Influence of Missing Explanatory Variables and Longitudinal Assessments in Breast Cancer Clinical Trials

Marion J. Procter

Department of Mathematics & Statistics

University of Strathclyde

Glasgow

UK

June 2016

This thesis is submitted to the University of Strathclyde for the degree of Doctor of Philosophy in the Faculty of Science

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Signed: Marian J Proctor Date: 8th June 2016

Acknowledgements

I would like to thank my supervisor Professor Chris Robertson for his help and assistance, in particular his advice on the structure of my thesis. I also acknowledge the help of the late Professor George Gettinby.

I would like to thank my employer Frontier Science Scotland Ltd. for the opportunity to work on this thesis. I thank Ms Karen Price, Professor Richard D. Gelber and the IBCSG for providing the IBCSG data and the HERA Steering Committee for permission to use the HERA data. I would also like to thank Ms Eleanor McFadden, Professor Rich Gelber, Dr. Fergus Daly and Dr. Ian Bradbury for their willingness to share their knowledge and experience.

Finally, I would like to thank my parents and my family for their support and encouragement throughout the time I have been working on this thesis.

Preface

I have worked part-time on my thesis while employed as a statistician at Frontier Science Scotland Ltd (FSS). During this time, my main work as a statistician at FSS has been on the HERA trial, which investigated adjuvant treatment of trastuzumab in early stage breast cancer. Unpublished work considering the influence of missing LVEF assessments in the HERA trial forms the basis of Chapter 6 in this thesis. The publications on the main efficacy endpoints and cardiac safety that have resulted from my work at FSS are listed below. These publications are related to work presented in this thesis, but none of the published work is part of this thesis. This is due to the fact that these publications are the result of group work and my individual contribution could not be separated out, and also due to the group agreements on publications.

Piccart-Gebhart, M.J., Procter, M., Leyland-Jones, B., et al. (2005). Trastuzumab after Adjuvant Chemotherapy in HER2-Positive Breast Cancer, *New England Journal of Medicine*, 353, 1659-1672.

Smith, I., Procter M., Gelber RD., et al. (2007). 2-year follow-up of trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer: a randomised controlled trial, *Lancet*, 369, 29-36.

Suter, T.M., Procter M., van Veldhuisen D.J., et al. (2007). Trastuzumab-Associated Cardiac Adverse Effects in the Herceptin Adjuvant Trial, *Journal of Clinical Oncology*, 25, 3859-3865.

Procter, M., Suter, T.M., de Azambuja, E. et al. (2010). Longer-Term Assessment of Trastuzumab-Related Cardiac Adverse Events in the Herceptin Adjuvant (HERA) Trial, *Journal of Clinical Oncology*, 28, 3422-4328.

Gianni, L., Dafni U., Gelber R.D. et al. (2011). Treatment with trastuzumab for 1 year adjuvant chemotherapy in patients with HER2-positive early breast cancer: a 4-year follow-up of a randomised controlled trial, *Lancet Oncology*, 12, 236-244.

Goldhirsch, A., Gelber, R.D., Piccart-Gebhart M.J. et al. (2013). 2 years versus 1 year of adjuvant trastuzumab for HER2-positive breast cancer (HERA): an open-label, randomised controlled trial, *Lancet*, 382, 1021-1028.

de Azambuja, E., Procter, M.J., van Veldhuisen D.J. et al. (2014). Trastuzumab-Associated Cardiac Events at 8 Years of Median Follow-Up in the Herceptin Adjuvant Trial (BIG 1-01), *Journal of Clinical Oncology*, 32, 2159-2165.

Abstract

Clinical trials in breast cancer assess treatment regimens based on a balance of efficacy and adverse effects. To achieve high-quality evidence for these assessments, it is important to minimise potential sources of bias. Therefore, potential bias in the parameter estimates resulting from missing observations is an important concern.

In this thesis, the influence of missing data on explanatory variables in time-dependent Cox model analysis is explored, with application to breast cancer clinical trials. In particular, imputation in the context of time-dependent covariates that may be informative missing data which is described has not been studied in detail in the statistical literature. Standard imputation methods from the statistical literature are described, which involve assumptions about the missing data mechanism. Missing observations of quality of life (QoL) are imputed by standard methods before analysis of disease-free survival (DFS) and the performance of the imputation methods is considered. Then the influence of missing observations of an outcome variable assessing safety is considered. Repeated measures analysis of a safety assessment is performed. The insights into the influence of missing data could be generalised.

Two clinical trials are considered; the International Breast Cancer Study Group (IBCSG) Trials VI and VII and the Herceptin Adjuvant (HERA) trial. Both investigated adjuvant treatment in breast cancer. There was no evidence in Trials VI and VII that the patient's QoL is related to the patient's DFS, though such a relationship could be masked by the missing observations. Simulation was performed in the context of a positive relationship between QoL and DFS. The simulation study suggested that the performance of the standard imputation methods was influenced by the missing data mechanism. There was no benefit from imputing LVEF values in the HERA trial. It was appropriate to perform the repeated measures analysis of LVEF values using observed LVEF values only.

Contents

List of Tables

List of Figures

1 Features of Breast Cancer Clinical Trials

1.1 Introduction

The focus of this thesis is the influence of missing data on explanatory variables that may be related to disease-free survival (DFS), such as quality of life, and also missing longitudinal assessments which are performed to assess major safety endpoints. The application is to breast cancer clinical trials. The potential problems associated with missing observations include bias of the parameter estimates and loss of power to detect clinically important differences among treatment groups over time (Fairclough 2010, Chapter 6; Little and Rubin 2002, Chapters 1 and 3). Imputation is among the methods proposed in the statistical literature to address these concerns (e.g. Rubin 1987; Molenberghs and Kenwood 2007). Standard imputation methods are reviewed and outlined in this thesis. Missing values of quality of life are imputed using standard imputation methods before analysis investigating the possible relationship between quality of life and DFS. The performance of standard imputation methods is considered. Then the influence of missing values of an outcome variable assessing safety is investigated. Repeated measures analysis of a safety assessment is performed.

This chapter begins by considering general features of breast cancer clinical trials. The two main aspects, efficacy and safety, are outlined, as well as the considerations of most relevance to this thesis: i) quality of life, ii) potential sources of bias, and iii) controlling bias. The subsection on general features concludes by describing considerations relating to the statistical analysis of breast cancer clinical trials. Then the clinical trials considered in this thesis are introduced. Next, quality of life and the main aspects of breast cancer clinical trials, efficacy and safety, are described in the context of the two clinical trials considered in this thesis. The standard analysis for time to event endpoints, the Cox proportional hazards model (Cox 1972), is described. Missing observations in longitudinal data are then considered. This includes i) the problem of missing data

in longitudinal data, ii) methods to deal with missing data, such as imputation, and iii) its prevalence. At the end of the chapter, the layout of thesis is described.

1.1.1 General Features of Breast Cancer Clinical Trials

Clinical trials in breast cancer are performed to assess the effects of treatment regimens. Clinical judgment of treatment regimens for breast cancer is based on balancing efficacy with adverse effects. In addition, the risk of taking part in a clinical trial is of importance to patients. Efficacy and safety (see section 1.1.2) are the two main aspects of breast cancer clinical trials and are assessed in order to provide evidence for making clinical judgments and to protect patients. It is usual that the main treatment comparisons in breast cancer clinical trials are based on DFS and overall survival (OS). Traditional endpoints such as these do not reflect the patient's sense of well-being and thus it is becoming increasingly common for quality of life to be assessed throughout the study (Fairclough 2010, p.1).

In order to achieve high-quality evidence on the benefits and risks of treatment regimens, careful attention must be given to the study design, execution and analysis. Potential sources of bias and controlling bias are among considerations during the design stage of a breast cancer clinical trial (Fairclough 2010, Chapter 6; Fayers and Machin 2007, p.355; Little and Rubin 2002, Chapter 1) (see section 1.1.3). The potential bias of most interest in this thesis is the potential bias in parameter estimates associated with missing observations and this is described in detail in section 1.6. Considerations related to the statistical analysis required to meet the trial goals generally include Type I and Type II errors and power related to hypothesis tests (Piantadosi 2005, Chapters 6 and 7) (see section 1.1.4). The power of hypothesis tests in the context of analysis of data with missing observations is of interest in this thesis (see section 1.6).

1.1.2 Efficacy and Safety

The evidence to make clinical judgments of treatment regimens for breast cancer comes from therapeutic clinical trials. Based on earlier experience of treatment

regimens in clinical trials, confirmatory trials evaluate the key clinical question relevant to efficacy or safety, which is formulated in advance as a hypothesis. The primary endpoint assessed in the key hypothesis of interest should be the variable which provides the most clinically relevant and convincing evidence directly related to the clinical question of most interest (International Conference on Harmonisation 1998, p.7). Generally, the endpoints of most interest in breast cancer clinical trials are DFS and OS. The standard analysis for these time to event endpoints is a Cox proportional hazards model. Explanatory variables which may be related to efficacy may be included as covariates in the Cox proportional hazards model analysis for the main analysis.

