clinic appointment and thus, should have been complete. The issue of missing items also came into play here. Some baseline scores were not calculable. Missing baseline scores were a problem when it came to analysis, as the patient could not be included in the ANCOVA. Following some work by White and Thompson, it is now common practice in CHaRT to impute missing baseline scores where necessary, using simple mean imputation (White, Thompson 2005). This was not the case at the time when the example trials were undertaken. This process ensures that everybody would have at least one QoL assessment and, if a repeated measures approach was used, every participant could then be included in the analysis.

An alternative model-based strategy to adjust for non-randomly missing outcomes has been proposed by Baker *et al.* (Baker et al. 2006). This approach generates a propensity-to-be-missing score for each randomisation group arising from informative covariates. In the context of longitudinal outcomes, a monotone missing data pattern is required. The idea behind this approach is that the use of informative covariates transforms a non-ignorable missing data mechanism into an ignorable one. This avoids the need for strong assumptions about the missing data mechanism. Some further research is needed into the practical application of this method, as an alternative to the methods outlined in this thesis.

All datasets were obtained from completed trials and I was not involved in the undertaking of these trials. Future work could include being involved in a trial from the outset, taking all the issues surrounding missing data into consideration. Primarily, avoiding missing data is of paramount importance. The reminder system is one way of remedying this, but potentially, there are others. The use of telephone reminders and completion of questionnaires over the telephone was not considered here. This alternative to the postal reminder clearly has different cost

implications. It may be possible to compare these strategies using a second randomisation, following non-response, to assess which method is more effective in terms of additional data and more cost-effective.

Six of the seven example datasets came from the same research institution. These particular trials are all of a similar nature, in that a single intervention was given and QoL assessed periodically thereafter. The early impetus for QoL assessment arose alongside cancer clinical trials where there were repeated treatments, for example, chemotherapy. In these trials, QoL was assessed at both on and off treatment time points. This type of trial differs from the example trials, but the principle behind appropriate strategies for missing data remains the same. However, one would need to be careful about using 'on treatment' values to impute 'off treatment' ones. Rather than using last value carried forwards in the traditional sense, you could use last 'on treatment' value carried forwards for a missing value 'on treatment'. In a similar manner, if using repeated measures either a variable indicating on or off treatment could be included in the model, or one could use only the 'on treatment' values in the model. Despite the differences that the example trials may have to other clinical trial settings, strategies behind dealing with the missing data are the same. The conclusions from this thesis could be applied.

The use of imputation needs to be explored further, in particular with regard to its 'quality'. More sophisticated economic techniques could be used to elicit a value for the quality of imputation in order to better inform the cost-effectiveness calculations. In addition, the 'quality' of reminder data was assumed to be perfect in the work presented. The accuracy of this assumption could be explored using similar techniques to that which were applied to the quality of imputation. In some of the example trials, there was evidence of MNAR data. Further work is needed on the analysis methods appropriate in this situation. One suggestion which would be applicable to the intermittent missing data pattern was the model proposed by Troxel *et al.* (Troxel, Lipsitz & Harrington 1998). This is an extension of the model proposed by Diggle and Kenward (Diggle, Kenward 1994) .

11.5 Conclusion

All of the work presented in this thesis has highlighted the need for trial researchers to consider the issues surrounding missing data more formally. There are a wide range of methods for handling missing data and clearly no single method is best for all situations. In fact, the recommended approach is to use several methods and compare the results across varying assumptions (Huntington, Dueck 2005). Determining the correct method to handle missing data depends primarily on the applicability of the assumptions behind the method. In other words, the presence of similar results regardless of the methods used (or more importantly the assumptions made) gives validity to the conclusions. In situations in which results vary depending on the methods used, a more careful consideration of the assumptions is necessary. Results must be interpreted with respect to the assumptions of each method. The work presented has an advantage over the current literature in that the reminder-response data are utilised. The accuracy of the different methods could be assessed as the true values were known. This unique approach has shown that the reminder-responses are extremely important and have a significant role to play in the collection of follow up data. This will ultimately improve the reliability of the trial conclusion.

The work has shown that the mechanism of missing data should be investigated and that the conclusions should be used to inform the most appropriate analysis. In the unlikely situation that data can be confirmed as being MCAR, complete case analysis or simple methods of imputation could be used. In the more likely situation of MAR data, multiple imputation is useful (Carpenter, Kenward 2007). An alternative method would be available case analysis. In the longitudinal setting a repeated measures model would be appropriate. When data is thought likely to be MNAR, more sophisticated approaches such as joint modelling or pattern mixtures models should be considered (Fairclough 2002). The reminder-responses play an important role in determining this missing data mechanism. Without them, you may get a distorted view of the mechanism, which could result in an inappropriate choice of analysis strategy.

The results showed that the choice of imputation method can have a bearing on the result. It is important to consider which method will be most appropriate. Huson *et al*. suggested that there is no one imputation technique which is applicable in all situations (for all missing data patterns and missing data mechanisms) (Huson, Chung & Salgo 2007). It is generally regarded that imputation is a tool to assess sensitivity of results rather than a primary analysis approach (Fairclough 2002, Fayers, Machin 2007, Carpenter, Kenward 2007). This is particularly prominent when the amount of missing data is significant. The credibility of trial results will be reduced if more than about 20% of data is imputed. Trial researchers should collect as much data as possible through the initial wave of questionnaires and through reminders. If any missing data remains, the reminder-responses can be used to help inform the most appropriate imputation method.

The reminder strategy was shown to be of economic benefit in smaller samples (<1000). The benefit in larger samples remains unclear. The proportion of missing data will be a major factor in the cost-effectiveness. There is still uncertainty in the 'quality' and cost of imputation. Thus, any potential benefit in terms of reduced cost is counter-acted by the potential for bias in the results. More sophisticated economic techniques such as the discrete choice experiment could be used to elicit the willingness to pay for perfect information and the quality of imputation. Where possible the reminder strategy is likely to be the better option, with imputation used as a tool for assessing sensitivity of trial results.

11.6 Recommendations

Following the work presented in this thesis, there are a number of recommendations I would make to researchers analysing trial data with missing QoL outcomes. A step by step strategy is provided in Figures 11.8 and 11.9, but in summary, the recommendations are:

1. Design the trial such that the proportion of missing data is kept to a minimum.
2. In trials of less than 1000 patients, a reminder-system should be used.
3. In larger trials, the use of reminders should be considered if the proportion of missing data is high but will be subject to financial constraints.
4. If missing data exists, the first step is to investigate reasons for missingness and identify the missing data mechanism. For an overall view, use Little's hypothesis test (Little 1988). If the mechanism of dropout at or after a given assessment is of interest, then use Fairclough (Fairclough 2002) or Ridout (Ridout 1991) logistic regression as appropriate. Make use of the reminder data as has been shown here.
5. To identify the most suitable method of imputation, make use of the data collected by reminder. Delete the reminder data, carry out imputation and assess the accuracy using the methods outlined in the thesis.
6. Simple imputation could be considered if the mechanism is found to be MCAR.
7. If there is evidence against the MCAR assumption a multiple imputation procedure is more suitable.
8. It is recommended that a sensitivity of the results is included using several methods of imputation.
9. Rather than carrying out a complete-case analysis on the final endpoint, use a repeated measures approach which makes use of all the available QoL assessments. This method makes the assumption of MAR which is more plausible in this setting.

The flow diagrams presented in Figures 11.8 and 11.9 can provide researchers with a step by step guide as how to deal with missing data. The reminder responses are an important part of this process. As has been discussed previously, the first stage in deciding how to deal with missing data is to consider the missing data mechanism (Figure 11.9). The findings from this process then feed into Figure 11.8 and the flow diagram is used to decide on the appropriate method of analysis and/or the use of imputation. The different stages of these flow diagrams are explained throughout this thesis and some direction as to where the information can be found is provided.

This thesis deals with longitudinal outcomes, but the idea is the same for a single endpoint (perhaps with baseline assessment). If a study contains only a single post intervention assessment, then the missing data mechanism would be investigated at this assessment using Fairclough's method. The result from this could then be fed into flow diagram one (Figure 11.8). Simple or multiple imputation could be considered, as appropriate. The number of available methods would be reduced, as there may or may not be a baseline assessment. In studies with a single endpoint, repeated measures would not be appropriate, but if the data were found to be MAR a multiple imputation procedure could be used. Any data collected through reminders could help inform the choice of model, as has been outlined in this thesis.

The choice between different approaches for missing data will also depend on the amount of data missing. Schulz and Grimes give a general rule of thumb with regard to missing data (Schulz, Grimes 2002). They state that in a trial with less than 5% missing, the bias will be minimal. A trial with over 20% missing poses a serious threat to the validity of the study. In between 5% and 20% missing leads to intermediate levels of problems. This general rule can be applied alongside the approaches set out in this thesis. Imputation is often only regarded as a plausible option when the amount of missing data is less than 20%. Undertaking imputation with more than 20% missing should be done so with caution, as it is likely that the result of the trial would not be accepted by the research community. This is provided as a guideline and not a rule for all scenarios. In conclusion,

there is no single way of dealing with missing data that is applicable in all situations. The approaches outlined in this thesis and the recommendations above provide the researcher with some tools to develop a strategy to deal with the problem of missing QoL outcomes when analysing clinical trial data.

The role of reminders is extremely important when collecting follow-up data within a clinical trial. The use of a reminder system is a cost-effective use of resources to maintain the sample size. Using reminders to minimize the amount of missing data also reduces the threat of bias. Finally, data collected by reminders enables a more informed selection of potential imputation methods, which again reduces the risk of bias. Ultimately, the aim of any trial is to obtain an unbiased as possible estimate of treatment difference to help inform and improve clinical practice to the benefit of patients; the use of reminders is pivotal in this.

Figure 11.8: Flow diagram 1 – Strategy for dealing with missing longitudinal QoL data

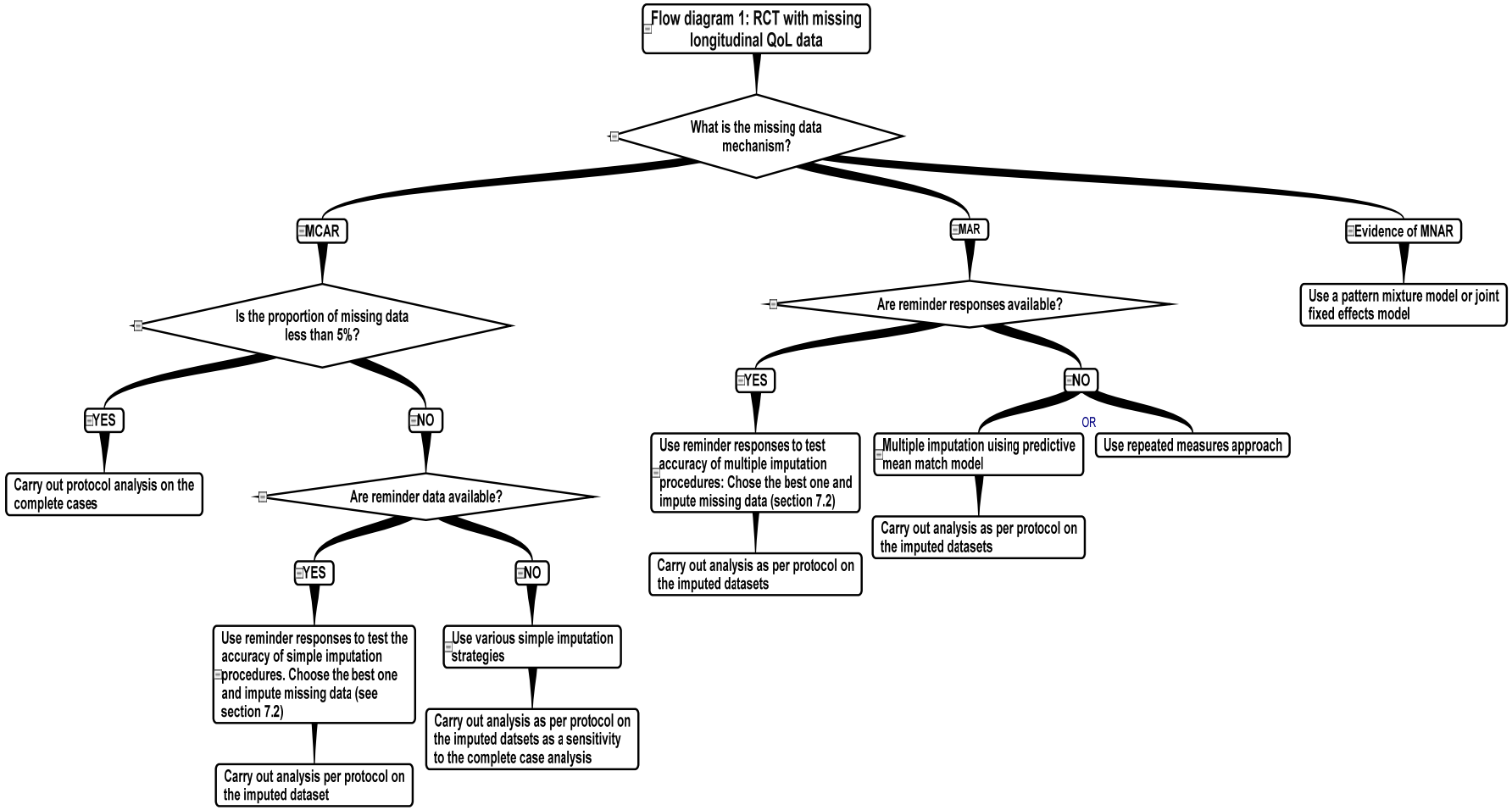
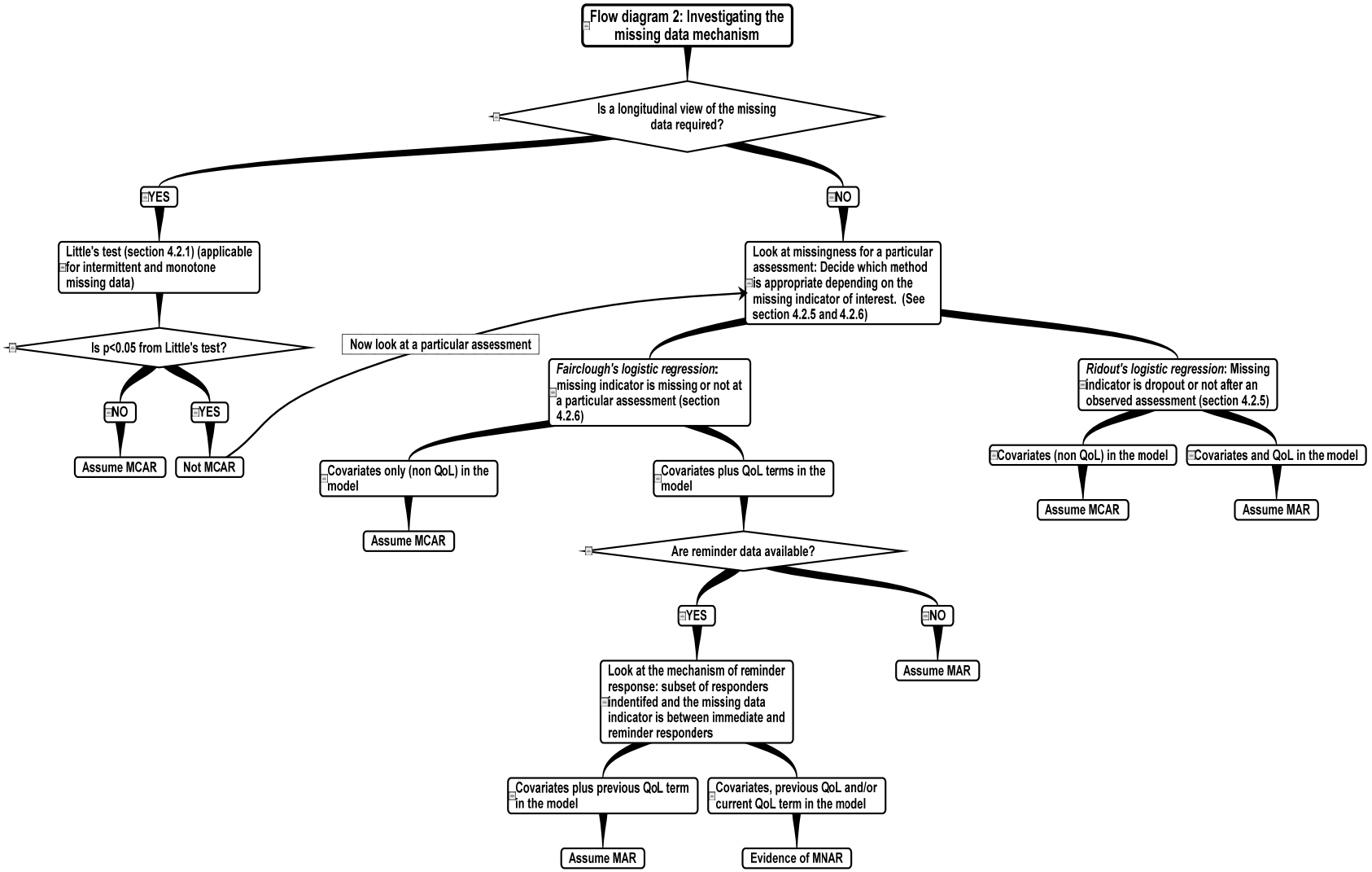