In well designed and executed studies evaluating time to event endpoints, for patients who do not have the event being considered at the time of the analysis, there is an appropriate censoring date available. In addition, there are few missing observations in the explanatory variables included as covariates in the Cox proportional hazards model for the main analysis. Therefore, it is not expected that missing observations cause difficulties in the main analysis of time to event endpoints.

Breast cancer has a highly variable prognosis and benefit from a treatment regimen is unpredictable for the individual patient. Increasing attention is being paid to the molecular factors of the tumour with the aim of providing early and accurate information on outcome and benefit from a treatment regimen. (Urruticoechea et al. 2005). Thus, it is becoming increasingly common in breast cancer clinical trials to assess explanatory variables which may be related to efficacy, such as the molecular marker Ki-67, throughout the study (Urruticoechea et al. 2005). When covariates are measured throughout the study and so vary over time they are referred to as time-dependent covariates.

In addition to providing evidence used in making clinical judgements, patient safety is monitored in clinical trials in order to protect patients. Safety can often be summarised in terms of the risk of clinically relevant adverse events

(Piantadosi 2005, p. 411). It is common for medical assessments which monitor patient safety to be repeated throughout the breast cancer trials. An example is assessments of left ventricular ejection fraction (LVEF) performed to assess cardiac function (e.g. Suter et al. 2007). Medical assessments repeated throughout the study may be used as time-dependent covariates in an analysis of time to occurrence of major safety endpoints.

1.1.3 Other Considerations in the Design and Execution of Breast Cancer Clinical Trials

While the endpoints of DFS and OS are generally of most interest in breast cancer clinical trials, it is becoming increasingly common for quality of life also to be assessed. The goal of a clinical trial is to achieve high-quality evidence on the benefits and risks of treatment regimens. Therefore, among the considerations in breast cancer trials are: i) quality of life, ii) potential sources of bias, and iii) controlling bias, which are described in this subsection. Considerations relating to the statistical analysis of breast cancer clinical trials, such as the Type I error, are described in section 1.1.4.

Quality of Life

The patient's ability to carry out day to day activities and how the patient feels will influence the patient's perception of whether a treatment is beneficial and the patient's perception of his or her health (Fairclough 2010, p.1). However, these factors are not reflected in traditional endpoints of efficacy and increasingly endpoints which address the patient's perception of his or her health are included in clinical trials. The question of whether good quality of life is associated with good prognosis is of clinical interest in breast cancer trials (e.g. Epplein et al. 2011; Keene Sarenmalm et al. 2009; Coates et al. 2000). The fact that quality of life assessments are commonly repeated throughout the study makes quality of life assessments suitable as a time-dependent covariate in a time-dependent Cox model.

Potential Sources of Bias

Error has two components, a completely random one and a systematic one with a net direction and magnitude which is bias. Bias may arise from many sources, for example selection effects, uncontrolled prognostic factors, procedural errors and statistical methods. Age, severity of disease or comorbidities are examples of covariates that can appear as selection effects. (International Conference on Harmonisation 1998, p.5; Piantadosi 2005, Chapter 7). Another potential source of bias is missing observations (see section 1.6). Bias can be large relative to the treatment effect being investigated and cannot be resolved by replication of observations. The implications of bias may be difficult to know in advance, but it is often possible to understand factors that may contribute to bias and take steps to control these factors by careful design and conduct of the trial (Piantadosi 2005, Chapter 7).

Controlling Bias

The main methods for controlling bias described in the statistical literature are: randomisation, blinding, concurrent controls, objective assessments, endpoint ascertainment (International Conference on Harmonisation 1998, p.10-12; Piantadosi 2005, Chapter 7) and these methods are briefly described below.

Randomisation

Bias in selecting the patients in the clinical trial may affect the external validity of the results from the trial. Randomisation is the main method available for reducing selection bias. It is the only reliable method for controlling the effects of unknown covariates among the patients in the treatment groups (Piantadosi 2005, p.179-180). Randomisation allows the assumption to be made that any differences in the observed and unknown covariates between treatment groups are attributable to chance. If other sources of bias have been removed, the difference in the outcome between the treatment groups can then be attributed to the treatments and thus providing evidence of causality.

Blinding

Blinding reduces assessment bias, where the investigator's or patient's assessment is influenced by knowledge of the treatment the patient received. Generally, this is in the scenario of drug trials.

Single blinding refers to the situation where the patient is unaware of which treatment he or she receives. In order to achieve this, the active treatment and the placebo must look, feel or taste the same and the investigator must be careful not to reveal the treatment group to the patient. Blinding can improve the objectivity of partially subjective endpoints. For example, if patients in a trial of analgesics know they are receiving the investigational treatment, they may be biased in favour of the treatment, causing them to overstate the treatment's efficacy (Piantadosi 2005, p.180). However, there are scenarios where single blinding is not possible, for example if comparing chemotherapy treatment with surgery (International Conference on Harmonisation 1998, p.8-9).

Double blinding refers to the situation where both the patient and the investigator responsible for assessing the outcome do not know which treatment the patient receives. Double blinding can further increase the usefulness of partially subjective endpoints. For example, if the investigator has knowledge of preclinical results in favour of the investigational treatment, he or she may be influenced by his or her expectations of the performance of the investigational treatment (Piantadosi 2005, p.180).

Concurrent Controls

Concurrent controls are an effective method of reducing bias, by removing the confounding of treatment effect with calendar time. Concurrent controls also make the use of randomisation more straightforward. For example, improvements in survival among cancer patients over calendar time make it impossible to determine the treatment effect from historical controls.

Objective Assessments

An objective assessment is an assessment for which independent reviewers would agree on the result. Whenever possible, the methods used to assess trial endpoints should be objective as this reduces bias and increases the reproducibility of the trial results.

However, there are circumstances where the most objective assessment is not the most appropriate. For example, when assessing pain intensity, the assessment of the investigator is more objective than the patient's assessment but may be poorly correlated with the assessment of the patient (Grossman et al. 1991). It is likely health professionals underestimate pain and overestimate functional abilities (Piantadosi 2005, p.181). In a clinical trial, it would be more appropriate to consider the patient's assessment of pain intensity, with the patient blinded to the treatment received.

Endpoint Ascertainment

It is important during the design stage of a trial that investigators plan follow-up procedures in order to ascertain trial endpoints. For example, it is not reasonable to assume patients who do not attend the clinic for a scheduled visit are alive and well. By actively determining the follow-up status of patients, by such methods as scheduled clinic follow-up visits and phone interviews, then the chance of ascertainment bias is reduced. The precise date of a breast cancer recurrence is difficult to determine. In breast cancer clinical trials, it is standard for the date of breast cancer recurrences to be reviewed centrally by the data management and medical staff (e.g. The International Breast Cancer Group 1997). The final adjudication of an endpoint may be made by a committee independent of the investigator (e.g. Chlebowski et al. 2003).

1.1.4 Other Considerations Relating to the Statistical Analysis of Breast Cancer Clinical Trials

As well as the design and execution of a clinical trial, attention must be given to the statistical analysis to meet the goals of the trial. Generally, in clinical trials the primary analysis will be based on hypothesis tests and/or confidence intervals (International Conference on Harmonisation 1998, p.3). The primary analysis may be based on a stratified or unstratified hypothesis test. When making an inference from hypothesis tests, there are two random errors possible; "a false positive" (Type I) and a "false negative" (Type II). The consequences of type I and type II errors are different (Piantadosi 2005, Chapter 7). This subsection begins by discussing stratification. Then the relationship between type I and type II errors and power, and factors influencing the probability of a type II error are described. Finally, p-values and confidence intervals are compared.

Stratification

While unrestricted randomisation is an acceptable approach, even in randomised trials, imbalances between treatment groups can occur by chance. Stratification by important prognostic factors measured at baseline may sometimes be valuable to help increase the comparability of the treatment groups. To be most advantageous, stratification should only be used for prognostic factors with relatively strong effects. The use of more than two or three stratification factors is rarely necessary (International Conference on Harmonisation 1998, p.10; Piantadosi 2005, p. 337-338).

When stratification is used in a clinical trial, the factors on which randomisation is stratified should be accounted for in the statistical analysis (International Conference on Harmonisation 1998, p.10). An example would be using a stratified log-rank test (Mantel 1966) for the comparison of treatment groups. The statistical section of the protocol should indicate if the primary analysis is based on a stratified or an unstratified analysis.

Type I and Type II Errors and Power

The type I error is a "false positive" result and occurs if there is no treatment effect but the investigator wrongly rejects the null hypothesis and concludes that there is a treatment effect. The type II error is a "false negative" and occurs when the investigator fails to reject the null hypothesis and so fails to find a treatment effect that exists. The power of a hypothesis test is the probability of finding a treatment effect of a given size to be statistically significant when the alternative hypothesis is correct and is therefore is 1 – the probability of a type II error. Conventionally, the power is set at 80% and 90% in clinical trials (Piantadosi 2005, p.277). Loss of statistical power due to missing observations in considered in section 1.6.