Figure 11.9: Flow diagram 2 – Identifying the missing data mechanism



References

Aaronson, N.K., Ahmedzai, S., Bergman, B., Bullinger, M., Cull, A., Duez, N., Filiberti, A., Flechtner, H., Fleishman, S., de Haea, J., Kaasa, S., Klee, M., Osoba, D., Razavi, D., Rofe, P., Schraub, S., Sneeuw, K., Sullivan, M. & Takeda, F. 1993, "The European Organisation 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,* , no. 85, pp. 365-376.

Avenell, A., Campbell, M.K., Cook, J.A., Hannaford, P.C., Kilonzo, M.M., McNeill, G., Milne, A.C., Ramsay, C.R., Seymour, D.G., Stephen, A.I. & Vale, L.D. Effect of multivitamin and multimineral supplements on morbidity from infections in older people (MAVIS trial): pragmatic, randomised, double blind, placebo controlled trial. *BMJ* 2005; **331** (7512): 324-329.

Baker, S.G., Fitzmaurice, G.M., Freedman, L.S. & Kramer, B.S. Simple adjustments for randomized trials with nonrandomly missing or censored outcomes arising from informative covariates. *Biostatistics* 2006; **7**(1): 29-40.

Ballard, C., Margallo-Lana, M., Juszczak, E., Douglas, S., Swann, A., Thomas, A., O'Brien, J., Everratt, A., Sadler, S., Maddison, C., Lee, L., Bannister, C., Elvish, R. & Jacoby, R. Quetiapine and rivastigmine and cognitive decline in Alzheimer's disease: randomised double blind placebo controlled trial. *BMJ* 2005; **330**(7496): 874.

Berry, M.A., Hargadon, B., Shelley, M., Parker, D., Shaw, D.E., Green, R.H., Bradding, P., Brightling, C.E., Wardlaw, A.J. & Pavord, I.D. Evidence of a role of tumor necrosis factor alpha in refractory asthma. *The New England journal of medicine* 2006; **354**(7): 697-708.

Blumenthal, J.A., Sherwood, A., Babyak, M.A., Watkins, L.L., Waugh, R., Georgiades, A., Bacon, S.L., Hayano, J., Coleman, R.E. & Hinderliter, A. Effects of exercise and stress management training on markers of

cardiovascular risk in patients with ischemic heart disease: a randomized controlled trial. *JAMA* 2005; **293**(13): 1626-1634.

Brooks, R with the EuroQoL Group. EuroQoL: the current state of play. *Health Policy* 1996; **37**: 53-72.

Brown, H. & Prescott, R. *Applied Mixed Models in Medicine* 1999. Wiley.

Brown, C.H. Protecting against nonrandomly missing data in longitudinal studies. *Biometrics* 1990; **46**(1): 143-155.

Burnham, K.P. & Anderson, D.R. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods Research* 2004; 33(2): 261-304.

Buszewicz, M., Rait, G., Griffin, M., Nazareth, I., Patel, A., Atkinson, A., Barlow, J. & Haines, A. Self management of arthritis in primary care: randomised controlled trial. *BMJ* 2006; **333**(7574): 879.

Carpenter, J.R. & Kenward, M.G. *Missing data in randomised controlled trials - a practical guide*. November 2007 Available: http://www.pcpoh.bham.ac.uk/publichealth/methodology/docs/invitations/Final_Report_RM04_JH17_mk.pdf [2007, 28/11].

Cook, N.R. An imputation method for non-ignorable missing data in studies of blood pressure. *Statistics in Medicine* 1997; **16**(23): 2713-2728.

Curran, D., Bacchi, M., Schmitz, S.F., Molenberghs, G. & Sylvester, R.J. Identifying the types of missingness in quality of life data from clinical trials. *Statistics in Medicine* 1998a; 17(5-7): 739-756.

Curran, D., Molenberghs, G., Fayers, P.M. & Machin, D. Incomplete quality of life data in randomized trials: missing forms. *Statistics in Medicine* 1998b: **17**(5-7): 697-709.

Curran, D., Molenberghs, G., Thijs, H. & Verbeke, G. Sensitivity analysis for pattern mixture models. *Journal of Biopharmaceutical Statistics* 2004: **14**(1): 125-143.

Dawson, J., Fitzpatrick, R., Murray, D. & Carr, A. Questionnaire on the perceptions of patients about total knee replacement. *Journal of Bone & Joint Surgery* 1998; **80-B**: 63-69.

De Gruttola, V. & Xin Ming, T. Modelling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics* 1994; **50**(4): 1003-1014.

Diggle, P.J. Testing for random dropouts in repeated measurements data. *Biometrics* 1989; **45**: 1255-1258.

Diggle, P.J., Heagerty, P., Liang, K.Y. & Zeger, S.L. *Analysis of Longitudinal Data (2nd edition)* 2002. Oxford University Press.

Diggle, P. & Kenward, M.G. Informative Drop-Out in Longitudinal Data-Analysis. *Applied Statistics-Journal of the Royal Statistical Society Series C* 1994; **43**(1): 49-93.

Donders, A.R., van der Heijden, G.J., Stijnen, T. & Moons, K.G. Review: a gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology* 2006; **59**(10): 1087-1091.

Drummond, M.F., O'Brien, B., Stoddart, G.L. & Torrance, G.W. *Methods for the Economic Evaluation of Health Care Programmes (2nd edition)* 1997. Oxford University Press, Oxford.

Efron, B. & Tibshirani, R.J. *An Introduction to the Bootstrap* 1993. Chapman and Hall, London.

Encrenaz, G., Rondeau, V., Messiah, A. & Auriacombe, M. Examining the influence of drop-outs in a follow-up of maintained opiate users. *Drug & Alcohol Dependence* 2005; **79**(3): 303-310.

Engels, J.M. & Diehr, P. Imputation of missing longitudinal data: a comparison of methods. *Journal of Clinical Epidemiology* 2003; **56**(10): 968-976.

Fairbank, J., Frost, H., Wilson-MacDonald, J., Yu, L.M., Barker, K., Collins, R. & Spine Stabilisation Trial, G. Randomised controlled trial to compare surgical stabilisation of the lumbar spine with an intensive rehabilitation

programme for patients with chronic low back pain: the MRC spine stabilisation trial. *BMJ* 2005; **330**(7502): 1233.

Fairclough, D.L. *Design and Analysis of Quality of Life Studies in Clinical Trials* 2002. Chapman and Hall.

Fairclough, D.L., Gagnon, D.D. & Zagari, M.J. Benefits of epoetin alfa for cancer patients' quality of life are confirmed after modelling to account for missing data. *Current Medical Research & Opinion* 2005; 21(Suppl 2): S6-8.

Fairclough, D.L., Gagnon, D.D., Zagari, M.J., Marschner, N., Dicato, M. & Epoetin Alfa Study, G. Evaluation of quality of life in a clinical trial with nonrandom dropout: the effect of epoetin alfa in anemic cancer patients. *Quality of Life Research* 2003; **12**(8): 1013-1027.

Fairclough, D.L., Peterson, H.F. & Chang, V. Why are missing quality of life data a problem in clinical trials of cancer therapy? *Statistics in Medicine* 1998; **17**(5-7): 667-677.

Fairclough, D.L., Thijs, H., Huang, I.C., Finnern, H.W. & Wu, A.W. Handling missing quality of life data in HIV clinical trials: what is practical? *Quality of Life Research* 2008; **17**(1): 61-73.

Fay, R.E. When are inferences from multiple imputations valid? *Proceedings of the Survey Research Methods Section of the American Statistical Association* 1992: pg 227.

Fayers, P.M., Aaronson, N.K., Bjordal, K., Groenvold, M., Curran, D., Bottomley, A. & on behalf of the EORTC Quality of Life Group. *The EORTC QLQ-C30 Scoring Manual (3rd edition)* 2001. European Organisation for Research and Treatment of Cancer, Brussels.

Fayers, P.M. & Machin, D. *Quality of Life: The assessment, analysis and interpretation of patient-reported outcomes (2nd edition)* 2007. Wiley, UK.

Fayers, P.M. & Machin, D. *Quality of Life: Assessment, Analysis and Interpretation* 2001. Wiley.

Feagan, B.G., Greenberg, G.R., Wild, G., Fedorak, R.N., Pare, P., McDonald, J.W., Dube, R., Cohen, A., Steinhart, A.H., Landau, S., Aguzzi, R.A., Fox, I.H. & Vandervoort, M.K. Treatment of ulcerative colitis with a humanized antibody to the alpha4beta7 integrin. *The New England Journal of Medicine* 2005; **352**(24): 2499-2507.

Fielding, S., Fayers, P.M., McDonald, A., McPherson, G. & Campbell, M.K. Simple imputation methods were inadequate for missing not at random (MNAR) quality of life data. *Health & Quality of Life Outcomes 2008a*; **6**(57).

Fielding, S., Maclennan, G., Cook, J.A. & Ramsay, C.R. A review of RCTs in four medical journals to assess the use of imputation to overcome missing data in quality of life outcomes. *Trials* 2008b; **9**(51).

Gadbury, G.L., Coffey, C.S. & Allison, D.B. Modern statistical methods for handling missing repeated measurements in obesity trial data: beyond LOCF. *Obesity Reviews* 2003; **4**(3): 175-184.

Grant, A.M., Wileman, S.M., Ramsay, C.R., Mowat, N.R., Krukowski, Z.H., Heading, R.C., Thursz, M.R., Campbell, M.K. & and the REFLUX Trial Group. Minimal access surgery compared with medical management for chronic gastro-oesophageal reflux disease: UK collaborative randomised trial. *BMJ* 2009; **337**: a2664.

Grant, A., Wileman, S.M., Ramsay, C., Bojke, L., Epstein, D., Sculpher, M., Macran, S., Kilonzo, M., Vale, L., Francis, J., Mowat, A., Krukowski, Z., Heading, R.C., Thursz, M., Russell, I., Campbell, M.K. & on behalf of the REFLUX trial group. The effectiveness and cost-effectiveness of minimal access surgery amongst people with gastro-oesophageal reflux disease - a UK collaborative study. The REFLUX trial. *Health Technology Assessment* 2008; **12**(31): 1-204.

Heitjan, F. & Little, R.J.A. Multiple imputation for the fatal accident reporting system. *Applied Statistics* 1991; 40: 13-29.

Hosmer, D.W. & Lemeshow, S. *Applied Logistic Regression* 1989. Wiley.

Houck, P.R., Mazumdar, S., Koru-Sengul, T., Tang, G., Mulsant, B.H., Pollock, B.G. & Reynolds, C.F. Estimating treatment effects from

longitudinal clinical trial data with missing values: comparative analyses using different methods. *Psychiatry Research 2004;* **129**(2): 209-215.

Hsieh, L.L., Kuo, C.H., Lee, L.H., Yen, A.M., Chien, K.L. & Chen, T.H. Treatment of low back pain by acupressure and physical therapy: randomised controlled trial. *BMJ* 2006; **332**(7543): 696-700.

Hunkeler, E.M., Katon, W., Tang, L., Williams, J.W.,Jr, Kroenke, K., Lin, E.H., Harpole, L.H., Arean, P., Levine, S., Grypma, L.M., Hargreaves, W.A. & Unutzer, J. Long term outcomes from the IMPACT randomised trial for depressed elderly patients in primary care. *BMJ* 2006; **332**(7536): 259-263.