Generally, the only factor that controls the type I error is the critical value of the hypothesis test. The type I error does not depend on the sample size. However, the type I error can become inflated when the multiple tests are performed, for example when several treatment group combinations are compared using multiple hypothesis tests. If multiple hypothesis tests are performed at the same significance level, the type I error of each individual hypothesis test will be controlled but the overall type I error across the clinical trial will increase. In this situation, methods for controlling the overall type I error should be carefully considered during the study design. Often the probability of a type I error in a clinical trial is set to 0.05, but there are circumstances where a higher or lower type I error is more appropriate (Piantadosi 2005, p.277).

In particular, careful consideration should be given to the type I and type II error rates when designing a trial to demonstrate equivalence (a non-inferiority trial). In such trials the null hypothesis may be framed as "the treatments are different" and the alternative hypothesis framed as "the treatments are the same". This is the reverse of what is standard in trials designed to demonstrate superiority of a treatment (Piantadosi 2005, p.290). For example, in the PhARE trial, the null hypothesis was "6 months of adjuvant trastuzumab treatment is not inferior to 12-

month treatment by a pre-specified acceptable margin in terms of disease-free survival" (Pivot et al. 2013). Such hypothesis tests are naturally one sided, and the roles of the Type I and Type II error must be considered when selecting the values (Piantadosi 2005, p.290).

Factors Influencing the Type II Error

The probability of a type II error is influenced by three factors. These factors are the critical value for the rejection of the null hypothesis, the variance of the estimator under the null hypothesis and the treatment effect assumed in the study design. In principle, investigators have control over the critical value of the hypothesis test and the variance of the estimator under the null hypothesis, which is directly related to the sample size.

The treatment effect assumed in the study design is generally based on the declared minimum clinically important difference. In this scenario, the probability of a type II error for a given sample size will increase as the assumed minimum clinically important difference increases. The advantage of a large clinical trial is that it can reliably find a treatment difference which is realistically small but clinically important (Piantadosi 2005, Chapter 7).

P-values vs Confidence Intervals

The p-value from a hypothesis test is the probability of obtaining a value of the test statistic equal to or more extreme than the observed test statistic when the null hypothesis is true. Therefore, p-values are probability statements made under the null hypothesis. It is important to remember that p-values are poor summaries of treatment effects (Royall 1986). In particular, p-values do not provide information about the magnitude of the estimated treatment effect which is required to assess its clinical importance. In addition, the p-value partially reflects the sample size through the variability of the estimated treatment effect and the sample size is not of clinical importance.

Confidence intervals are centred on the estimated treatment effect and provide information on the magnitude and precision of the estimate. This makes confidence intervals more useful than p-values. (Piantadosi 2005, Chapter 16). For example, the 95% confidence interval for the hazard ratio for a DFS event in the trastuzumab group vs the observation group of 0.43 to 0.67 described in Piccart et al. (2005) is more informative than the corresponding p-value <0.0001. However, it should be remembered that confidence intervals and hypothesis tests are related. In particular, a confidence interval is a collection of hypotheses that cannot be rejected at a specified significance level.

This subsection discussed statistical considerations relating to breast cancer trials. These would have been amongst the considerations in the published analyses for the breast cancer trials illustrated in this thesis. These breast cancer trials are described in the next section.

1.2 Breast Cancer Clinical Trials Considered in this Thesis

Adjuvant treatments with chemotherapy, endocrine therapy and combinations of both have been shown to increase DFS and OS in breast cancer patients (e.g. The International Breast Cancer Study Group [IBCSG] 1997). The breast cancer trials considered in this thesis, IBCSG Trials VI and VII and the more recent Herceptin Adjuvant (HERA) Trial, were large, international studies that investigated adjuvant treatment in early stage breast cancer.

IBCSG Trials VI and VII

IBCSG Trial VI was designed to examine different durations and timing of adjuvant chemotherapy in premenopausal and perimenopausal patients. IBCSG Trial VII compared tamoxifen alone or together with different durations and timing of chemotherapy among postmenopausal patients (see Figure 3.1). There were 1554 patients randomised to Trial VI between July 1986 and April 1993 and 1266 patients randomised to Trial VII during the same time period. Quality of life was an important consideration in Trials VI and VII and data on patient's self-

assessed quality of life were prospectively collected throughout the study. Hürny et al. (1996) found that between baseline and 18 months, there was a significant improvement of quality of life over time. There was a significant adverse impact of delayed chemotherapy on all quality of life measures. Herring et al. (2004) noted that poor baseline quality of life was associated with improved prognosis in postmenopausal patients. This may reflect the fact that chemotherapy treatment is toxic.

The further analysis of quality of life in Trials VI and VII presented in Chapter 3 of this thesis investigated the hypothesis that the patient's quality of life as measured by coping/perceived adjustment ("coping score") was related to the patient's DFS. The high proportion of missing coping scores and findings from previous statistical analysis of quality of life indicated that imputation is appropriate. The IBCSG dataset was the basis for simulated datasets with a known relationship between quality of life and DFS considered in Chapters 4 and 5. These simulated datasets are used to investigate if the performance of the standard imputation methods given different missing data mechanisms is influenced by the relationship between quality of life and DFS.

HERA Trial

The HERA trial was designed to investigate whether trastuzumab treatment was effective as adjuvant treatment for early stage breast cancer if used after the completion of the primary treatment and benefit in DFS and OS has been shown (Piccart-Gebhart et al. 2005; Smith et al. 2007). Cardiac function was monitored in all patients as trastuzumab treatment is associated with congestive heart failure (CHF) in patients with metastatic breast cancer. Assessments of LVEF were performed throughout the study as part of the cardiac monitoring and the percentage of missing LVEF assessments was low. Statistical analysis of cardiac safety in the HERA trial found a low incidence of CHF and suggested that cardiac dysfunction associated with adjuvant trastuzumab treatment is generally reversible (Suter et al. 2007; Procter et al. 2014).

The influence of missing LVEF assessments is investigated in the further analysis of cardiac safety in the HERA trial. As part of the further cardiac analysis, the LVEF values were used as a time-dependent covariate in a time-dependent Cox model for occurrence of a cardiac endpoint. Data were available for 3386 patients randomised between December 2001 and March 2005 with a median follow-up of 1 year. In this setting, it was appropriate to perform the repeated measures analysis of the LVEF values over time in a mixed model using observed LVEF values only.