Hunsberger, S., Murray, D., Davis, C.E. & Fabsitz, R.R. Imputation strategies for missing data in a school-based multi-centre study: the Pathways study. *Statistics in Medicine* 2001; **20**(2): 305-316.

Huntington, J.L. & Dueck, A. Handling missing data. *Current problems in cancer* 2005: **29**(6): 317-325.

Huson, L.W., Chung, J. & Salgo, M. Missing data imputation in two phase III trials treating HIV1 infection. *Journal of Biopharmaceutical Statistics* 2007; **17**: 159-172.

Jordhøy, M.S., Fayers, P.M., Loge, J.H., Ahlner-Elmqvist, M. & Kaasa, S. Quality of life in palliative cancer care: results from a cluster randomized trial. *Journal of Clinical Oncology* 2001; **19**(18): 3884-3894.

Jordhøy, M.S., Kaasa, S., Fayers, P., OVreness, T., Underland, G. & Ahlner-Elmqvist, M. Challenges in palliative care research; recruitment, attrition and compliance: Experience from a randomized controlled trial. *Palliative Medicine* 1999; **13**(4): 299-310.

Kaplan, S.A., Roehrborn, C.G., Rovner, E.S., Carlsson, M., Bavendam, T. & Guan, Z. Tolterodine and tamsulosin for treatment of men with lower urinary tract symptoms and overactive bladder: a randomized controlled trial. *JAMA* 2006; **296**(19): 2319-2328.

Keller, S.D., Majkut, T.C., Kosinski, M. & Ware, J.E.,Jr. Monitoring health outcomes among patients with arthritis using the SF-36 Health Survey: overview. *Medical Care* 1999; **37**(Suppl 5): MS1-9.

Kennedy, T., Jones, R., Darnley, S., Seed, P., Wessely, S. & Chalder, T. Cognitive behaviour therapy in addition to antispasmodic treatment for irritable bowel syndrome in primary care: randomised controlled trial. *BMJ* 2005; **331**(7514): 435.

Kenward, M.G. & Carpenter, J. Multiple imputation: Current perspectives. *Statistical Methods in Medical Research* 2007; **16**(3): 199-218.

Korzenik, J.R., Dieckgraefe, B.K., Valentine, J.F., Hausman, D.F., Gilbert, M.J. & Sargramostim in Crohn's Disease Study Group. Sargramostim for active Crohn's disease. *The New England Journal of Medicine* 2005; **352**(21): 2193-2201.

Listing, J. & Schlittgen, R. A nonparametric test for random dropouts. *Biometrical Journal* 2003; **45**(1): 113-127.

Listing, J. & Schlittgen, R. Tests if dropouts are missed at random. *Biometrical Journal* 1998; **40**(8): 929-935.

Little, R.J.A. A class of pattern-mixture models for normal incomplete data. *Biometrika* 1994; **81(**3): 471-483.

Little, R.J.A. A Test of Missing Completely at Random for Multivariate Data With Missing Values. *Journal of American Statistical Association* 1988; **83**(404): 1198-1202.

Little, R.J.A. & Rubin, D.B. *Statistical Analysis with Missing Data (2nd edition)* 2002. Wiley.

Liu, G. & Gould, A.L. Comparison of alternative strategies for analysis of longitudinal trials with dropouts. *Journal of Biopharmaceutical Statistics* 2002; **12**(2): 207-226.

Macran, S., Wileman, S., Barton, G., Russell, I. & REFLUX trial group. The development of a new measure of quality of life in the management of

gastro-oesophageal reflux disease: the Reflux questionnaire. *Quality of Life Research* 2007; **16**(2): 331-343.

McIntosh, E., Donaldson, C. & Ryan, M. Recent advances in the methods of cost-benefit analysis in healthcare. Matching the art to the science. P*harmacoEconomics* 1999; **15**(4): 357-367.

McManus, R.J., Mant, J., Roalfe, A., Oakes, R.A., Bryan, S., Pattison, H.M. & Hobbs, F.D. Targets and self monitoring in hypertension: randomised controlled trial and cost effectiveness analysis. *BMJ* 2005; **331**(7515): 493.

Meggitt, S.J., Gray, J.C. & Reynolds, N.J. Azathioprine dosed by thiopurine methyltransferase activity for moderate-to-severe atopic eczema: a double-blind, randomised controlled trial. *Lancet* 2006; **367**(9513): 839-846.

Moher, D., Schulz, K.F. & Altman, D.G. The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet* 2001; **357**(9263): 1191-1194.

Molenberghs, G. & Kenward, M.G. *Missing Data in Clinical Studies* 2007. Wiley.

Molenberghs, G., Thijs, H., Jansen, I., Beunckens, C., Kenward, M.G., Mallinckrodt, C. & Carroll, R.J. Analyzing incomplete longitudinal clinical trial data. *Biostatistics* 2004; **5**(3): 445-464.

Morita, S., Kobayashi, K., Eguchi, K., Matsumoto, T., Shibuya, M., Yamaji, Y. & Ohashi, Y. Analysis of incomplete quality of life data in advanced stage cancer: A practical application of multiple imputation. *Quality of Life Research* 2005; **14**(6): 1533-1544.

Musil, C.M., Warner, C.B., Yobas, P.K. & Jones, S.L. A comparison of imputation techniques for handling missing data. *Western Journal of Nursing Research* 2002; **24**(7): 815-829.

Myers, W.R. Handling missing data in clinical trials: An overview. *Drug Information Journal* 2000; **34**(2): 525-533.

Nair, K.S., Rizza, R.A., O'Brien, P., Dhatariya, K., Short, K.R., Nehra, A., Vittone, J.L., Klee, G.G., Basu, A., Basu, R., Cobelli, C., Toffolo, G., Dalla Man, C., Tindall, D.J., Melton, L.J.,3rd, Smith, G.E., Khosla, S. & Jensen, M.D. DHEA in elderly women and DHEA or testosterone in elderly men. *The New England Journal of Medicine* 2006; **355**(16): 1647-1659.

Osoba, D., Rodrigues, G., Myles, J., Zee, B. & Pater, J. Interpreting the significance of changes in health-related quality-of- life scores. *Journal of Clinical Oncology* 1998; **16**(1): 139-144.

Patrician, P.A. Multiple imputation for missing data. *Research in Nursing & Health* 2002; **25**(1): 76-84.

Petersen, L., Jeppesen, P., Thorup, A., Abel, M.B., Ohlenschlaeger, J., Christensen, T.O., Krarup, G., Jorgensen, P. & Nordentoft, M. A randomised multicentre trial of integrated versus standard treatment for patients with a first episode of psychotic illness. *BMJ* 2005a; **331**(7517): 602.

Petersen, R.C., Thomas, R.G., Grundman, M., Bennett, D., Doody, R., Ferris, S., Galasko, D., Jin, S., Kaye, J., Levey, A., Pfeiffer, E., Sano, M., van Dyck, C.H., Thal, L.J. & Alzheimer's Disease Cooperative Study Group. Vitamin E and donepezil for the treatment of mild cognitive impairment. *The New England Journal of Medicine* 2005b; **352**(23): 2379-2388.

Pocock, S.J. *Clinical Trials: A Practical Approach* 1983. John Wiley & Sons.

Ralston, S.H., Langston, A.L., Campbell, M.K., MacLennan, G., Selby, P.L. & Fraser, W.D. Preliminary results from the PRISM study: a multicentre randomised controlled trial of intensive vs. symptomatic management for Paget's disease of bone. *Endocrine Abstracts* 2006; **12:** no. OC15.

Ridout, M.S. Testing for random dropouts in repeated measurement data. *Biometrics* 1991; **47**(4): 1617-1619.

Royston, P. Multiple imputation of missing values: update of **ice.** *Stata Journal* 2005; **5**: 527.

Royston, P. Multiple Imputation of Missing Values. *The Stata Journal* 2004; **4**(3): 227-241.

Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys* 1987. John Wiley & Sons, Inc, New York.

Rubin, D.B. Inference and missing data. *Biometrika* 1976; **72**: 359-364.

Rubin, D.B. & Schenker, N. Multiple imputation in health-care databases: an overview and some applications. *Statistics in Medicine* 1991; **10**(4): 585-598.

Ryan, M., Bate, A., Eastmond, C.J. & Ludbrook, A. Use of discrete choice experiments to elicit preferences. *Quality in Health Care* 2001; **10**(Suppl 1): 55-60.

SAS Institute Inc. *SAS/STAT 9.1 User's Guide* 2004. Cary, NC.

Schafer, J.L. *Analysis of Incomplete Multivariate Data (1st edition)* 1997. Chapman and Hall.

Schluchter, M.D. Methods for the analysis of informatively censored longitudinal data. *Statistics in Medicine* 1992; **11**(14-15): 1861-1870.

Schmitz, N. & Franz, M. A bootstrap method to test if study dropouts are missing randomly. *Quality & Quantity* 2002; **36**(1): 1-16.

Schulz, K.F. & Grimes, D.A. Sample size slippages in randomised trials: exclusions and the lost and wayward. *Lancet* 2002; **359**: 781-785.

Simes, R.J., Greatorex, V. & Gebski, V.J. Practical approaches to minimize problems with missing quality of life data. *Statistics in Medicine* 1998; **17**(5-7): 725-737.

Siris, E.S. Paget's disease of bone. *Journal of Bone & Mineral Research* 1998; **13**(7): 1061-1065.

StataCorp. *Stata Statistical Software: Release 10* 2007. College Station, TX: StataCorp LP.

Stinnett, A.A. & Mullahy, J. Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analysis. *Medical Decision Making* 1998; **18**(no. 2 Suppl): S68-80.

Tang, L., Song, J., Belin, T.R. & Unutzer, J. A comparison of imputation methods in a longitudinal randomized clinical trial. *Statistics in Medicine* 2005; **24**(14): 2111-2128.

The KAT trial group. The Knee Arthroplasty Trial (KAT) Design Features, Baseline Characteristics and Two-Year Functional Outcomes after Alternative Approaches to Knee Replacement. *Journal of Bone and Joint Surgery American* 2009; **91**: 134-141.

The RECORD Trial Group. Oral vitamin D3 and calcium for the secondary prevention of low-trauma fractures in elderly people (Randomised Evaluation of Calcium Or Vitamin D, RECORD): a randomised placebo-controlled trial. *Lancet* 2005; **365**: 1621-1628.

Thomas, K.J., MacPherson, H., Thorpe, L., Brazier, J., Fitter, M., Campbell, M.J., Roman, M., Walters, S.J. & Nicholl, J. Randomised controlled trial of a short course of traditional acupuncture compared with usual care for persistent non-specific low back pain. *BMJ* 2006; **333**(7569): 623.

Troxel, A.B., Lipsitz, S.R. & Harrington, D.P. Marginal models for the analysis of longitudinal measurements with nonignorable non-monotone missing data. *Biometrika* 1998; **85**(3): 661-672.

Troxel, A.B., Fairclough, D.L., Curran, D. & Hahn, E.A. Statistical analysis of quality of life with missing data in cancer clinical trials. *Statistics in Medicine* 1998; **17**(5-7): 653-666.

Twisk, J. & de Vente, W. Attrition in longitudinal studies. How to deal with missing data. *Journal of Clinical Epidemiology* 2002; **55**(4): 329-337.

Ware, J.R., Snow, K.K., Kosinski M. & Gandek B. *SF-36 Health Survey Manual and Interpretation Guide* 1993. New England Medical Centre, Boston, MA.

Ware, J.E.,Jr, Keller, S.D., Hatoum, H.T. & Kong, S.X. The SF-36 Arthritis-Specific Health Index (ASHI): I. Development and cross-validation of scoring algorithms. *Medical Care* 1999; **37**(no. 5 Suppl): MS40-50.

White, I.R. & Thompson, S.G. Adjusting for partially missing baseline measurements in randomized trials. *Statistics in Medicine* 2005; **24**(7): 993-1007.

Whynes, D.K., Woolley, C., Philips, Z. & for the TOMBOLA Group 2008. Management of low-grade cervical abnormalities detected at screening: which method do women prefer? *Cytopathology* 2008; **19**: 355.

Winblad, B., Kilander, L., Eriksson, S., Minthon, L., Batsman, S., Wetterholm, A.L., Jansson-Blixt, C., Haglund, A. & Severe Alzheimer's Disease Study Group. Donepezil in patients with severe Alzheimer's disease: double-blind, parallel-group, placebo-controlled study. *Lancet* 2006; **367**(9516): 1057-1065.

Wood, A.M., White, I.R. & Thompson, S.G. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials* 2004; **1**(4): 368-376.

World Health Organization. *Constitution of the World Health Organization* 1948, WHO Basic Documents, Geneva.

Wright, J.G., Wang, E.E., Owen, J.L., Stephens, D., Graham, H.K., Hanlon, M., Nattrass, G.R., Reynolds, R.A. & Coyte, P. Treatments for paediatric femoral fractures: a randomised trial. *Lancet* 2005; **365**(9465): 1153-1158.

Wu, M.C. & Bailey, K.R. Estimation and comparison of changes in the presence of informative right censoring: Conditional linear model. *Biometrics* 1989; **45**(3): 939-955.

Yang, X. & Shoptaw, S. Assessing missing data assumptions in longitudinal studies: an example using a smoking cessation trial. *Drug & Alcohol Dependence* 2005; **77**(3): 213-225.

Yu, L.-., Burton, A. & Rivero-Arias, O. Evaluation of software for multiple imputation of semi-continuous data. *Statistical Methods in Medical Research* 2007; **16**(3): 243-258.

Zigmond, A. & Snaith, R. The Hospital Anxiety and Depression Scale. *Acta Psychiatrica Scandinavica* 1983; **67**: 361-370.

Appendices

Appendix 1.1: SF-36

Please fill in all the questions again by putting a cross in the relevant box of the answer that applies to you. These questions ask for your views about your health and how you feel about life in general.

1. In general, would you say your health is:

Excellent	Very good	Good	Fair	Poor
☐	☐	☐	☐	☐

2. Compared to one year ago, how would you rate your health in general now?

Much better now than one year ago	Somewhat better now than one year ago	About the same as one year ago	Somewhat worse than one year ago	Much worse than one year ago
☐	☐	☐	☐	☐

3. The following questions are about activities you might do during a typical day. Does your health now limit you in these activities? If so, how much?