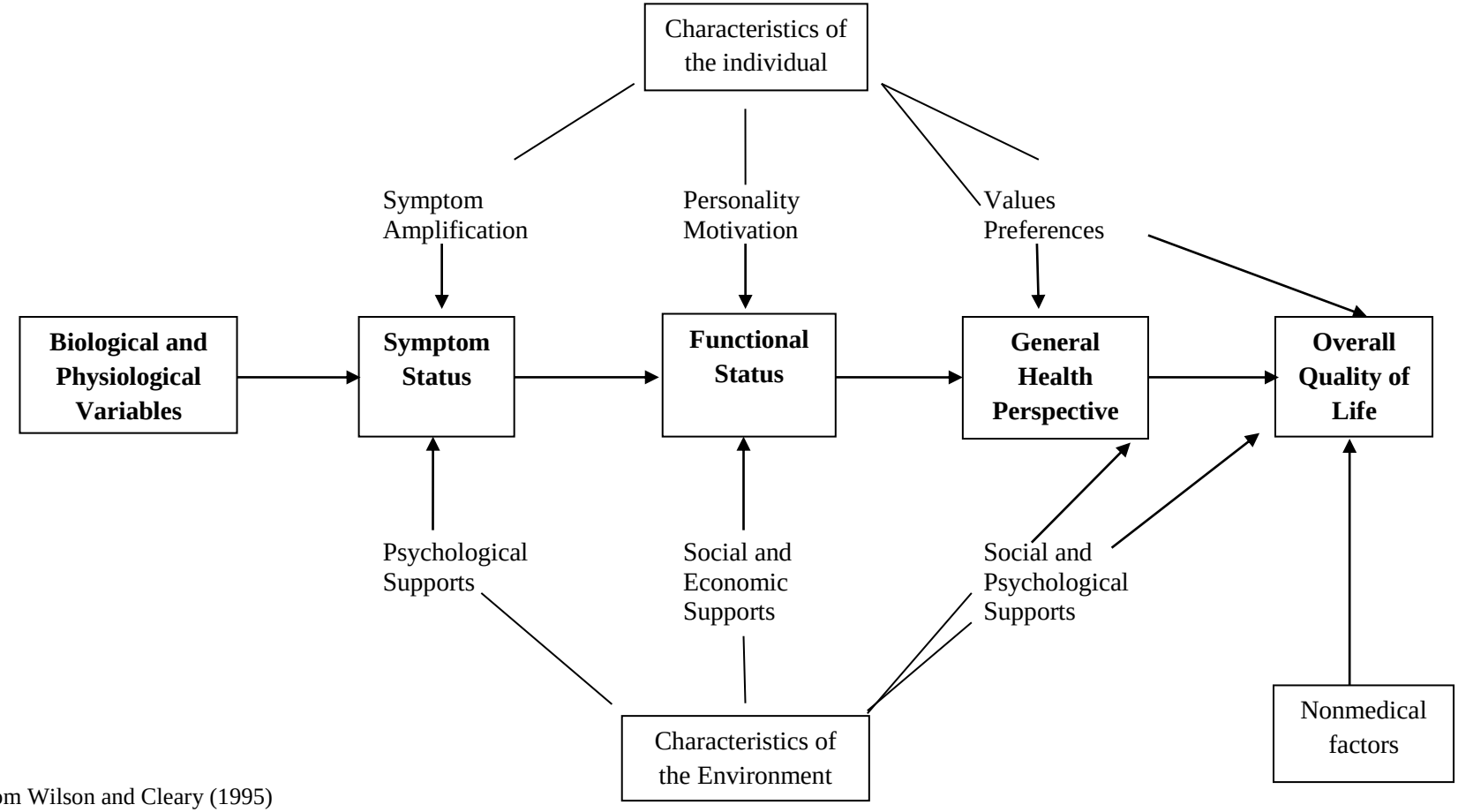
1.3 Quality of Life

Adjuvant therapy for early breast cancer can have substantial adverse effects, but improves DFS and overall survival. An important clinical question is whether the benefits in efficacy are worth the adverse effects on quality of life (Hürny et al. 1996). Therefore, quality of life was an important endpoint in IBCSG Trials VI and Trials VII. The relationship between quality of life and DFS and delayed chemotherapy and DFS is explored in later chapters of this thesis (Chapters 3-5). Some important aspects of quality of life are described in this section. These aspects are: the definition of quality of life, the reason for and background to assessing quality of life and considerations in measuring quality of life.

1.3.1 Definition of Quality of Life

The World Health Organisation (WHO, 1948 and 1958) defined health as a "state of complete physical, mental and social well-being and not merely the absence of infirmity or disease." Measures of health can be thought of as existing on a continuum of increasing biological, social and psychological complexity. Wilson and Cleary (1995) proposed the conceptual model described in Figure 1.1. The main purpose of the figure was to distinguish among conceptually distinct measures of quality of life and for the authors to make explicit what they consider the dominant causal associations.

Figure 1.1 Conceptual Model of Quality of Life



Taken from Wilson and Cleary (1995)

Most conceptualisations of quality of life include the dimensions of physical functioning, social functioning, role functioning, mental health and general health perceptions (Ware 1987; Ware et al. 1981). Important concepts such as energy/fatigue, pain and cognitive functioning are incorporated in these broader categories. Clinical data, such as patient-reported symptoms, are generally not included in conceptualisations of quality of life (Wilson and Cleary 1995). While symptoms and adverse events impact on quality of life and will form part of a patient's assessment of his or her assessment of quality of life, they are not equivalent to quality of life (Fairclough 2010, p.3).

The term quality of life is not well defined. However, there is general agreement that it is a multi-dimensional concept that focuses on the impact of disease and its treatment on the well-being of a person (Fayers and Machin 2007, p.4; Fairclough 2010, p.2). In its broadest definition, quality of life is influenced by our environmental and social conditions. Kaplan and Bust (1982) proposed the term health-related quality of life to distinguish health effects from other environmental and social factors influencing a person's perception of quality of life, for example job satisfaction. Proposed definitions of health-related quality of life include those of Cella and Bonomi (1995) and Patrick and Erickson (1993). The definition proposed by Cella and Bonomi is: "Health-related quality of life refers to the extent to which one's usual or expected physical, emotional and social well-being are affected by a medical condition or its treatment." The definition of health-related quality of life proposed by Patrick and Erickson (1993) is broader and considers quantity: "The value assigned to duration of life as modified by impairments, functional states, perceptions and social opportunities that are influenced by disease, injury, treatment or policy." In general, the scope of clinical trials is limited to the assessment of health-related quality of life.

1.3.2 Reason for and Background to Assessing Quality of Life

Reason for Assessing Quality of Life

The endpoints of clinical trials have traditionally focused on endpoints that are physical or laboratory measures of response. The main treatment comparisons in breast cancer clinical trials are generally based on DFS and OS. However, traditional endpoints do not reflect how the patient feels or the patient's ability to carry out day to day activities. These factors are likely to influence the patient's perception on whether the treatment is beneficial and the patient's perception of his or her health. More recently, clinical trials have been designed with endpoints which include the patient's perception of his or her well-being, monitored throughout the study.

Historical Development of Quality of Life Instruments

In the late 1970s and early 1980s, questionnaires were developed that focused on physical functioning, physical and psychological symptoms, impact of illness, perceived distress and life satisfaction. However, these instruments were used for general evaluation of health and were not designated as quality of life instruments by the authors. An example of these questionnaires is the Sickness Impact Profile (Gilson et al. 1975). During the same time period, Priestman and Baum (1976) adapted visual analogue scales (VAS) to assess quality of life in breast cancer patients. VAS consist of a line with descriptive anchors at the extremes and often the length of the line is 10cm. Several subjective effects, such as the patient's opinion as to "Is the treatment helping?" were assessed.

Building on these early developments, more recent quality of life instruments have tended to emphasis subjective aspects of quality of life, such as emotional functioning, while continuing to consider functional capacity. It is common for one of more questions relating to overall quality of life also to be included.
It is important to note that a person's quality of life cannot be directly measured and it is only possible to make inferences from measurable indicators of symptoms and reported perceptions (Fairclough 2010, p.3).

1.3.3 Health Status Assessment Measures

Characteristics of Health Assessment Measures

In health status assessment measures, multiple aspects of the patient's perceived well-being are assessed when the patient responds to a series of questions from which a score is derived. This score provides a relative comparison of the patient's quality of life to the quality of life in the same patient at other times and to that of other patients. The assessment measures contained in the quality of life instrument often include a series of questions about particular aspects of the patient's daily life during a recent period of time and a global question asking the patient to rate his or her current quality of life. Some questions among health status assessment measures focus on the perceived impact of the disease and treatment, for example "Has your physical condition or medical treatment interfered with your family life?" (Not at All, A Little, Quite a Bit, Very Much) in European Organization for Research and Treatment of Cancer (EORTC) QLQ-C30 (Aaronson et al. 1993). Other questions focus on the frequency and severity of symptoms, for example "Have you had pain?"(Not at All, A Little, Quite a Bit, Very Much) in EORTC QLQ-C30. The main purpose of health status assessment measures is to compare quality of life among patients in different treatment groups or to identify changes in quality of life over time among patients in the same treatment group.

Objective vs Subjective

Health status assessment measures vary in the extent to which they assess events which can be observed or require the patient to make inferences. Some health status assessment measures assess symptoms or functional abilities which can be measured objectively, for example the frequency or severity of symptoms or whether the patient can perform certain tasks such as walking a mile. In contrast, some health status assessment measures assess the impact of symptoms or conditions, such as how much symptoms interfere with daily activities. Quality of

life instruments generally contain a combination of different types of health status assessment measures (Fairclough 2010, p.5).

It may be that objective health status assessments are inappropriately considered more valid. As an example, while patient ratings have sometimes been found not to agree with the ratings of medical staff (Grossman et al. 1991), consideration should be given to the fact that the patient has more complete information than medical staff about the patient's perception of his or her health and quality of life. The expectation that an objective health status assessment has less measurement error than a subjective assessment may be a reason for considering objective assessments more valid. This reasoning may not be correct when the objective endpoint has a high degree of measurement error. It should also be noted that some traditional endpoints contain a high degree of measurement error, for example blood pressure, or have poor predictive and prognostic validity, for example pulmonary function tests (Wiklund 1990).

Generic and Disease-Specific Instruments

Quality of life instruments can be designed to be generic or disease-specific. Generic instruments are designed to assess quality of life in people with and without active disease, for example EuroQoL (EQ-5D) (The EuroQoL Group 1990). This is an advantage when following patients for an extended period of time after treatment has ended. Disease-specific instruments are designed to be narrower in scope and assess in detail the impact of a particular disease or treatment, for example EORTC QLQ-C30. Therefore, they are more likely to detect a small but clinically important change in quality of life related to treatment.

Global Measurements and Aspect-Specific Measurements

Some quality of life instruments are designed to provide a single global indicator of quality of life. An example is the question "Please mark with an X the appropriate place within the bar to indicate your rating of this person's quality of

life during the past week." (Lowest Quality – Highest Quality) which forms the Spitzer Uniscale intended to be completed by physicians (Spitzer et al. 1981). Others are designed to provide a profile of the dimensions such as the physical, emotional, functional and social well-being of patients, for example Functional Assessment of Cancer-General Version (FACT-G) (Cella et al. 1993). Many quality of life instruments attempt to assess both, for example EORTC QLQ-C30.

Profiles of different aspects of quality of life are useful when the objective of the clinical trial is to investigate the possibly different treatment effect on several dimensions of quality of life. The advantage of a single indicator of quality of life is that it provides a simple approach to decision making. However, there are considerable challenges in constructing a single score that combines multiple aspects of quality of life and is valid in all contexts. A single indicator of quality of life has a potential for loss of information. It is possible that a particular treatment produces a benefit in one aspect of quality of life and a reduction in another aspect of quality of life which cancel each other out. (Fairclough 2002, p.5). Though these impacts on different aspects of quality of life may be important in the patient's assessment of whether the treatment is beneficial, they would not be observed in a single indicator of quality of life.

1.3.4 Reliability and Responsiveness and Response Format

Reliability and Responsiveness

The question "Would the patient give the same response at another time if he or she was experiencing the same quality of life?" is referred to as reliability. If the same patient experiencing the same quality of life has a large variation in responses, then it is difficult to discriminate between different levels of quality of life experienced by the patient or to identify any change in quality of life over time (Fairclough 2010, p.41; Fayers and Machin 2007, p.91).

The question "Is the quality of life instrument sensitive to changes that are considered important to the patient?" is referred to as responsiveness. In a clinical

trial, responsiveness directly affects the ability to identify treatment-related changes in quality of life (Fairclough 2010, p.41; Fayers and Machin 2007, p.101).

Response Format

Quality of life instruments vary in their response format. The Likert scale (Likert 1932) and the Visual Analogue Scale (VAS) are two of the most commonly used formats. The Likert scale contains a limited number of ordered responses and each response has an associated descriptive label. Individuals can discriminate at most 7 to 10 ordered categories and reliability reduces at 5 or fewer levels (Miller 1956; Streiner and Norman, p.48-49).

The VAS consists of a line with descriptive extreme anchors and often the length of the line is 10cm. The person is asked to mark the appropriate place on the line. Priestman and Baum (1976) developed VAS to assess quality of life in breast cancer patients. The concept behind the VAS is that the measure is continuous and potentially discriminates better than a Likert scale, however this has not generally been true in validation studies where both formats have been used (Fairclough 2010, p.7). It may be that patients are more likely to judge how far along the line applies to him or her in round values and then mark on the line the place that he or she considers as the appropriate round value. The VAS format has the following limitations: i) it requires a level of eye-hand co-ordination that may be unrealistic in certain patients, ii) it requires a data management step where the position of the mark on the line is measured and iii) it excludes telephone assessment and interview formats (Fairclough 2010, p.7).

1.3.5 Summary

In clinical trials, the scope is generally limited to health-related quality of life, representing the impact of disease and treatment on a patient's perception of his or her well-being. Quality of life is multidimensional including physical, emotional, functional and social components. It is subjective and represents the patient's

perspective. Quality of life is not well defined and different instruments use different definitions. There is a wide range of instruments, however this range is expected to be reduced once the purpose of evaluating quality of life and the disease area has been considered. They are designed to be completed by the patient and frequently have several subscales.

Quality of life will influence the patient's perception of whether a treatment is beneficial. Adverse events are the most common reason for patients to discontinue treatment. Patient safety is monitored throughout the breast cancer trials and is described in the next section.

1.4 Safety in Breast Cancer Clinical Trials

The risk in taking part in a clinical trial is important to the patient. Patient safety is monitored in clinical trials in order to protect patients and inform clinical judgement. This section considers the scope of evaluation of safety data and statistical analysis of safety data.

Scope of Evaluation of Safety Data

The safety and tolerability profile of a treatment regimen for breast cancer is established by therapeutic breast cancer trials. These trials provide an important opportunity to explore potential adverse effects that were not previously associated with the treatment regimen, even though these investigations may have a low power (International Conference on Harmonisation 1998, p.29). Ongoing safety evaluation is performed throughout the trial to protect patients. Procedures should be in place to promptly notify all concerned investigator(s)/site(s) of findings that could adversely affect the safety of subjects and to expedite the reporting of all adverse drug reactions that are both serious and unexpected (International Conference on Harmonisation 1996, p.25-26). Reporting to regulatory authorities and/or institutional review board(s) may also be required (International Conference on Harmonisation 1996, p.25-26). An independent data monitoring committee (IDMC) may be established to assess at intervals the

progress of a trial, including safety data. An IDMC makes recommendations to the sponsor on whether to continue, modify or terminate a trial after reviewing the trial information (International Conference on Harmonisation 1998, p.21).

The safety information collected is generally based on adverse events, laboratory tests concerning blood chemistry and haematology and vital signs. An example in breast cancer trials is collecting severity of adverse events that are associated with chemotherapy, such as nausea/vomiting. It is important to be able to identify the severity of an adverse event, adverse events leading to death, and whether an adverse event is classified as serious. Adverse events must be reported regardless of the investigator assessment of whether the adverse event was related to the treatment regimen. Procedures should be in place to promptly notify all concerned investigators expedite the reporting to all Another important issue to explore is the reasons for discontinuing the treatment regimen (International Conference on Harmonisation 1998, p.30).

Statistical Analysis of Safety Data

Patients who do not receive any of the study treatment regimen are generally excluded from the statistical analysis of safety data. The population groups for the statistical analysis of safety data may be defined by the treatments the patient received rather than the randomised treatment groups. It is common for safety data to be summarised in terms of the risk of clinically relevant adverse events and for the occurrence of adverse events of interest to be expressed as an incidence. When calculating an incidence, it is important to consider the appropriate specification of the denominator. For example, it may be appropriate to consider the denominator as the number of patients who received at least one dose of the investigational treatment or as the total exposure time (in patient-years) (International Conference on Harmonisation 1998, p.30).

Often the probability of the risk of an adverse event over time is of interest and therefore it is common for medical assessments monitoring patient safety to be

repeated throughout the study. In this situation, summaries of the change in medical assessments from baseline over time and modelling the result of the medical assessment over time is likely to be of clinical interest. For example, it is of interest to model LVEF assessment over time in the HERA trial (see Chapter 6). Models for longitudinal data are described in section 1.6.4. In the next section, modelling the time-to-event endpoints which are generally of most interest in breast cancer trials is described.

1.5 Cox Proportional Hazards Model and Time-Dependent Cox Model

The standard analysis for the time to event endpoints in breast cancer clinical trials is a Cox proportional hazards model. Covariates related to efficacy may be included in the Cox proportional hazards model. Breast cancer trials increasingly have covariates measured throughout the study. Such explanatory variables may be included as time-dependent covariates in a time-dependent Cox model analysis. The analysis of some time to event endpoints may be complicated by another type of event which prevents the event of interest occurring. This section considers the proportional hazards assumption, the assumption of a parametric form for the baseline hazard function, parameter estimates, time-dependent covariates and competing risks.

Proportional Hazards Assumption

In a survival model, the hazard function describes the instantaneous risk of failure for subjects that are surviving at time *t*. The effect parameters describe how the hazard varies in response to explanatory variables, such as the treatment group (Kalbfleisch and Prentice 2002, Chapter 2). The hazard ratio describing the treatment effect is assumed to be constant over time. In addition to an indicator for the treatment group, covariates may be included in the survival model in an attempt to control for confounding effects. An example of covariates that may be included in survival models are age and gender. Under the proportional hazards

assumption, the covariates affect the hazard function in a multiplicative fashion. The hazard function given covariate values is derived by multiplying a function of the covariates values by the baseline hazard function. The Cox proportional hazards model uses the assumption of proportional hazards to estimate the effects of the covariates without specifying the baseline hazard function.

In the Cox proportional hazard model, the hazard function at time *t* for the *i*th subject (*i*=1,..,*n*) is assumed to be:

$$\lambda(t \mid \boldsymbol{Z}_i) = \lambda_0(t) e^{\boldsymbol{\beta}^T \boldsymbol{Z}_i}, \qquad (1.1)$$

where $\boldsymbol{Z}_i$ is a vector of covariates, $\boldsymbol{\beta}$ is a vector of regression coefficients and $\lambda_0(t)$ is the unspecified baseline hazard function.

For an subject characterised by the covariates $\boldsymbol{Z}_i$ the ratio of his/her hazard to the baseline hazard is $e^{\boldsymbol{\beta}^T \boldsymbol{Z}_i}$. This relationship can be written as:

$$\log\left\{\frac{\lambda(t)}{\lambda_0(t)}\right\} = \boldsymbol{\beta Z}_i \qquad (1.2)$$

Parameter Estimates from Cox Proportional Hazards Model

Parameter estimates in the Cox proportional hazards model are obtained by maximizing the partial likelihood rather than the full likelihood. This is in order to eliminate the unknown baseline hazard function and to account for censored values. When the sample contains ties, an approximation of the partial likelihood is used. The confidence interval for the hazard ratio or the parameter estimate will generally provide useful information in addition to the p-value.