		Yes limited a lot	Yes limited a little	No, not limited at all
a)	Vigorous activities, such as running, lifting heavy objects, participating in strenuous sport	☐	☐	☐
b)	Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling or playing golf	☐	☐	☐
c)	Lifting or carrying groceries	☐	☐	☐
d)	Climbing several flights of stairs	☐	☐	☐
e)	Climbing one flight of stairs	☐	☐	☐
f)	Bending, kneeling or stooping	☐	☐	☐
g)	Walking more than one mile	☐	☐	☐
h)	Walking several hundred yards	☐	☐	☐
i)	Walking one hundred yards	☐	☐	☐
j)	Bathing or dressing yourself	☐	☐	☐

4. During the past 4 weeks, how much of the time have you had any of the following problems with your work or other regular daily activities as a result of your physical health?

	All of the time	Most of the time	Some of the time	A little of the time	None of the time
a) Cut down on the amount of time you spent on work or other activities	☐	☐	☐	☐	☐
b) Accomplished less than you would like	☐	☐	☐	☐	☐
c) Were limited in the kind of work or other activities	☐	☐	☐	☐	☐
d) Had difficulty performing the work or other activities (for example, it took extra effort)	☐	☐	☐	☐	☐

5. During the past 4 weeks, how much of the time have you had any of the following problems with your work or other regular daily activities as a result of any emotional problems (such as feeling depressed or anxious)?

	All of the time	Most of the time	Some of the time	A little of the time	None of the time
a) Cut down on the **amount of time** you spent on work or other activities	☐	☐	☐	☐	☐
b) **Accomplished less** than you would like	☐	☐	☐	☐	☐
c) Did work or other **activities less carefully than usual**	☐	☐	☐	☐	☐

6. During the past 4 weeks, to what extent has your physical health or emotional problems interfered with your normal social activities with the family, friends, neighbours, or groups?

Not at all	Slightly	Moderately	Quite a bit	Extremely
☐	☐	☐	☐	☐

7. How much bodily pain have you had during the past 4 weeks?

None	Very mild	Mild	Moderate	Severe	Very severe
☐	☐	☐	☐	☐	☐

8. During the past 4 weeks, how much did pain interfere with your normal work (including both outside the home and housework)?

Not at all	A little bit	Moderately	Quite a bit	Extremely
☐	☐	☐	☐	☐

9. These questions are about how you feel and how things have been with you during the past 4 weeks. For each question, please give the one answer that comes closest to the way you have been feeling. How much of the time during the past 4 weeks…

	All of the time	Most of the time	Some of the time	A little of the time	None of the time
a) Did you feel full of life?	☐	☐	☐	☐	☐
b) Have you been very nervous?	☐	☐	☐	☐	☐
c) Have you felt so down in the dumps that nothing could cheer you up?	☐	☐	☐	☐	☐
d) Have you felt calm and peaceful?	☐	☐	☐	☐	☐
e) Did you have a lot of energy?	☐	☐	☐	☐	☐
f) Have you felt downhearted and depressed?	☐	☐	☐	☐	☐
g) Did you feel worn out?	☐	☐	☐	☐	☐
h) Have you been happy?	☐	☐	☐	☐	☐
i) Did you feel tired?	☐	☐	☐	☐	☐

10. During the past 4 weeks, how much of the time has your physical health or emotional problems interfered with your social activities?

All of the time	Most of the time	Some of the time	A little of the time	None of the time
☐	☐	☐	☐	☐

11. How TRUE or FALSE is each of the following statements for you?

	Definitely true	Mostly true	Don't know	Mostly false	Definitely false
a) I seem to get sick a little easier than other people	☐	☐	☐	☐	☐
b) I am as healthy as anybody I know	☐	☐	☐	☐	☐
c) I expect my health to get worse	☐	☐	☐	☐	☐
d) My health is excellent	☐	☐	☐	☐	☐

Appendix 1.2: SF-12

Please fill in all the questions again by putting a cross in the relevant box of the answer that applies to you.

These questions ask for your views about your health and how you feel about life in general.

1. In general, would you say your health is:

Excellent	Very good	Good	Fair	Poor
☐	☐	☐	☐	☐

2. The following questions are about activities you might do during a typical day. Does your health now limit you in these activities? If so, how much?

		Yes limited a lot	Yes limited a little	No, not limited at all
a)	**Moderate activities**, such as moving a table, pushing a vacuum cleaner, bowling or playing golf	☐	☐	☐
b)	Climbing **several** flights of stairs	☐	☐	☐

3. During the past 4 weeks, how much of the time have you had any of the following problems with your work or other regular daily activities as a result of your physical health?

	All of the time	Most of the time	Some of the time	A little of the time	None of the time
a) Accomplished less than you would like	☐	☐	☐	☐	☐
b) Were limited in the kind of work or other activities	☐	☐	☐	☐	☐

4. During the past 4 weeks, how much of the time have you had any of the following problems with your work or other regular daily activities as a result of any emotional problems (such as feeling depressed or anxious)?

	All of the time	Most of the time	Some of the time	A little of the time	None of the time
a) Accomplished less than you would like	☐	☐	☐	☐	☐
b) Did work or other activities less carefully than usual	☐	☐	☐	☐	☐

5. During the past 4 weeks, how much did pain interfere with your normal work (including both outside the home and housework)?

Not at all	A little bit	Moderately	Quite a bit	Extremely
☐	☐	☐	☐	☐

6. These questions are about how you feel and how things have been with you during the past 4 weeks. For each question, please give the one answer that comes closest to the way you have been feeling. How much of the time during the past 4 weeks…

	All of the time	Most of the time	Some of the time	A little of the time	None of the time
a) Have you felt calm and peaceful?	☐	☐	☐	☐	☐
b) Did you have a lot of energy?	☐	☐	☐	☐	☐
c) Have you felt downhearted and depressed?	☐	☐	☐	☐	☐

7. During the past 4 weeks, how much of the time has your physical health or emotional problems interfered with your social activities?

All of the time	Most of the time	Some of the time	A little of the time	None of the time
☐	☐	☐	☐	☐

Appendix 1.3: EuroQoL EQ5D

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 ☐

Appendix 1.4: EORTC QLQ-C30 (version 3)

EORTC QLQ-C30

We are interested in some things about you and your health. Please answer all of the questions yourself by circling the number that best applies to you. There are no "right" or "wrong" answers. The information that you provide will remain strictly confidential.

		Not at all	A little	Quite a bit	Very much
1.	Do you have any trouble doing strenuous activities, like carrying a heavy shopping bag or a suitcase?	1	2	3	4
2.	Do you have any trouble taking a long walk?	1	2	3	4
3.	Do you have any trouble taking a short walk outside of the house?	1	2	3	4
4.	Do you need to stay in bed or a chair during the day?	1	2	3	4
5.	Do you need help with eating, dressing, washing yourself or using the toilet?	1	2	3	4

During the past week:		Not at all	A little	Quite a bit	Very much
6.	Were you limited in doing either your work or other daily activities?	1	2	3	4
7.	Were you limited in pursuing your hobbies or other leisure time activities?	1	2	3	4
8.	Were you short of breath?	1	2	3	4
9.	Have you had pain?	1	2	3	4
10.	Did you need to rest?	1	2	3	4
11.	Have you had trouble sleeping?	1	2	3	4
12.	Have you felt weak?	1	2	3	4
13.	Have you lacked appetite?	1	2	3	4
14.	Have you felt nauseated?	1	2	3	4
15.	Have you vomited?	1	2	3	4
16.	Have you been constipated?	1	2	3	4

Please go on to the next page

During the past week:		Not at all	A little	Quite a bit	Very much
17.	Have you had diarrhoea?	1	2	3	4
18.	Were you tired?	1	2	3	4
19.	Did pain interfere with your daily activities?	1	2	3	4
20.	Have you had difficulty in concentrating on things, like reading a newspaper or watching television?	1	2	3	4
21.	Did you feel tense?	1	2	3	4
22.	Did you worry?	1	2	3	4
23.	Did you feel irritable?	1	2	3	4
24.	Did you feel depressed?	1	2	3	4
25.	Have you had difficulty remembering things?	1	2	3	4
26.	Has your physical condition or medical treatment interfered with your family life?	1	2	3	4
27.	Has your physical condition or medical treatment interfered with your social activities?	1	2	3	4
28.	Has your physical condition or medical treatment caused you financial difficulties?	1	2	3	4

For the following questions please circle the number between 1 and 7 that best applies to you

29. How would you rate your overall health during the past week?

1	2	3	4	5	6	7
Very poor						Excellent

30. How would you rate your overall quality of life during the past week?

1	2	3	4	5	6	7
Very poor						Excellent

Appendix 1.5: Oxford Knee Score

The purpose of the Oxford Knee Score is to help assess the impact that your knee pain has had on your daily life in the past four weeks.

The following questions must ALL be answered on your experiences over the past 4 weeks.

1. How would you describe the pain you usually have from your knee?

None	Very mild	Mild	Moderate	Severe
☐	☐	☐	☐	☐

2. For how long have you been able to walk before the pain in your knee became severe (with or without a walking aid)?

No pain for 30 minutes or more	16-30 minutes	5-15 minutes	Around the house only	Not at all
☐	☐	☐	☐	☐

3. After a meal (sat at a table), how painful is it been for you to stand up from a chair because of your knee?

Not at all Painful	Slightly Painful	Moderately Painful	Very Painful	Unbearable
☐	☐	☐	☐	☐

4. Have you been troubled by pain from your knee in bed at night?

No nights	Only 1 or 2 Nights	Some nights	Most nights	Every night
☐	☐	☐	☐	☐

5. How much has pain from your knee interfered with your usual work, including housework?

Not at all	A little bit	Moderately	Greatly	Totally
☐	☐	☐	☐	☐

6. Could you walk down a flight of stairs?

Yes, easily	With little difficulty	With moderate difficulty	With extreme difficulty	No, Impossible
☐	☐	☐	☐	☐

7. Have you been limping when walking, because of your knee?

Rarely / Never	Sometimes or just at first	Often, not just at first	Most of the time	All of the time
☐	☐	☐	☐	☐

8. Have you felt that your knee might suddenly 'give way' or let you down?

Rarely / Never	Sometimes or just at first	Often, not just at first	Most of the time	All of the time
☐	☐	☐	☐	☐

9. Could you kneel down and get up again afterwards?

Yes, easily	With little difficulty	With moderate difficulty	With extreme difficulty	No, Impossible
☐	☐	☐	☐	☐

10. Have you had any trouble with washing and drying yourself (all over) because of your knee?

No trouble at all	Very little trouble	Moderate trouble	Extreme trouble	Impossible to do
☐	☐	☐	☐	☐

11. Have you had any trouble getting in and out of a car or using public transport because of your knee?

No trouble at all	Very little trouble	Moderate trouble	Extreme trouble	Impossible to do
☐	☐	☐	☐	☐

12. Could you do the household shopping on your own?

Yes, easily	With little difficulty	With moderate difficulty	With extreme difficulty	No, Impossible
☐	☐	☐	☐	☐

Appendix 1.6: REFLUX Questionnaire

For the questions in section A - F, please put a cross in the box which best describes how often your symptoms have occurred and the effect they have had on your quality of life**.**

SECTION A - HEARTBURN

A1. In the last two weeks, how often have you experienced heartburn (a burning sensation which moves up from your chest to your throat)?

Not at all ☐
Once a week ☐
Two or three times a week ☐
Most days ☐
Everyday ☐

A2. In the last two weeks, how often have you experienced any discomfort or pain in your chest?

Not at all ☐
Once a week ☐
Two or three times a week ☐
Most days ☐
Everyday ☐

A3. In the last two weeks, how much has the heartburn or discomfort/pain in your chest affected your quality of life?

Not at all ☐
A little ☐
Moderately ☐
A lot ☐
Extremely ☐

SECTION B - ACID REFLUX

B1. In the last two weeks, how often have you experienced acid reflux and/or had an acid taste in your mouth?

- Not at all ☐
- Once a week ☐
- Two or three times a week ☐
- Most days ☐
- Everyday ☐

B2. In the last two weeks, how often have you been sick (vomited)?

- Not at all ☐
- Once a week ☐
- Two or three times a week ☐
- Most days ☐
- Everyday ☐

B3. In the last two weeks, how often have you regurgitated (brought up) quantities of liquid or solids into your mouth?

- Not at all ☐
- Once a week ☐
- Two or three times a week ☐
- Most days ☐
- Everyday ☐

B4. In the last two weeks, how often have you experienced a feeling of nausea (without actually being sick or regurgitating)?

Not at all ☐
Once a week ☐
Two or three times a week ☐
Most days ☐
Everyday ☐

B5. In the last two weeks, how often have you wanted to be sick but physically been unable to?

Not at all ☐
Once a week ☐
Two or three times a week ☐
Most days ☐
Everyday ☐

B6. In the last two weeks, how much have these acid reflux symptoms affected your quality of life?

Not at all ☐
A little ☐
Moderately ☐
A lot ☐
Extremely ☐

SECTION C – WIND

C1. In the last two weeks, how often have you experienced a lot of wind from the lower bowel?

Not at all ☐
Once a week ☐
Two or three times a week ☐
Most days ☐
Everyday ☐

C2. In the last two weeks, how often have you experienced a lot of burping/belching?

Not at all ☐
Once a week ☐
Two or three times a week ☐
Most days ☐
Everyday ☐

C3. In the last two weeks, how often have you experienced bloatedness and/or a feeling of trapped wind, in your stomach?

Not at all ☐
Once a week ☐
Two or three times a week ☐
Most days ☐
Everyday ☐

C4. In the last two weeks, how often have you experienced loud gurgling noises from your stomach?

Not at all ☐
Once a week ☐
Two or three times a week ☐
Most days ☐
Everyday ☐

C5. In the last two weeks, how much have these wind problems affected your quality of life?

Not at all ☐
A little ☐
Moderately ☐
A lot ☐
Extremely ☐

SECTION D - EATING AND SWALLOWING

D1. In the last two weeks, how often have you experienced difficulty swallowing food or have you actually choked on food?

Not at all ☐
Once a week ☐
Two or three times a week ☐
Most days ☐
Everyday ☐

D2. In the last two weeks, how often have your eating habits been restricted because of your condition? Examples might be eating more slowly, having smaller portions or eating different foods.