Let the sample comprise of *r* uncensored times $t_1 < \ldots < t_r$. The remaining *n*-*r* subjects are right censored. Let *i* denote the subject failing at t_i and *j* denote the subject failing at time t_j. $\boldsymbol{Z}_i$ and $\boldsymbol{Z}_j$ are the vectors of covariates for these subjects respectively. The log partial likelihood (Cox 1972) is given by:

$$l(\beta) = \sum_{i=1}^{r} \left\{ Z_i^T \beta - \log\left[\sum\nolimits_{t_j \geq i} \exp(Z_j^T \beta) \right] \right\} \qquad (1.3)$$

The treatment effect is often presented as a hazard ratio. In considering a binary (active vs placebo) treatment effect, the hazard ratio (active vs placebo) is the exponential of the parameter of the treatment effect. It is common in breast cancer trials that the hazard ratio corresponding to the minimum treatment effect of clinical interest is considered in the sample size calculations. A parametric form may be assumed as part of the sample size calculations.

The Assumption of a Parametric Form for the Baseline Hazard Function

The baseline hazard function is not specified in the Cox proportional hazards model. However, the proportional hazards model may be fitted by assuming the baseline hazard function follows a parametric form. For example, an exponential form for the baseline hazard function may be assumed during the sample size calculations in a breast cancer trial, as was done in the HERA trial. The hazard function is assumed to be constant. This assumption implies that the survival times follow an exponential distribution rescaled by the covariates. A generalisation of the exponential distribution allows for a power dependence of the hazard on time. This yields the two-parameter Weibull distribution.

Stratified Cox Proportional Hazards Model

In a clinical trial, a stratified analysis may be used to account for prognostic factors rather than including them as covariates in the survival model. The stratified Cox proportional hazards model allows the form of hazard function to vary across levels of stratification variables. Suppose *X* is a secondary categorical predictor, with *Q* levels, that we want to adjust for when making inferences about the relationship between the covariates $\boldsymbol{Z}$ and the outcome. Then a stratified analysis can be performed by fitting

$$\lambda_q(t|\boldsymbol{Z}_i, X) = \lambda_{0q}(t)e^{\boldsymbol{\beta}^T Z_i}, q = 1, .., Q \qquad (1.4)$$

The baseline hazard functions for the *Q* strata are allowed to be arbitrary. The parameter estimates are again found by maximising the partial likelihood. In

general, the approximate partial likelihood of β is the product of the Q partial likelihoods arising from the qth stratum alone (Kalbfleish and Prentice 2002, Chapter 4).

Time-dependent Covariates

It is becoming increasingly common in breast cancer trials to have covariates measured throughout the study and thus are defined as time-dependent covariates. Examples of time-dependent covariates in breast cancer trials are covariates which may be related to efficacy, quality of life assessments and medical assessments performed to monitor patient safety repeated throughout the study.

Let D be a time-dependent covariate and let

$$\bar{D}_i(t) = \{D_i(u);\ 0 \leq u < t\} \qquad (1.5)$$

denote the covariate history up to time t for the ith patient.

In the time-dependent Cox proportional hazard model, the derived covariates $Z_i(t)$ are functions of the covariate history $D_i(t)$ and the time t. An internal covariate is defined as a value over time, such as a quality of life assessment, generated by the subject under study (Kalbfleish and Prentice, Chapter 6.). The parameter estimation for such covariates is based on the partial likelihood score function:

$$U(\beta) = \sum_{i=1}^{n} \delta_i \left\{ D_i(t) - \frac{\sum_{t_j \geq t_i} \exp\left(\beta^T D_j(t_i)\right) D_j(t_i)}{\sum_{t_j \geq t_i} \exp\left(\beta^T D_j(t_i)\right)} \right\} \qquad (1.6)$$

where δ_i indicates, by values 1 versus 0, whether the ith subject fails or is censored at time t_i.

Competing Risks

In the analysis of some time to event endpoints, the possibility that the patient has another type of event which prevents the occurrence of the event of interest has to be considered. An example is when considering death due to breast cancer, death due to other causes is a competing event. In the competing risks analogue of the

Cox proportional hazard model (Holt 1976), the cause-specific hazard of cause *c* for patient *i* with covariate vector Z_i is modelled as

$$\lambda_c(t \mid \boldsymbol{Z}_i) = \lambda_{c,0}(t) e^{\boldsymbol{\beta}_c^T \boldsymbol{Z}_i}, \qquad (1.7)$$

where $\lambda_{c,0}(t)$ is an unspecified baseline cause-specific hazard function and the vector $\boldsymbol{\beta}_c$ represents the covariate effects on cause *c*.

The calculation of cause-specific hazard ratio follows standard methods. An important assumption is that the failure processes of the competing risks are independent (Piantadosi 2005, p. 199). Therefore, the interpretation of cause-specific hazard ratios, such as the specific hazards of a cardiac endpoint, requires caution.

1.6 Missing Observations in Longitudinal Data

Longitudinal data refers to variables measured repeatedly throughout a study. As expected given that longitudinal data involves the patient returning repeatedly for an assessment, missing observations are common (Diggle et al. 2002, p.21). These missing observations complicate the analysis of longitudinal data. Thus, methods for dealing with missing longitudinal data, such as imputation, have been proposed in the statistical literature. In the analysis of longitudinal data, repeated measures models (e.g. Searle 1971) can be used to account for the structure of the measurements. This section begins by considering the problem of missing data in longitudinal data. Next, methods to deal with missing data and its prevalence are described. The last part of this section describes repeated measures models and a method for exploring the relationship among imputed values.

1.6.1 Problem of Missing Data, Missing Data Mechanism and Prevention of Missing Data

Problems Arising From Missing Observations

The potential problems associated with missing observations assessments include loss of power to detect clinically important differences among treatment groups or over time. Generally, the sample size of breast cancer clinical trials is based on DFS or OS and the sample size will be large enough to detect clinically important differences in quality of life assessments or patient safety. However, the reduction in power due to missing observations could potentially lead to the failure to detect a clinically important difference.

Another problem associated with missing observations is the potential bias of the parameter estimate. If the probability of missingness is associated with the unknown value or with covariates not considered in the analysis model, then the parameter estimate will be over- or underestimated (Fairclough 2010, p.149).

Missing Data Patterns

The missing data pattern describes which values of variables in the dataset are observed and which are missing. The first missing data pattern to be considered was univariate missing data when only a single variable has missing observations. An assessment repeated throughout a clinical trial is an example of information collected repeatedly in a longitudinal study and may have a monotone missing data pattern. This refers to situation where all assessments were observed up until the time the patient was lost to follow-up (dropped out) and then no further assessments are observed (Little and Rubin 2002, Chapter 1).

It is common for missing observations from a clinical trial to have a general missing data pattern. When a patient has intermittent missing observations in assessments repeated throughout the study, this is an example of a general missing data pattern (Little and Rubin 2002, Chapter 1). Methods for dealing with missing monotone data may be easier to apply than methods for dealing with a general

missing data pattern. These missing data patterns are illustrated below in Table 1.1:

Table 1.1 Examples of Missing Data Patterns

A (Monotone)

Patient	Time 1	Time 2	Time 3	Time 4
1	O	O	O	O
2	O	O	O	
3	O	O	O	O
4	O			
5	O	O		

B (General)

Patient	Time 1	Time 2	Time 3	Time 4
1	O	O	O	O
2	O	O	O	
3	O	O	O	O
4	O	O		
5	O		O	O

The spaces corresponding to missing observations are highlighted in blue;
O corresponds to observed

Missing Data Mechanism

There are three major categories of missing data, depending on whether the reason for the observation being missing is related to the patient's quality of life (Rubin 1976; Little and Rubin 1987, Chapter 1; Little and Rubin 2002, Chapter 1).

i) Missing Completely at Random (MCAR)

For a quality of life assessment to be MCAR, the assumption is that the probability that an observation is missing is independent of the unknown value of the missing quality of life assessment and values of quality of life assessments at other times. Fielding et al. (2009) note that a scenario for MCAR quality of life assessment is if the form was lost in the post after being sent by the patient.

ii) Missing at Random (MAR)

For a quality of life assessment to be MAR, the assumption is that the probability that an observation is missing is independent of the unknown value of the missing

quality of life assessment but may depend upon the covariates and quality of life assessments made at other times. Fielding et al. (2009) note that a scenario for MAR quality of life assessments is if patients with poorer baseline quality of life are less likely to complete subsequent assessments than patients with better baseline quality of life. If quality of life assessments are MAR, then parameter estimates based on the analysis of patients with complete data (complete case analysis) will be biased (Little and Rubin 2002, Chapter 3; Molenberghs and Kenward 2007, Chapter 4).

Unlike complete case analysis, available case analysis considers all available values. Fairclough (2010, Chapter 6) notes that it is possible to obtain unbiased parameter estimates from available case analysis of quality of life assessments. When the missing data mechanism is ignorable, unbiased estimates of the parameters $\boldsymbol{\theta}$ can be obtained from likelihood based methods using all observed data and covariates associated with the probability of missingness. Little and Rubin (2002, Chapter 6) defined when the missing data mechanism is ignorable, as outlined in the paragraph below.

The complete data Y is portioned into observed values, Y^{obs}, and missing values, Y^{miss}, and the missing-data indicator matrix M indicates the missing data pattern. They formulate models in terms of a probability distribution of Y with the density $f(Y|\theta)$ indexed by unknown vector parameter θ, and a probability distribution function $f(M|Y, \psi)$ for M given Y indexed by a vector parameter ψ. ψ is the distribution of the missing data mechanism. The missing data mechanism is ignorable for likelihood inference if:

a) the missing data are MAR;
b) the parameters θ and ψ are distinct, in the sense that the joint parameter space of (θ, ψ) is the product of the parameter space of θ and the parameter space of ψ. (Therefore the parameters θ are functionally independent of the missing data mechanism)

iii) Informative Missing Data (or Missing Not At Random)

If the probability that an observation is missing depends on the unknown value of the missing quality of life assessment, then this quality of life assessment is informative missing data. The French 2003 decennial health study illustrates informative missing data from a quality of life questionnaire. Peyre et al. (2010) note that for half of items on the of the SF-36 questionnaire, low scores on their subscale was associated with missingness of the item. Therefore, these items were considered informative missing data. If quality of life assessments are informative missing data, then parameter estimates based on complete case and available case analysis will be biased (Little and Rubin 2002, Chapter 3).

Selection Models and Pattern Mixture Models

Schafer and Graham (2002) note that to model missing data without the MAR assumption, a distribution for the missingness must be explicitly specified as well as the model for the complete data. The two different ways to do this are selection models, reviewed by Little (1995), and pattern mixture models (see section 2.3.6). In typical applications of selection models, it is assumed that measurements over time ($Y_1, \ldots, Y_T$) follow a well-recognised distribution such as a multivariate Normal and allow the probability of dropout at occasion t to follow a logistic regression on the previous and current values ($Y_1, \ldots, Y_t$) but not on future values (Schafer and Graham 2002).

Identification of Informative Missing Data

As the values of the missing observations are unknown, it is not possible to formally test a hypothesis that the probability that an observation is missing is independent of the unknown value. Formal comparisons of MAR versus the alternative informative missing data should be considered with caution (Jansen et al. 2006; Schafer and Graham 2002). For example, in selection models, the logistic coefficients for the drop-out model can be set to give special cases of MCAR and MAR. This allows the possibility of testing the MAR hypothesis by considering the confidence interval for the coefficient (Schafer and Graham

2002), but the results of such tests depend on untestable assumptions about the population distribution (e.g. Kenward 1998). However, it is possible to investigate whether the assumption that data are MAR is reasonable, though the evidence should be interpreted carefully (Molenberghs and Kenward 2007, p.185). For example, if the probability that an observation is missing is associated with the fact the patient has a poor quality of life due to suffering a serious adverse event, then it is unlikely that the assumption that the quality of life assessment is MAR is valid.

Preventing Missing Observations

Even though statistical methods for dealing with missing data exist, it is important to minimise the amount of missing data in a clinical trial. This is of particular importance when assessing quality of life, where often there is a high proportion of missing data. Potential factors contributing to missing quality of life assessments are: i) the assessments seem burdensome to a patient focused on his or her disease and ii) inadequate processes for collecting the assessments.

Consideration should be given during the design stage of a study to the schedule and frequency of study assessments to minimise the burden on patients. When possible, quality of life assessments should be scheduled at the same time as other medical assessments. It is important that the study procedures are clearly described in the protocol. This should include details about how to complete quality of life assessments if the patient's treatment does not follow the protocol schedule or if the patient is unable to complete the assessment without assistance (Fairclough 2010, Chapter 2).

The quality of life questionnaire should be single-sided. A patient information sheet describing the reason for quality of life endpoints and the reason why completing the full quality of life assessment is important may help reduce missing quality of life assessments. However, the patient must be aware that his or her medical treatment will not be affected if he or she decides not to complete the

quality of life assessment (Fairclough 2010, Chapter 2). As not all quality of life assessments will be completed, it is useful to record the reason the patient did not complete the quality of life assessment and if possible to record if the reason was related to the patient's quality of life (Fairclough 2010, p.46).

1.6.2 Dealing with Missing Data-Methods

Methods proposed for analysis of data with missing observations, can be grouped into the following categories, which are not mutually exclusive:

i) procedures based on completely recorded units
ii) imputation-based procedures
iii) weighting procedures
iv) model-based procedures

Overview of Methods

Imputation-based procedures, where missing values are replaced, are the focus of this thesis. Standard methods are described in Chapter 2. Following imputation, it may be of interest to explore the relationship among the imputed values. Hierarchical cluster analysis (e.g. Everitt [1993]) allows this relationship to be explored by constructing a dendrogram, a tree-like structure describing the distance between the observations.

Complete case analysis may be satisfactory when there are only small amounts of missing data. However, it may lead to substantial bias and lacks efficiency (Little and Rubin 2002, Chapter 3). Randomisation inferences from sample survey design without missing data commonly weight sample units by their design weights, which are inversely proportional to their probability of selection. Weighting procedures for missing data modify the weights with the aim of adjusting for missing data as if it were part of the sample design (Little and Rubin 2002, p.19-20).

Model based procedures are a broad class of procedures. A model is defined for the observed data and inferences are based on the likelihood or posterior distribution under the defined model, with parameters estimated by procedures such as maximum likelihood. An example is the expectation maximisation (EM) algorithm (Dempster et al. 1977), which has an expectation step and maximisation step. Next in this subsection, two different types of missing data, namely missing outcome data and missing explanatory variables, are considered.

Missing Outcome Data in Clinical Trials

Clinical trials are usually designed to allow statistical analysis which can be meaningfully interpreted to be performed using straightforward calculations. When considering standard classical designs, there is a standard least squares analysis, which gives estimates of parameters, standard errors for contrasts of parameters and the analysis of variance (ANOVA) table.

Consider a missing data pattern where the covariates $\boldsymbol{X}$ are fully observed and the outcome variable Y contains missing observations. In this situation and when the MCAR assumption holds, the patients with Y missing provide no information for the regression of Y on $\boldsymbol{X}$ and analysis of the complete cases is fully efficient. However, the balance of the original design is lost and more complex calculations are required to compute the correct least squares analysis (Little and Rubin 2002, Chapter 2).

The advantages of imputing the missing data in a clinical trial rather than analysing the available data include the fact the required statistical analysis is easier to interpret and compute as standard statistical methods can be applied. The aim is to have simple rules for imputing data in order to achieve one or more completed dataset(s) which represents the unknown complete data.

Assuming the probability of missingness is unrelated to the missing values of Y and so the MAR assumption holds, there are several methods for imputing

missing data that lead to correct estimates of all parameters which can be estimated. In addition, it is simple to correct the residual (error) mean square, standard errors and sums of squares that have one degree of freedom. Though more complicated to compute, it is also possible to provide correct sums of squares with more than one degree of freedom. This information yields the ANOVA table (Little and Rubin 2002, Chapter 2). Imputation methods have also been developed to deal with informative missing data. However, as the missing values of *Y* are unknown, these methods often make assumptions that cannot be tested with the available data.

Missing Explanatory Variables versus Missing Outcome Data

An example of missing explanatory variables is where, in a clinical trial, treatment group and stratification variables are fully observed but some other variables, such as the patient's weight at randomisation, are missing. The missing covariates may be part of a general linear model. The basic statistical assumption underlying general linear modelling is that the outcome variable can be described as the sum of a fixed component which is a linear function of the explanatory variables and a random error component. In the scenario where some patient's weight at randomisation is missing, complete case analysis would exclude these patients, allowing standard statistical analysis to be applied. The disadvantage would be loss of precision and potential bias due to excluding patients. Bias would arise when the missing data mechanism is not MCAR and so the complete patients are not a random sample of all patients. It is likely that the balance in a treatment comparison would be lost making the analysis less efficient. Imputation would generally be useful in this scenario. Considering covariates that are part of a general linear model, methods for dealing with missing covariates are generally applicable with minor modifications to missing outcome data (Wang and Chen 2001).

A challenge in dealing with missing covariates in the context of non-linear regression is that the likelihood-based score is not a linear function of the

covariates while it is linear function of the outcomes. This non-linearity property makes dealing with missing covariates more complicated than dealing with missing outcome data. Among the complications is modelling the covariate distribution when the maximum likelihood estimator is considered (Wang and Chen 2001). For survival data, further challenges in dealing with missing covariates are added due to censoring. For example, the conditional probability of missing covariates given observed covariate and outcome data is generally a function of the unknown baseline hazard function and some covariate distributions (Wang and Chen 2001).