Not at all ☐
Once a week ☐
Two or three times a week ☐
Most days ☐
Everyday ☐

D3. In the last two weeks, how much have these problems with eating affected your quality of life?

Not at all ☐
A little ☐
Moderately ☐
A lot ☐
Extremely ☐

SECTION E – BOWEL MOVEMENTS

E1. In the last two weeks, how often have you experienced diarrhoea and/or loose stools?

Not at all ☐
Once a week ☐
Two or three times a week ☐
Most days ☐
Everyday ☐

E2. In the last two weeks, how often have you experienced constipation and/or hard stools?

- Not at all ☐
- Once a week ☐
- Two or three times a week ☐
- Most days ☐
- Everyday ☐

E3. In the last two weeks, how often have you had a feeling of an urgent need to have a bowel movement?

- Not at all ☐
- Once a week ☐
- Two or three times a week ☐
- Most days ☐
- Everyday ☐

E4. In the last two weeks, how often have you had a feeling of not emptying your bowels?

- Not at all ☐
- Once a week ☐
- Two or three times a week ☐
- Most days ☐
- Everyday ☐

E5. In the last two weeks, how much have these bowel problems affected your quality of life?

Not at all ☐
A little ☐
Moderately ☐
A lot ☐
Extremely ☐

SECTION F – SLEEP

F1. In the last two weeks, how often have you experienced difficulty in lying down to sleep?

Not at all ☐
Once a week ☐
Two or three times a week ☐
Most nights ☐
Every night ☐

F2. In the last two weeks, how often have you experienced difficulty getting to sleep because of your reflux symptoms?

Not at all ☐
Once a week ☐
Two or three times a week ☐
Most nights ☐
Every night ☐

F3. In the last two weeks, how often have you been woken up because of your reflux symptoms?

Not at all ☐
Once a week ☐
Two or three times a week ☐
Most nights ☐
Every night ☐

F4. In the last two weeks, how much have these sleep related problems affected your quality of life?

Not at all ☐
A little ☐
Moderately ☐
A lot ☐
Extremely ☐

SECTION G – WORK, PHYSICAL AND SOCIAL ACTIVITIES

For the following section, please put a cross in the box which best applies to you.

G1. In the last two weeks, have your reflux symptoms affected you at work (paid or voluntary)?

Not applicable (I do not do paid or voluntary work) ☐

No, my symptoms do not affect me ☐

Yes, my symptoms have affected me but I still work ☐

Yes, I have worked less often because of my symptoms ☐

Yes, I have not worked in the last two weeks because of my symptoms ☐

I no longer work because of my symptoms ☐

G2. In the last two weeks, have your reflux symptoms affected your ability to perform less strenuous activities (such as going for a gentle walk, shopping or housework)?

Not applicable (I do not perform these activities, though this is not due to my reflux Symptoms) ☐

No, my symptoms do not affect me ☐

Yes, my symptoms have affected me but I still perform these activities as often as ever ☐

Yes, I perform these activities less often because of my symptoms ☐

Yes, I have not performed these activities in the last two weeks ☐

I no longer perform these activities at all because of my symptoms ☐

G3. In the last two weeks, have your reflux symptoms affected your ability to perform strenuous activities (such as brisk walking or swimming)?

Not applicable (I do not perform these activities, though this is not due to my reflux Symptoms) ☐

No, my symptoms do not affect me ☐

Yes, my symptoms have affected me but I still perform these activities as often as ever ☐

Yes, I perform these activities less often because of my symptoms ☐

Yes, I have not performed these activities in the last two weeks ☐

I no longer perform these activities at all because of my symptoms ☐

G4. In the last two weeks, have you found that your reflux symptoms have affected any of your social activities (such as going out for meals, going out for drinks or socializing with other people)?

Not applicable (I do not perform these activities, though this is not due to my reflux Symptoms) ☐

No, my symptoms do not affect me ☐

Yes, my symptoms have affected me but I still perform these activities as often as ever ☐

Yes, I perform these activities less often because of my symptoms ☐

Yes, I have not performed these activities in the last two weeks ☐

I no longer perform these activities at all because of my symptoms ☐

G5. In the last two weeks, how much has the effect of your reflux symptoms on your work, physical or social activities affected your quality of life?

Not at all ☐

A little ☐

Moderately ☐

A lot ☐

Extremely ☐

Appendix 2.1: Description of trials with imputation of the quality of life outcomes

First Author	Number of participants	Main QoL Outcomes	% missing at final endpoint	Method of Imputation	Imputation primary or sensitivity analysis?	Method of Analysis
Ballard et al.	82	QoL – agitation and cognition	14%	LVCF	Primary	ANCOVA
Berry *et al.*	30	Juniper asthma QoL scale	unclear	1. Worst value if missing due to asthma 2. LVCF if missing not related to asthma	Primary	Paired t-test
Blumenthal *et al.*	134	GHQ – general health Beck depression inventory	7%	LVCF	Primary	GLM
Buszewicz *et al.*	812	SF36	24%	1. Hotdeck at baseline 2. Multiple imputation for follow up	Primary	ANCOVA
Fairbank *et al.*	349	Oswestry disability index SF36	19%	Multiple imputation	Sensitivity	ANCOVA
Feagan *et al.*	181	Inflammatory Bowel Disease Questionnaire (IBDQ)	5%	LVCF	Primary	ANCOVA
Hsieh *et al.*	129	Roland and Morris disability questionnaire	15%	Baseline carried forward	Primary	ANCOVA
Hunkeler *et al.*	1801	SF12	Not clear	Not specified	Sensitivity	t-test

Kaplan *et al.*	879	International Prostate Symptom Scores (IPSS)	8%	LVCF	Primary	ANCOVA
Kennedy *et al.*	149		26%	Based on score changes *	Primary	Generalised Estimating Equations (GEEs)
Korzenik *et al.*	124	Inflammatory Bowel Disease Questionnaire (IBDQ)	Not clear	LVCF	Primary	Stratified rank test
McManus *et al.*	441	State anxiety inventory	1%	1. LVCF 2. Mean imputation	Sensitivity	GLM repeated measures
Meggitt *et al.*	63	Dermatology life quality index (DLQI)	14%	LVCF	Sensitivity	Adjusted regression model
Nair *et al.*	144	Health Status Questionnaire (adaption of SF36)	Not clear	LVCF	Primary	Multiple regression
Petersen, L. *et al.*	547	Global assessment of functioning and symptoms (GAF)	32%	1. LVCF 2. Zero value	Sensitivity	Repeated measures
Petersen, R. *et al.*	769	Alzheimer's Disease Assessment Scale	Not clear	Projection method #	Primary	ANCOVA
Thomas *et al.*	239	SF36 Oswestry Pain disability index McGill present pain index	9%	LVCF	Sensitivity	ANCOVA
Winblad *et al.*	248	Alzheimer's Disease Cooperative Study activities of daily living inventory for severe Alzheimer's disease (ADCS-ADL-Severe)	11%	LVCF	Primary	ANCOVA
Wright *et al.*	109	RAND Physical Function child health questionnaire	7%	Not specified	Primary	ANCOVA

* Impute a score based on changes in other items when at least 75% of those items were present (such as IBS severity scale)

Projection method appropriate for assessing responses among subjects with neurodegenerative disease

Appendix 3.1: KAT - Patient assessed outcome at baseline, 3, 12 and 24 months

	Metal Backed		Non Metal Backed		Patellar Resurfacing		No Patellar Resurfacing		Mobile Bearing		Fixed Bearing	
Oxford knee score	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Baseline	**17.87**	**(7.78)**	**17.35**	**(7.67)**	**18.49**	**(7.39)**	**18.15**	**(7.66)**	**17.18**	**(7.64)**	**16.49**	**(7.35)**
3 month	**31.01**	**(9.93)**	**29.28**	**(9.41)**	**31.19**	**(9.56)**	**30.49**	**(9.45)**	**30.39**	**(9.77)**	**29.37**	**(9.58)**
1 year	**34.66**	**(10.19)**	**32.74**	**(9.75)**	**34.66**	**(9.44)**	**34.53**	**(10.16)**	**33.43**	**(10.45)**	**32.59**	**(10.73)**
2 year	**35.41**	**(10.67)**	**33.26**	**(10.53)**	**35.61**	**(9.83)**	**35.25**	**(10.15)**	**33.61**	**(10.49)**	**32.81**	**(10.38)**
Difference at 1 year (95% CI)		***0.91***	***(-1.13, 2.95)***			***0.02***	***(-1.02, 1.06)***			***-0.25***	***(-2.31, 1.80)***	
Difference at 2 years (95% CI)		***1.33***	***(-0.91, 3.57)***			***0.27***	***(-0.86, 1.39)***			***-0.13***	***(-2.20, 1.93)***	
EQ-5D												
Baseline	**0.40**	**(0.31)**	**0.36**	**(0.32)**	**0.40**	**(0.30)**	**0.39**	**(0.31)**	**0.32**	**(0.32)**	**0.34**	**(0.31)**
3 month	**0.68**	**(0.25)**	**0.64**	**(0.24)**	**0.70**	**(0.23)**	**0.69**	**(0.24)**	**0.66**	**(0.27)**	**0.66**	**(0.24)**
1year	**0.72**	**(0.26)**	**0.69**	**(0.24)**	**0.74**	**(0.23)**	**0.73**	**(0.25)**	**0.72**	**(0.29)**	**0.69**	**(0.29)**
2 year	**0.72**	**(0.26)**	**0.69**	**(0.27)**	**0.74**	**(0.24)**	**0.72**	**(0.27)**	**0.71**	**(0.28)**	**0.67**	**(0.27)**
Difference at 1 year (95% CI)		***0.02***	***(-0.03, 0.07)***			***0.01***	***(-0.02, 0.03)***			***0.02***	***(-0.03, 0.07)***	
Difference at 2 years (95% CI)		***0.01***	***(-0.04, 0.07)***			***0.01***	***(-0.01, 0.04)***			***0.03***	***(-0.02, 0.08)***	
SF-12 Physical component score (PCS)												
Baseline	**30.53**	**(8.08)**	**29.78**	**(7.42)**	**31.07**	**(8.05)**	**31.26**	**(8.50)**	**31.04**	**(8.12)**	**30.39**	**(7.95)**
3 month	**38.90**	**(10.08)**	**37.76**	**(9.23)**	**39.42**	**(9.35)**	**38.68**	**(9.06)**	**38.50**	**(9.52)**	**38.13**	**(9.68)**
1 year	**40.37**	**(11.02)**	**38.00**	**(10.02)**	**40.82**	**(10.51)**	**40.72**	**(10.39)**	**40.36**	**(10.63)**	**38.69**	**(10.80)**
2 year	**40.26**	**(10.88)**	**38.08**	**(10.67)**	**40.66**	**(10.99)**	**40.84**	**(10.39)**	**40.38**	**(11.44)**	**38.70**	**(10.61)**
Difference at 1 year (95% CI		***1.71***	***(-0.38, 3.79)***			***0.14***	***(-0.89, 1.17)***			***0.86***	***(-1.09, 2.80)***	
Difference at 2 years (95% CI		***1.69***	***(-0.56, 3.93)***			***-0.10***	***(-1.19, 1.00)***			***0.81***	***(-1.27, 2.89)***	
SF-12 mental component score (MCS)												
Baseline	**49.14**	**(12.56)**	**49.52**	**(12.23)**	**50.70**	**(11.37)**	**49.73**	**(11.20)**	**48.09**	**(11.97)**	**48.62**	**(11.94)**
3 month	**50.73**	**(11.21)**	**49.96**	**(11.66)**	**51.21**	**(10.60)**	**51.14**	**(10.97)**	**48.19**	**(11.80)**	**49.47**	**(10.99)**
1 year	**51.16**	**(11.57)**	**51.38**	**(10.49)**	**52.31**	**(10.20)**	**51.47**	**(11.10)**	**50.55**	**(11.20)**	**50.14**	**(12.14)**
2 year	**51.44**	**(10.19)**	**50.98**	**(10.19)**	**51.64**	**(9.95)**	**50.87**	**(11.07)**	**49.78**	**(10.76)**	**50.77**	**(11.25)**
Difference at 1 year (95% CI)		***-0.11***	***(-2.10, 1.88)***			***0.54***	***(-0.47, 1.56)***			***0.05***	***(-1.90, 1.99)***	
Difference at 2 years (95% CI)		***0.47***	***(-1,55 2.50)***			***0.29***	***(-0.74, 1.31)***			***-0.85***	***(-2.76, 1.06)***	

Appendix 4.1: Syntax for Little's test of MCAR

The syntax below represents an example for the calculation of Little's test statistic using SAS version 9.1 {{3513 SAS Institute Inc. 2004; }}.

Firstly the directory in which the data and any formatting files can be found must be provided.

```
LIBNAME PERM 'T:\People\s.fielding\CSO fellowship\Data\KAT';
libname library 'T:\People\s.fielding\CSO fellowship\Data\KAT';
```

Secondly the data file needed must be obtained from the permanent SAS datasets and here is saved as a temporary file 'temp1'. In this example the EQ5D scores measured at baseline, three months, one year and two years are renamed to score1 to score4 to enable the generic program to be used easily.

```
data temp1;
set perm.katdata;
rename eq5d_0=score1 eq5d_3m=score2 eq5d_1y=score3 eq5d_2y=score4;
run;
```

The pattern specific means and pooled estimates of the covariance are calculated using PROC MI.

```
ods output MissPattern=pattern EMestimates=estimate;
proc mi data=temp1 Nimpute=1;
EM outem=outem;
var score:;
run;
```

The next stage is to calculate the matrices with the pattern specific means and pooled estimates using PROC IML. The results from the output datasets created in the last step are read into matrices for the next step.

```
proc iml;
use work.estimate;
score={score1 score2 score3 score4};
read all where(_Type_="MEAN") var score into mean[colname=varname] ;
read all where (_Type_="COV") var score into
                covar[rowname=_Name_ colname=varname];
print "pooled mean", mean;
print "pooled covariance", covar;
```

```
use work.pattern;
read all var score into patmean;
read all var {Freq} into M;
print 'observed means for pattern', M patmean [format=5.2];
```

The next piece of code combines all this information to calculated Little's test statistic

```
stat=0; df=0;

do p=1 to Nrow(patmean);
        patp = ((patmean[p,])>0);     * indicator of non-missing value *
        locp=loc(patp>0);                    *  location of non-missing value *
        jp=sum(patp);                        *  Number of non-missing values *
if jp^=0 then do;
        dp=diag(patp)[locp,];
        vp=dp*covar*dp`;                     *  covariance for pattern *
        rp=mean[,locp]-patmean[p,locp];   *  difference in means *
        stat=stat+M[p]*(rp*inv(vp)*rp`);  *  sums statistic *
        df=df+jp;                            *  sums degrees of freedom *
        end;
end;

df=df-ncol(mean);                                * adjusts DF *
pvalue=1-probchi(stat,df);                * calculates p-value *

print stat df pvalue;                  * prints the test-statistic, DF and
                                                    p-value *
quit;
```

This code can easily be adapted by changing the input file in the first step and remembering to change the definition of the 'score' matrix in the PROC IML.

Appendix 4.2: Syntax for the Listing and Schlittgen test

The syntax below represents an example for the calculation of Listing and Schlittgen's test if dropouts are random {{3268 Listing,J. 1998; }} using the statistical software STATA {{3564 StataCorp 2007; }}. This program can be adapted to calculate the test statistic for any number of assessments.

First set the working directory and open the required data file.

```
cd " T:\People\s.fielding\CSO fellowship\Data\KAT\"
use KATdata, clear
```

Rename the EQ5D scores at each assessment to v1-v4 for use later in the program

```
rename eq5d_0 v1
rename eq5d_3m v2
rename eq5d _1y v3
rename eq5d_2y v4
```

Generate the variable pattern which indicates which patients have a monotone missing data pattern.

```
gen pattern=.
replace pattern=1 if (v1!=. & v2!=. & v3!=. & v4!=.)
replace pattern=2 if (v1!=. & v2!=. & v3!=. & v4==.)
replace pattern=3 if (v1!=. & v2!=. & v3==. & v4==.)
replace pattern=4 if (v1!=. & v2==. & v3==. & v4==.)
replace pattern=5 if (v1==. & v2==. & v3==. & v4==.)
tab pattern, missing
```

Select only those patients with monotone missing pattern

```
keep if (pattern==1 | pattern==2 | pattern==3 | pattern==4)
```

Next generate the mean QoL score at each assessment for each of the patterns.

```
foreach x of numlist 1 2 3 4 {
sum v1 if pattern==`x'
local N`x' = r(N)
local mean1`x' = r(mean)
```

```
display `N`x''
display `mean1`x''
}

foreach x of numlist 1 2 3 {
sum v2 if pattern==`x'
local mean2`x' = r(mean)
display `mean2`x''
}

foreach x of numlist 1 2 {
sum v3 if pattern==`x'
local mean3`x' = r(mean)
display `mean3`x''
}

sum v4 if pattern==1
local mean41 = r(mean)
display `mean41'

local m = `N2'+`N3'+`N4'
```

Next D is calculated using information from the above steps

```
local D = (1/(`m'))*(`N4'*(`mean11'-`mean14') + `N3'*(`mean21'-
`mean23')+`N2'*(`mean31'-`mean32'))
display `D'
```

The variance of D is needed and first some other quantities are calculated and the correlations estimated.

```
matrix accum R = v1 v2 v3 if pattern==1, nocons dev
matrix R = corr(R)
matrix m = (`N2',`N3',`N4')
matrix V = m*R*m'
matrix list V
local V = V[1,1]

corr v1 v2 if pattern==1, cov
local s11 = r(Var_1)
local s22 = r(Var_2)
corr v1 v3 if pattern==1, cov
local s33 = r(Var_2)

local sigma2 = (`s11'+`s22'+`s33')/3
```

```
local varD = `sigma2'/`m' + (`sigma2'/(`m'*`N1'^2))*`V'
```

The test statistic S is calculated using the estimate of D and variance of D. The p-value is obtained.

```
local S = `D'/sqrt(`varD')
local p =normalden(`S')
```

Finally each of the quantities D, var(D), S and the p-value are displayed.

```
display `D'
display `varD'
display `S'
display `p'
```

Appendix 5.1: REFLUX - Mean (SD) QoL scores by missing data pattern

Pattern	N	Baseline	3 months	12 months
		EQ5D		
1	281	0.73 (0.25)	0.74 (0.27)	0.74 (0.25)
2	12	0.65 (0.26)	0.86 (0.15)	-
3	28	0.69 (0.27)	-	0.72 (0.30)
4	23	0.65 (0.33)	-	-
5	5	-	0.55 (0.27)	0.45 (0.33)
6	4	-	0.37 (0.45)	-
7	2	-	-	0.86 (0.19)
8	2	-	-	-
	Total	0.72 (0.26)	0.74 (0.27)	0.73 (0.26)
		SF12 physical component score		
1	268	45.3 (9.3)	47.1 (9.9)	46.9 (10.1)
2	15	46.0 (9.5)	44.9 (9.8)	
3	31	44.1 (10.9)		44.1 (9.5)
4	22	45.0 (9.1)		
5	7		45.4 (8.5)	45.3 (6.5)
6	2		38.0 (23.5)	
7	5			40.0 (11.6)
8	7			
	Total	45.2 (9.4)	46.9 (9.9)	46.5 (10.0)
		SF12 mental component score		
1	268	46.2 (11.2)	45.9 (12.8)	46.4 (12.5)
2	15	42.8 (12.3)	46.2 (14.2)	
3	31	41.1 (15.2)		40.7 (15.5)
4	22	42.3 (12.7)		
5	7		48.4 (8.8)	50.6 (9.7)
6	2		36.9 (23.5)	
7	5			38.1 (16.6)
8	7			
	Total	45.3 (11.8)	45.9 (12.8)	45.8 (13.0)
		RQLS		
1	239	65.6 (24.4)	76.8 (23.9)	78.7 (21.8)
2	26	68.1 (20.1)	79.6 (19.2)	
3	37	62.2 (27.4)		80.2 (21.8)
4	25	62.8 (24.3)		
5	18		79.1 (19.5)	81.9 (14.3)
6	3		70.8 (12.2)	
7	5			60.0 (27.4)
8	4			
	Total	65.2 (24.4)	77.1 (23.1)	78.8 (21.6)

Appendix 5.2: MAVIS - Mean (SD) QoL scores by missing data pattern

Pattern	N	Baseline	6 months	12 months
EQ5D				
1	825	0.77 (0.21)	0.79 (0.21)	0.78 (0.21)
2	28	0.65 (0.25)	0.66 (0.30)	
3	4	0.82 (0.13)		
4	51	0.70 (0.25)		0.85 (0.20)
5	1		0.52 (-)	0.52 (-)
6	1		0.66 (-)	
	Total	0.77 (0.22)	0.78 (0.21)	0.78 (0.21)
SF12 physical component score				
1	814	43.6 (10.9)	44.4 (10.4)	44.0 (10.8)
2	32	41.4 (12.2)	38.7 (12.1)	
3	9	39.1 (10.4)		43.8 (8.0)
4	51	39.4 (11.6)		
5	4		43.7 (16.1)	40.2 (10.0)
	Total	43.3 (11.0)	44.2 (10.5)	44.0 (10.8)
SF12 mental component score				
1	814	54.0 (8.6)	53.6 (8.7)	53.5 (9.1)
2	32	50.3 (9.4)	45.7 (11.4)	
3	9	47.6 (9.2)		51.0 (11.2)
4	51	52.4 (9.0)		
5	4		44.4 (9.0)	48.5 (8.2)
	Total	53.7 (8.7)	53.3 (9.0)	53.4 (9.2)

Appendix 5.3: RECORD - Mean (SD) QoL scores by missing data pattern

Pattern	N	4 months	12 months	24 months
EQ5D				
1	2606	0.74 (0.23)	0.75 (0.24)	0.74 (0.26)
2	545	0.67 (0.27)	0.66 (0.78)	
3	273	0.72 (0.25)		0.72 (0.25)
4	483	0.63 (0.30)		
5	240		0.73 (0.25)	0.70 (0.27)
6	97		0.66 (0.26)	
7	85			0.67 (0.28)
8	963			
	Total	0.72 (0.25)	0.73 (0.25)	0.73 (0.26)
SF12 physical component score				
1	2406	41.8 (10.7)	42.1 (10.8)	41.5 (11.1)
2	475	38.9 (11.2)	39.3 (11.4)	
3	289	40.2 (11.6)		38.8 (11.2)
4	474	37.6 (11.9)		
5	348		41.6 (10.5)	40.6 (11.0)
6	139		36.9 (11.8)	
7	106			41.4 (10.2)
8	1055			
	Total	40.8 (11.1)	41.4 (11.0)	41.2 (11.1)
SF12 mental component score				
1	2406	51.3 (10.0)	51.1 (9.9)	50.7 (10.2)
2	475	49.4 (10.2)	47.9 (11.8)	
3	289	49.6 (11.1)		49.3 (10.9)
4	474	46.0 (11.4)		
5	348		49.5 (10.0)	50.1 (10.8)
6	139		45.9 (11.6)	
7	106			47.8 (11.1)
8	1055			
	Total	50.2 (10.5)	50.3 (10.4)	50.4 (10.4)

Appendix 5.4: RECORD - Mean (SD) QoL of continuers and dropouts (scenario two)

		Continuers		Drop outs			
Assessment	Total N	N (%)	Mean (SD)	N (%)	Mean (SD)	Diff (95% CI)	p-value
EQ5D							
4 months	3907	3424 (88)	0.73 (0.24)	483 (12)	0.63 (0.30)	0.10 (0.08,0.12)	<0.001
12 months	3488	2846 (82)	0.75 (0.24)	642 (18)	0.66 (0.27)	0.08 (0.06,0.11)	<0.001
SF12 physical component score							
4 months	3644	3170 (87)	41.2 (10.0)	474 (13)	37.6 (11.9)	3.7 (2.5,4.8)	<0.001
12 months	3368	2754 (82)	42.0 (10.7)	614 (18)	38.7 (11.5)	3.3 (2.3,4.3)	<0.001
SF12 mental component score							
4 months	3644	3170 (87)	50.9 (10.2)	474 (13)	46.0 (11.4)	4.9 (3.8,6.0)	<0.001
12 months	3368	2754 (82)	50.9 (9.9)	614 (18)	47.4 (11.8).	3.45 (2.5,4.5)	<0.001

Appendix 5.5: RECORD - Mean (SD) QoL for responders and subsequent dropout

		Continuers		Drop outs		Unadjusted OR	
Assessment	Total N	N	Mean (SD)	N	Mean (SD)	(95% CI)	p-value
EQ5D							
4 months	2606	2362	0.76 (0.24)	244	0.70 (0.25)	0.44 (0.26,0.75)	0.002
12 months	2059	1804	0.75 (0.25)	255	0.70 (0.27)	0.49 (0.31,0.79)	0.003
SF12 physical component score							
4 months	2406	2200	42.1 (10.7)	206	39.4 (10.8)	0.98 (0.96,0.99)	0.011
12 months	1908	1666	42.8 (10.6)	242	41.1 (10.9)	0.98 (0.97,0.99)	0.019
SF12 mental component score							
4 months	2406	2200	51.6 (9.9)	206	48.4 (10.6)	0.97 (0.96,0.98)	<0.001
12 months	1908	1666	51.8 (9.3)	242	49.7 (10.8)	0.98 (0.96,0.99)	0.001

Appendix 5.6: KAT - Mean (SD) QoL scores by missing data pattern

Pattern	N	Baseline	3 months	1 year	2 years
EQ5D					
1	1621	0.40 (0.30)	0.70 (0.23)	0.74 (0.24)	0.72 (0.27)
2	180	0.33 (0.32)	0.62 (0.30)	0.64 (0.31)	
3	67	0.42 (0.30)	0.66 (0.23)		0.68 (0.25)
4	88	0.30 (0.31)	0.59 (0.29)		
5	102	0.40 (0.31)		0.76 (0.22)	0.78 (0.19)
6	24	0.24 (0.33)		0.57 (0.29)	
7	19	0.29 (0.37)			0.68 (0.27)
8	94	0.38 (0.31)			
9	43		0.67 (0.31)	0.66 (0.32)	0.70 (0.32)
10	5		0.57 (0.08)	0.44 (0.33)	
11	3		0.65 (0.37)		0.54 (0.24)
12	1		0.80 (-)		
13	11			0.61 (0.31)	0.61 (0.30)
14	4			0.49 (0.53)	
15	19				0.81 (0.24)
16	75				
Total	2356	0.39 (0.31)	0.68 (0.24)	0.73 (0.25)	0.72 (0.26)
SF12 Physical component score					
1	1522	50.7 (11.3)	51.6 (10.6)	52.3 (10.3)	51.4 (10.4)
2	170	47.0 (11.5)	48.8 (12.0)	48.0 (12.4)	
3	83	46.5 (11.2)	48.2 (12.0)		48.0 (11.7)
4	90	46.8 (12.6)	44.3 (12.1)		
5	126	50.3 (11.9)		50.3 (12.4)	49.8 (11.3)
6	43	46.7 (11.3)		46.2 (12.1)	
7	29	46.6 (12.8)			47.8 (10.5)
8	99	46.9 (10.7)			
9	58		48.5 (10.7)	48.6 (10.8)	49.0 (10.8)
10	14		41.6 (13.3)	42.9 (12.1)	
11	3		33.7 (14.4)		29.0 (12.3)
12	4		56.0 (12.7)		
13	16			53.1 (10.9)	54.5 (9.2)
14	5			42.4 (17.4)	
15	19				52.0 (11.9)
16	75				
Total	2356	49.8 (11.5)	50.7 (11.0)	51.5 (10.9)	51.0 (10.6)
SF12 Mental component score					
1	1522	31.4 (8.2)	39.3 (9.1)	40.8 (10.4)	40.7 (10.7)
2	170	30.1 (8.3)	37.1 (10.3)	37.2 (11.6)	
3	83	31.0 (7.8)	36.5 (9.7)		36.7 (11.2)
4	90	28.6 (8.0)	36.2 (8.6)		
5	126	31.2 (8.5)		41.0 (11.4)	40.3 (11.1)
6	43	30.4 (6.5)		35.7 (10.4)	
7	29	30.7 (9.3)			37.1 (12.0)
8	99	30.6 (8.7)			
9	58		38.9 (10.3)	40.8 (10.4)	38.3 (10.1)
10	14		35.0 (7.6)	33.9 (9.1)	
11	3		33.2 (8.8)		27.9 (4.1)
12	4		39.0 (12.2)		

13	16			37.7 (7.2)	36.1 (11.1)
14	5			35.2 (8.5)	
15	19				41.0 (9.9)
16	75				
Total	2356	31.1 (8.2)	38.8 (9.3)	40.3 (10.6)	40.3 (10.8)
Oxford Knee Score					
1	1158	18.6 (734)	31.0 (9.3)	34.3 (10.0)	34.5 (10.4)
2	286	17.3 (7.4)	30.0 (9.8)	33.0 (10.1)	
3	130	17.3 (7.6)	30.1 (10.1)		34.9 (10.1)
4	167	16.5 (7.7)	29.2 (10.3)		
5	169	18.2 (7.7)		36.0 (9.3)	36.6 (9.7)
6	61	17.4 (7.6)		32.6 (11.2)	
7	69	18.1 (7.8)			36.6 (10.1)
8	152	17.2 (8.2)			
9	32		30.2 (10.4)	35.1 (10.4)	35.3 (9.9)
10	11		27.2 (9.3)	30.6 (11.7)	
11	6		28.3 (12.0)		26.3 (12.9)
12	8		31.4 (13.0)		
13	14			30.6 (10.1)	34.1 (7.9)
14	3			25.3 (18.0)	
15	16				37.6 (10.5)
16	74				
Total	2356	18.0 (7.6)	30.5 (9.6)	34.1 (10.1)	34.9 (10.3)

Appendix 5.7: KAT - Fairclough logistic regression results (scenario one)

Assessment	Complete N (%)	Missing N (%)	Adjusted OR for previous QoL OR(95% CI)	p-value
EQ5D				
3 months	1978 (76)	558 (24)	0.66 (0.47,0.92)[1]	0.015
1 year	1701 (72)	655 (28)	0.24 (0.17,0.33)[2]	<0.001
2 years	1560 (66)	796 (34)	0.15 (0.11,0.22)[3]	<0.001
SF12 physical component score				
3 months	1744 (74)	612 (26)	0.99 (0.98,1.01)	0.36
1 year	1663 (71)	693 (29)	0.96 (0.95,0.97)[2]	<0.001
2 years	1541 (65)	815 (35)	0.96 (0.95,0.97)[3]	<0.001
SF12 mental component score				
3 months	1744 (74)	612 (26)	0.987 (0.979,0.996)	0.004
1 year	1663 (71)	693 (29)	0.98 (0.97, 0.99)[2]	<0.001
2 years	1541 (65)	815 (35)	0.97 (0.96,0.98)[3]	<0.001
Oxford Knee score				
3 months	1608 (68)	748 (32)	0.99 (0.98,1.00)[1]	0.093
1 year	1477 (63)	879 (37)	0.96 (0.95,0.97)	<0.001
2 years	1312 (56)	1044 (44)	0.957 (0.95,0.96)	<0.001

Adjusted for: [1] any readmissions; [2] type of knee arthritis and post-operative complications; [3] ASA grade.

Appendix 5.8: KAT Fairclough logistic regression results (scenario two)

Assessment	Complete N (%)	Missing N (%)	Adjusted OR for previous QoL OR(95% CI)	p-value
EQ5D				
3 months	2195 (93)	180 (8)	0.78 (0.47,1.28)[1]	0.32
1 year	2195 (93)	181 (8)	0.57 (0.34,0.96)[2]	0.034
2 years	2247 (95)	314 (14)	0.17 (0.11,0.25)[3]	<0.001
SF12 physical component score				
3 months	2162 (92)	181 (8)	1.00 (0.98,1.02)[1]	0.712
1 year	2162 (92)	205(9)	0.976 (0.956,0.996)[2]	0.02
2 years	2241 (95)	314 (14)	0.96 (0.94,0.97)[3]	<0.001
SF12 mental component score				
3 months	2162 (92)	181 (8)	0.983 (0.976,0.995)[1]	0.010
1 year	2162 (92)	205(9)	0.98 (0.96,0.99)[2]	<0.001
2 years	2241 (95)	314 (14)	0.96 (0.95,0.98)[3]	<0.001
Oxford Knee Score				
3 months	2192 (93)	187 (9)	0.99 (0.97,1.01)	0.32
1 year	2192 (93)	206 (9)	0.96 (0.94,0.98)	<0.001
2 years	2249 (95)	321 (14)	0.95 (0.94,0.97)	<0.001

Adjusted for: [1] extent of knee arthritis and readmissions; [2] extent of knee arthritis, post-operative complications, readmissions and ASA grade; [3] ASA grade and type of knee arthritis.

Appendix 5.9: PRISM - Mean (SD) QoL score by missing data pattern

Pattern	N	Baseline	1 year	2 years	3 years	4 years
			EQ5D			
1	55	0.64 (0.27)	0.63 (0.27)	0.64 (0.27)	0.61 (0.29)	0.63 (0.31)
2	366	0.61 (0.30)	0.61 (0.29)	0.59 (0.31)	0.59 (0.30)	
3	3	0.40 (0.37)	0.35 (0.37)	0.26 (0.44)		0.35 (0.37)
4	367	0.60 (0.29)	0.60 (0.28)	0.59 (0.29)		
5	3	0.70 (0.11)	0.68 (0.08)		0.67 (0.07)	0.67 (0.11)
6	64	0.54 (0.35)	0.54 (0.35)		0.58 (0.35)	
7	1	0.73 (-)	0.69 (-)			0.69 (-)
8	175	0.61 (0.29)	0.56 (0.31)			
9	2	0.64 (0.07)		0.25 (0.38)	0.29 (0.43)	0.34 (0.50)
10	32	0.59 (0.31)		0.63 (0.27)	0.62 (0.30)	
11	20	0.42 (0.36)		0.33 (0.41)		
12	1	0.73 (-)			0.36 (-)	0.69 (-)
13	6	0.46 (0.26)			0.31 (0.25)	
14	155	0.47 (0.33)				
15	2		0.90 (0.03)	1.00 (0)	1.00 (-)	0.94 (0.08)
16	23		0.68 (0.27)	0.68 (0.26)	0.65 (0.29)	
17	16		0.63 (0.35)	0.61 (0.40)		
18	4		0.62 (0.22)		0.54 (0.37)	
19	11		0.53 (0.34)			
20	2			0.31 (0.54)	0.31 (0.54)	
21	1			0.62 (-)		
22	15					
Total	1324	0.58 (0.30)	0.59 (0.29)	0.59 (0.31)	0.59 (0.31)	0.62 (0.31)
			SF36 Physical component score			
1	51	41.0 (12.3)	38.5 (12.1)	38.8 (12.2)	38.6 (12.0)	38.4 (11.5)
2	306	37.8 (11.7)	36.0 (11.0)	36.4 (10.8)	36.2 (10.4)	
3	3	31.1 (7.1)	29.6 (6.0)	27.1 (5.8)		29.3 (4.7)
4	332	37.3 (11.1)	36.0 (10.7)	36.5 (11.1)		
5	37	38.2 (10.2)	36.8 (11.1)		35.6 (11.2)	
6	1	43.2 (-)	43.7 (-)			41.9 (-)
7	154	35.3 (11.6)	34.4 (10.0)			
8	3	39.0 (2.5)		35.9 (1.5)	35.3 (9.5)	34.5 (8.1)
9	38	35.2 (11.8)		34.7 (11.4)	34.6 (11.5)	
10	65	32.6 (9.7)		32.6 (10.1)		
11	2	31.7 (9.6)			24.0 (0.4)	23.7 (3.6)
12	25	30.6 (10.1)			29.2 (8.9)	
13	1	33.4 (-)				32.5 (-)
14	180	33.3 (10.4)				
15	1		27.7 (-)	17.9 (-)	31.9 (-)	29.7 (-)
16	28		37.7 (12.5)	34.6 (12.3)	35.1 (11.7)	
17	24		36.4 (11.2)	35.5 (11.4)		
18	4		32.9 (11.0)		31.7 (11.7)	
19	1		24.8 (-)			34.2 (-)
20	11		34.8 (9.8)			
21	5			45.8 (7.6)	42.2 (12.1)	
22	10			32.6 (8.9)		
23	4				28.1 (6.9)	
24	38					
Total	1324	36.3 (11.3)	35.9 (10.8)	36.1 (11.0)	35.8 (10.8)	37.1 (11.0)
			SF36 mental component score			
1	51	50.3 (21.1)	50.2 (10.4)	48.7 (11.0)	48.0 (10.9)	49.0 (11.4)
2	306	50.5 (11.1)	49.1 (11.4)	48.9 (11.0)	47.8 (12.1)	

3	3	30.8 (6.7)	37.0 (3.7)	34.2 (1.3)		31.3 (7.0)
4	332	49.7 (11.0)	47.3 (11.9)	46.7 (12.3)		
5	37	50.5 (11.3)	49.4 (10.8)		47.9 (11.1)	
6	1	47.1 (-)	47.5 (-)			37.1 (-)
7	154	48.5 (12.0)	43.6 (13.2)			
8	3	50.8 (10.8)		42.0 (11.9)	39.9 (0.57)	42.3 (6.3)
9	38	46.9 (14.2)		45.6 (14.4)	45.8 (13.2)	
10	65	44.3 (12.4)		42.6 (12.9)		
11	2	57.0 (6.1)			42.9 (5.1)	50.0 (7.7)
12	25	42.6 (12.0)			34.4 (11.6)	
13	1	35.7 (-)				28.7 (-)
14	180	46.0 (12.5)				
15	1		45.5 (-)	47.4 (-)	43.1 (-)	54.4 (-)
16	28		46.8 (12.8)	46.5 (12.1)	47.1 (14.4)	
17	24		45.1 (10.2)	45.1 (11.4)		
18	4		43.0 (8.3)		53.6 (17.4)	
19	1		38.0 (-)			48.8 (-)
20	11		47.5 (14.5)			
21	5			41.9 (14.4)	45.7 (19.2)	
22	10			47.4 (14.3)		
23	4				32.1 (8.5)	
24	38					
Total	1324	48.7 (11.8)	47.4 (12.0)	47.1 (12.0)	46.8 (12.5)	47.4 (11.5)
			Arthritis Index			
1	51	41.2 (13.6)	39.2 (13.6)	39.5 (12.8)	39.0 (12.5)	38.5 (12.2)
2	306	38.0 (12.7)	36.3 (11.9)	36.5 (12.3)	35.8 (11.8)	
3	3	27.7 (9.0)	24.4 (7.8)	22.5 (8.9)		23.8 (7.2)
4	332	36.9 (12.4)	36.1 (11.6)	36.3 (12.6)		
5	37	38.1 (11.9)	37.8 (12.9)		35.8 (10.9)	
6	1	42.0 (-)	45.1 (-)			38.5 (-)
7	154	35.0 (13.0)	33.6 (11.8)			
8	3	39.3 (5.6)		34.4 (3.0)	31.8 (9.2)	32.0 (-)
9	38	33.4 (12.0)		33.8 (12.9)	33.2 (13.0)	
10	65	30.1 (10.8)		31.0 (10.8)		
11	2	37.8 (12.7)			28.6 (36.2)	27.5 (4.0)
12	25	27.9 (9.1)			26.4 (11.3)	
13	1	27.8 (-)				21.9 (-)
14	180	32.3 (11.6)				
15	1		28.3 (-)	21.4 (-)	30.7 (-)	33.3 (-)
16	28		37.6 (14.0)	34.8 (14.1)	24.8 (14.5)	
17	24		36.4 (11.1)	34.2 (12.3)		
18	4		31.1 (10.9)		30.4 (10.7)	
19	1		21.7 (-)			46.4 (-)
20	11		35.2 (13.0)			
21	5			40.3 (13.0)	40.5 (19.4)	
22	10			33.1 (11.1)		
23	4				21.7 (8.7)	
4	38					
Total	1324	35.8 (12.6)	35.9 (12.0)	35.9 (12.5)	35.2 (12.3)	36.9 (12.0)

Appendix 5.10: PRISM - Results for Ridout's logistic regression (scenario two)

		Continuers		Drop outs		Adjusted OR	
Assessment	Total N	N	Mean (SD)	N	Mean (SD)	(95% CI)	p-value
EQ5D							
Baseline	1250	1095	0.60 (0.30)	155	0.47 (0.33)	0.31 (0.18,0.52)[1]	<0.001
Year 1	1090	904	0.60 (0.29)	186	0.56 (0.31)	(0.58 (0.34,0.98)[2]	0.041
Year 2	889	485	0.59 (0.30)	404	0.57 (0.31)	(0.79 (0.51,1.22)[3]	0.29
Year 3	560	63	0.61 (0.29)	497	0.59 (0.31)	0.87 (0.30,2.51)[4]	0.8
SF36 Physical component score							
Baseline	1198	1018	36.8 (11.4)	180	33.1 (10.2)	0.97 (0.96,0.99)[5]	0.001
Year 1	953	788	36.2 (11.0)	165	34.4 (9.9)	0.99 (0.96,1.01)[6]	0.3
Year 2	866	435	36.4 (11.1)	431	35.8 (11.0)	1.00 (0.98,1.01)[7]	0.5
Year 3	504	57	37.8 (11.8)	447	35.5 (10.7)	0.98 (0.95,1.02)[8]	0.37
SF36 Mental component score							
Baseline	1198	1018	49.1 (11.6)	180	46.0 (12.5)	0.98 (0.97,0.99)[5]	0.004
Year 1	953	788	48.1 (11.5)	165	43.9 (13.3)	0.96 (0.94,0.98)[6]	0.001
Year 2	866	435	48.2 (11.5)	431	46.0 (12.4)	0.98 (0.97,0.99)[7]	0.006
Year 3	504	57	47.3 (10.5)	447	46.7 (12.7)	0.99 (0.97,1.02)[8]	0.75
Arthritis Index							
Baseline	1198	1018	36.4 (12.6)	180	32.3 (11.6)	0.98 (0.96,0.99)[5]	<0.001
Year 1	953	788	36.4 (12.0)	165	33.7 (11.9)	0.98 (0.96,1.00)[6]	0.08
Year 2	866	435	36.4 (12.6)	431	35.3 (12.4)	0.99 (0.98,1.00)[7]	0.24
Year 3	504	57	38.1 (12.2)	447	34.8 (12.3)	0.98 (0.95,1.01)[8]	0.16

Adjusted for: [1] marital status and existence of siblings; [2] marital status and pelvic bone involved with Paget's; [3] existence of siblings; [4] previous bone scan, previous femoral fractures, clinician reported bone pain; [5] marital status; [6] previous vertebral factures; [7] bone scan; [8] bone scan and previous femoral fractures.

Appendix 5.11: PRISM - Mean (SD) QoL scores for monotone missingness (scenario three)

Pattern	N	Baseline	3 months	12 months	Test Statistic	p-value
EQ5D						
1	637	0.61 (0.29)	0.61 (0.28)	0.60 (0.29)	S=1.36	0.16
2	115	0.60 (0.31)	0.57 (0.31)			
3	16	0.59 (0.25)				
SF36 physical component score						
1	562	37.6 (11.5)	36.0 (10.9)	36.3 (11.1)	S=-0.27	0.38
2	96	38.1 (11.3)	36.4 (11.4)			
3	14	37.1 (9.5)				
SF36 mental component score						
1	562	50.2 (11.1)	48.5 (11.4)	48.2 (11.5)	S=1.31	0.17
2	96	49.2 (11.1)	47.1 (12.3)			
3	14	48.8 (13.2)				
Arthritis Index						
1	562	37.6 (12.7)	36.3 (11.9)	36.4 (12.6)	S=0.22	0.39
2	96	38.2 (12.4)	36.3 (12.4)			
3	14	36.2 (12.3)				

Appendix 5.12: TOMBOLA – Mean (SD) EQ5D scores by missing data pattern

Pattern	N	Baseline	12 months	18 months	24 months	30 months
1	1422	0.63 (0.13)	0.63 (0.15)	0.63 (0.14)	0.64 (0.15)	0.63 (0.14)
2	188	0.63 (0.14)	0.63 (0.15)	0.62 (0.17)	0.63 (0.15)	
3	103	0.63 (0.12)	0.62 (0.13)	0.61 (0.14)		0.62 (0.15)
4	162	0.61 (0.14)	0.61 (0.16)	0.62 (0.16)		
5	89	0.62 (0.12)	0.64 (0.11)		0.63 (0.14)	0.63 (0.14)
6	66	0.61 (0.14)	0.59 (0.16)		0.60 (0.17)	
7	49	0.63 (0.14)	0.59 (0.14)			0.60 (0.16)
8	215	0.63 (0.14)	0.62 (0.14)			
9	71	0.62 (0.15)		0.62 (0.18)	0.60 (0.20)	0.61 (0.20)
10	45	0.60 (0.18)		0.61 (0.18)	0.62 (0.14)	
11	20	0.64 (0.12)		0.66 (0.10)		0.61 (0.14)
12	63	0.62 (0.12)		0.61 (0.16)		
13	33	0.61 (0.18)			0.61 (0.13)	0.62 (0.16)
14	53	0.62 (0.11)			0.60 (0.16)	
15	38	0.63 (0.12)				0.58 (0.17)
16	683	0.60 (0.15)				
17	25		0.62 (0.18)	0.59 (0.20)	0.60 (0.16)	0.57 (0.23)
18	3		0.62 (0.20)	0.66 (0.12)	0.63 (0.18)	
19	1		0.42 (-)	0.46 (-)		0.46 (-)
20	2		0.59 (0.19)		0.59 (0.19)	0.63 (0.14)
21	1		0.46 (-)		0.58 (-)	
22	1		0.58 (-)			0.58 (-)
23	10		0.64 (0.17)			
24	2			0.54 (0.001)	0.48 (0.09)	0.45 (0.04)
25	3				0.68 (0.09)	0.57 (0.16)
26	2				0.64 (0.13)	
27	1					0.58 (-)
28	48					
	Total	0.62 (0.14)	0.63 (0.14)	0.63 (0.15)	0.63 (0.14)	0.63 (0.15)

Appendix 5.13: TOMBOLA - results of Ridout logistic regression for scenario three

	Continuers		Drop outs		Unadjusted OR	
Assessment	N	Mean (SD)	N	Mean (SD)	(95% CI)	p-value
Baseline	428	0.64 (0.14)	491	0.62 (0.13)	0.61 (0.23,1.62)	0.32
6 weeks	428	0.65 (0.14)	350	0.66 (0.13)	1.46 (0.51,4.18)	0.48
12 months	481	0.63 (0.15)	268	0.64 (0.14)	1.46 (0.52, 4.07)	0.47
18 months	545	0.64 (0.14)	210	0.64 (0.15)	1.01 (0.34,3.03)	0.98
24 months	624	0.64 (0.14)	126	0.64 (0.14)	0.65 (0.16,2.59)	0.54

Appendix 5.14: NPC trial - Mean (SD) QoL by pattern of missing data

Pattern	N	Baseline	1 month	2 months	3 months	4 months
			Pain			
1	114	42.8 (33.7)	33.5 (31.6)	36.3 (31.8)	35.4 (32.6)	38.2 (33.5)
2	34	37.3 (35.5)	28.4 (30.6)	38.7 (33.3)	42.6 (33.1)	
3	10	38.3 (22.3)	26.7 (21.1)	30.0 (25.8)		48.3 (31.9)
4	33	44.9 (40.1)	31.8 (32.4)	40.9 (30.6)		
5	7	31.0 (31.0)	31.0 (20.2)		26.2 (31.7)	35.7 (31.1)
6	2	8.3 (11.8)	8.3 (11.8)		0 (0)	
7	69	46.4 (34.5)	49.0 (32.6)			
8	5	50 (37.3)		43.3 (27.9)	60.0 (25.3)	43.3 (19.0)
9	3	27.8 (25.5)		16.7 (28.9)	22.2 (38.5)	
10	1	16.7 (-)		66.7 (-)		
11	155	56.0 (36.1)				
12	1					
Total	434	47.3 (38.5)	36.1 (32.0)	37.2 (31.3)	36.6 (32.8)	39.0 (32.7)
			Physical functioning			
1	112	57.7 (27.5)	57.9 (28.0)	56.3 (30.0)	58.5 (31.2)	53.0 (33.5)
2	35	54.9 (38.3)	58.0 (33.2)	50.0 (31.4)	34.9 (31.9)	
3	10	56.0 (15.8)	62.0 (23.9)	62.0 (29.0)		46.0 (31.3)
4	33	47.3 (29.9)	39.2 (27.4)	35.8 (31.5)		
5	8	40 (28.3)	49.4 (27.6)		52.5 (26.0)	47.5 (28.2)
6	1	100 (-)	40.0 (-)		40.0 (-)	
7	69	44.5 (30.5)	27.0 (28.0)			
8	5	52 (41.5)		48.0 (30.3)	48.0 (22.8)	36.0 (26.1)
9	3	66.7 (30.6)		46.7 (46.2)	40.0 (52.9)	
10	1	80.0 (-)		0 (-)		
11	154	36.7 (28.2)				
12	1		80.0 (-)		60.0 (-)	
13	2					
Total	434	46.8 (30.6)	47.6 (31.3)	51.4 (31.4)	52.4 (32.3)	51.6 (32.7)
			Emotional functioning			
1	107	71.8 (23.3)	75.7 (22.6)	74.6 (23.2)	78.1 (18.5)	73.9 (22.4)
2	35	69.8 (24.8)	73.7 (22.7)	67.1 (28.7)	61.0 (28.2)	
3	10	71.7 (23.3)	683 (25.4)	73.3 (26.0)		64.2 (26.7)
4	35	67.0 (27.6)	74.0 (20.9)	68.8 (26.3)		
5	8	77.1 (24.7)	79.2 (24.4)		89.6 (12.8)	87.5 (17.3)
6	2	87.5 (5.9)	95.8 (5.9)		100 (0)	
7	69	66.1 (26.3)	66.7 (27.1)			
8	6	61.1 (20.9)		62.5 (32.8)	58.8 (30.0)	62.5 (28.7)
9	3	75.0 (14.4)		63.9 (12.7)	69.4 (12.7)	
10	1	83.3 (-)		66.7 (-)		
11	1	58.3 (-)			50.0 (-)	58.3 (-)
12	156	60.2 (26.5)				
13	1		50.0 (-)	41.7 (-)	16.7 (-)	33.3 (-)
Total	434	66.2 (25.7)	72.8 (23.9)	71.4 (25.1)	73.8 (23.0)	73.1 (23.1)

Appendix 5.15: NPC trial - Results of Ridout logistic regression (scenario one)

Assessment	Total N (%)	Dropouts N (%)	Unadjusted OR (95% CI)	p-value	Adjusted OR(95% CI)	p-value
			Pain			
Baseline	433 (99)	185 (43)	1.008 (1.003,1.014)	0.003	1.006 (1.00,1.12)[1]	0.045
1 month	203 (47)	63 (31)	1.01 (1.00,1.02)	0.023	1.01 (1.00,1.02)[2]	0.041
2 months	149 (34)	40 (27)	1.00 (0.99,1.01)	0.70	-	-
3 months	136 (31)	47 (35)	1.00 (0.99,1.01)	0.85	-	-
			Physical functioning			
Baseline	431 (99)	183 (42)	0.98 (0.97,0.99)	<0.001	0.987 (0.979,0.995)[1]	0.001
1 month	203 (47)	63 (31)	0.97 (0.96,0.98)	<0.001	0.97 (0.96,0.98)[2]	<0.001
2 months	148 (34)	40 (27)	0.99 (0.98,1.00)	0.047	0.99 (0.98,1.00)	0.047
3 months	130 (30)	48 (37)	0.98 (0.9,0.99)	0.003	0.985 (0.972,0.997)[3]	0.018
			Emotional functioning			
Baseline	433 (99)	186 (43)	0.99 (0.98,1.00)	0.005	0.99 (0.98,1.00)[1]	0.025
1 month	201 (46)	63 (31)	0.98 (0.97,0.99)	0.001	0.98 (0.97,099)[2]	0.001
2 months	148 (34)	40 (27)	0.99 (0.98,1.01)	0.41	-	-
3 months	130 (30)	49 (38)	0.981 (0.966,0.997)	0.023	0.98 (0.97,1.00)[3]	0.057

Adjusted for: [1] Karnofsky index, cluster pair; [2] sex; [3] Karnofsky index.

Appendix 5.16: NPC trial - Mean (SD) QoL scores and odds ratio for dropout

	Continuers		Drop outs		Unadjusted OR	
	N	Mean (SD)	N	Mean (SD)	(95% CI)	p-value
Pain						
Baseline	278	42.4 (34.4)	155	56.0 (36.1)	1.008 (1.003, 1.015)[1]	0.004
1 month	200	31.7 (30.6)	69	49.0 (32.6)	1.02 (1.01,1.03)	<0.001
2 month	166	36.2 (31.5)	34	41.7 (30.5)	1.005 (0.99,1.02)	0.36
3 month	126	35.8 (32.5)	39	38.9 (33.8)	1.00 (0.99,1.01)	0.61
Physical functioning						
Baseline	277	52.5 (30.4)	154	36.7 (28.2)	0.98 (0.97,0.99)[1]	0.001
1 month	200	54.7 (29.2)	69	27.0 (28.0)	0.97 (0.96,0.98)	<0.001
2 month	165	54.8 (30.3)	34	34.7 (31.6)	0.98 (0.97,0.99)	0.001
3 month	125	57.7 (30.5)	40	36.0 (32.4)	0.98 (0.97,0.99)	<0.001
Emotional functioning						
Baseline	277	69.6 (24.6)	156	60.2 (26.5)	0.98 (0.97,0.99)[1]	0.002
1 month	198	74.9 (22.4)	69	66.7 (27.1)	0.98 (0.97,0.99)	0.016
2 month	162	72.0 (24.9)	36	68.8 (25.9)	0.99 (0.98,1.01)	0.48
3 month	123	77.2 (20.2)	40	63.5 (27.9)	0.97 (0.96,0.99)	0.002

[1] Adjusted for Karnofsky status

Appendix 5.17: NPC trial - Baseline QoL scores between responder groups at follow- up (scenario two)

	Responders		Non-responders		Mean		
Assessment	N	Mean (SD)	N	Mean (SD)	Difference	95% CI	p-value
Pain							
1 month	269	42.6 (34.5)	164	55.1 (36.0)	-12.5	(-19.3,-5.7)	<0.001
2 months	200	41.8 (34.4)	233	52.0 (35.9)	-10.2	(-16.8,-3.5)	0.003
3 months	165	40.7 (33.8)	268	51.4 (36.0)	-10.7	(-17.5,-3.8)	0.002
4 months	136	42.2 (32.8)	297	49.7 (36.5)	-7.5	(-14.7,-0.3)	0.041
Physical functioning							
1 month	268	52.2 (30.3)	163	37.9 (29.0)	14.3	(8.4,20.1)	<0.001
2 months	200	55.5 (29.8)	231	39.3 (29.2)	16.2	(10.6,21.8)	<0.001
3 months	164	56.5 (30.6)	267	40.9 (29.1)	15.6	(9.8,21.4)	<0.001
4 months	136	56.6 (27.6)	295	42.3 (30.9)	14.4	(8.3,20.5)	<0.001
Emotional functioning							
1 month	268	69.7 (24.8)	165	60.6 (26.2)	9.0	(4.1,13.9)	<0.001
2 months	199	70.5 (24.1)	234	62.6 (26.5)	7.9	(3.1,12.7)	0.001
3 months	164	71.2 (23.3)	269	63.2 (26.6)	8.0	(3.0,12.9)	0.002
4 months	135	71.9 (23.0)	298	63.6 (26.4)	8.2	(3.0,13.4)	0.002

Peer-reviewed publications

1. Shona Fielding, Graeme MacLennan, Jonathan A Cook, Craig R Ramsay. 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.

2. Shona Fielding, Peter M Fayers, Alison McDonald, Gladys McPherson, Marion K Campbell for the RECORD study group. Simple imputation methods were inadequate for missing not at random (MNAR) quality of life data. Health and Quality of Life Outcomes 2008, 6: 57.

3. Fielding S, Fayers PM, Loge JH, Jordhøy MS, Kaasa S. Methods for handling missing data in palliative care research. Palliative Medicine 2006; 20: 791-798.

4. Shona Fielding, Peter M Fayers, Craig R Ramsay. Investigating the missingness mechanism in quality of life data: A comparison of approaches. Health and Quality of Life Outcomes 2009; 7: 57.