Missing Explanatory Variables in Survival Data

Assuming the probability of missingness is unrelated to the unknown values of the outcome variable Y, though may depend on completely observed covariates, estimates of β from the Cox proportional hazards model based on the complete case analysis is unbiased but lacks efficiency (Paik 1997). If data are completely observed, the full likelihood function contains an unspecified baseline hazard function making the estimate of β difficult to obtain and therefore the partial likelihood is considered. The maximum partial likelihood estimator of β is the solution to the log-partial likelihood equation involving derivatives with respect to β equalling 0 (Cox 1972; Andersen and Gill 1982). When covariates of a patient with an event are missing, a component of the log-partial likelihood is unknown. The baseline survival function, $S^{(0)}(\beta, X_i)$, cannot be calculated if any patients in the risk set have missing covariates.

Methods of dealing with missing covariates in the Cox proportional hazards model have been proposed. These include methods considering the baseline cumulative hazards and weighted estimators (e.g. Wang and Chen 2001; Qi et al. 2005; Bang and Robins 2005; White and Royston 2009), and methods when the covariates are missing at random (e.g. Lin and Ying 1993; Pugh et al. 1993; Zhou and Pepe 1995; Paik and Tsai 1997; Paik 1997; Zhou and Wang 2000) or when the missing covariates are informative missing data (e.g. Leong et al. 2001;

Herring et al. 2004). Methods for dealing with missing time-dependent covariates have also been proposed (e.g. Altman and De Stavola 1994; Collett 1994; Dupuy and Mesbah 2002; Bradshaw et al. 2010).

Augmented Inverse Selection Probability Weighted Estimators and Fully Augmented Weighted Estimators

Wang and Chen (2001) proposed an augmented inverse selection probability weighted (AIPW) estimator for parameter estimates of MAR covariates. The augmentation term of the parameter estimator depends on the baseline cumulative hazard and on a conditional distribution implemented by using an EM-type algorithm. The method proposed was developed considering missing covariates in survival data and extends the inverse probability weighted estimator proposed by Horvitz and Thompson (1952).

The complete case analysis is adjusted to give the simple inverse probability weighted (SIPW) estimator by using the inverse of the selection probability π_i as the weight. If π_i is unknown, it has to be estimated. The SIPW is simple to implement but may be inefficient. Adding an augmented term to the simple weighted estimating equation was discussed in Robins et al. (1994) in a general framework. The AIPW estimator is derived from the augmented inverse AIPW estimating equation and can be implemented by an EM-type algorithm.

Related to this AIPW estimator, Qi et al. (2005) proposed a kernel-assisted fully augmented weighted (FAW) estimator where both the selection probabilities and conditional expectation of the unobserved covariate are estimated non-parametrically. Here, non-parametric kernel smoothing techniques are adopted to estimate conditional expectations that depend on the cumulative baseline hazard function and the conditional distribution of the missing covariates given the observed covariates.

Doubly Robust Estimators

An estimator from a missing data model is doubly robust if it remains consistent when either a model for the missing data mechanism or a model for the distribution of the complete data is correctly specified. Scharfstein et al. (1999) showed that the orthogonal AIPW estimator proposed by Robins et al. (1994) and Rotnitzky et al. (1998) was doubly robust. They developed a general method to construct doubly robust estimators when data are missing at random. They also showed the orthogonal AIPW estimator had an alternative "regression representation". Bang and Robins (2005) extended previously developed methods in order to construct doubly robust estimators in longitudinal monotone missing data models. The method assumes that the data are MAR, and here $\mathbf{L} = \bar{\mathbf{L}}_{J+1} = (\mathbf{L}_1^{\mathrm{T}}, \ldots, \mathbf{L}_{J+1}^{\mathrm{T}})^{\mathrm{T}}$ represents the full data obtained at times $h = 1, \ldots, J+1$.

Suppose the parameter of interest μ is the mean of $Y = L_{J+1}$. The parameter of interest μ can be expressed in terms of regression functions defined recursively. Bang and Robins (2005) show that it not necessary to specify a parametric model for the entire joint distribution of **L**, Instead, parametric models are specified for the regression functions. The regression parameters are then estimated recursively based on the observed data.

Logistic or Linear Regression using the Censoring Indicator, the Cumulative Baseline Hazard and Other Covariates

White and Royston (2009) proposed logistic or linear regression using the censoring indicator, the cumulative baseline hazard and the other covariates as a suitable model for imputing missing binary or Normally distributed covariates. They considered regression analysis when covariates have missing data. They noted multiple imputation, which involves replacing each missing observation with K ($K > 1$) simulated values (Rubin 1987; Little and Rubin 2002), is typically more efficient than complete case analysis in this scenario. A regression model, referred to as the imputation model, may be used in multiple imputation, as opposed to the analysis model whose regression coefficients are of interest. In

survival data, it is common to include the censoring indicator and the log of the survival time in the imputation model. The choice of variables in the imputation model is very important in performing multiple imputation appropriately. They derived a number of exact and approximate results about the imputation model involving the censoring indicator and complete covariates in terms of the model parameters θ. These results motivate the regression models involving the censoring indicator, the cumulative baseline hazard and the complete covariates.

The parameter estimator is exact in the case of a single binary covariate and in other situations it is approximately valid for small covariate effects and/or small cumulative incidence. The method proposed assumes the missing covariates are MAR or MCAR. In addition to the Cox proportional hazards model for the survival time, an "exposure model" is also required to account for the missing covariates. The model parameters θ considered contain the parameters from the "exposure model", the regression coefficients to be estimated and the cumulative baseline hazard.

Covariates Missing At Random

Parameter estimators dealing with MAR covariates include Lin and Ying (1993), Pugh et al. (1993), Zhou and Pepe (1995), and Zhou and Wang (2000). Paik (1997) considered imputation of the missing covariate to give a completed dataset. However, to avoid specifying the baseline hazard function, estimates of β following imputation of the missing covariate were derived by adapting the methods proposed by Zhou and Pepe (1995) and Paik and Tsai (1997).

Informative Missing Data

Methods based on maximising a semiparametric likelihood such as Martinussen (1999) and Chen and Little (1999) do not address informative missing data with missing categorical and continuous covariates. The extension of parameter estimators of Lipsitz and Ibrahim (1998) by Leong et al. (2001) applies to informative missing data when the missing covariates are binary. Considering a

more general missing covariates situation, Herring et al. (2004) proposed a partial-likelihood-based method of parameter estimation which uses an EM approach. The method proposed by Herring et al. (2004) does not require the assumption the missing covariates are MAR and is general in the type and number of missing covariates.

Methods for Dealing with Missing Time-Dependent Covariates

It is possible that time-dependent covariates have missing observations. This is likely when a covariate is assessed at a number of study visits and a patient does not attend the study visit or the assessment is not done at the study visit. Consider a clinical trial which is assessing survival and where a time-dependent covariate *D* is measured at discrete times until a terminal event occurs. There may be some sequences of measurements that end prematurely, truncating the covariate history of *D*, and this is referred to as dropout (Diggle and Kenward 1994; Little 1995; Scharfstein 1999). The Cox model is a standard analysis to evaluate the relationship between time to dropout and the time-dependent covariate *D*. When performing this analysis, the assumption is that changes in the value of *D* occur at the times t_j of measurement of *D* and that the value of *D* during the interval [t_j, t_{j+1}] is observed at the end of the interval. Therefore, the value of *D* is missing at the time of dropout.

Altman and De Stavola (1994) and Collett (1994) propose imputing the missing value of *D* at the time of dropout with the last observed value. However, this approach is not appropriate if *D* changes in the instants before dropout. Dupuy and Mesbah (2002) propose a joint modelling approach for dropout time and longitudinal covariate data. This model allows possible changes in *D* just before dropout and can applied when dropout is dependent on the unknown value of *D*. Parameter estimation is performed using the EM-algorithm. Bradshaw et al. (2010) extends the approach to parameter estimation proposed by Herring et al. (2004) to allow time-dependent covariates. The selection model proposed also allows covariates to be informative missing data. It is defined by the joint

distribution of the event times, missing covariates and the missing data mechanism.

Summary

Missing data can be missing outcome data or missing explanatory variables. Dealing with missing covariates in survival data is more complex than dealing with missing outcome data. Methods of dealing with missing covariates in the Cox proportional hazards model have been proposed. These methods include weighted estimators and generally do not consider imputation. The methods extend to missing time-dependent covariates and informative missing data and the combination of both. In this thesis, time-dependent covariates that may be informative missing data are considered. Imputation in this context has not been studied in detail in the literature and is the focus on later chapters in this thesis.

1.6.3 Prevalence of Missing Data and Imputation Methods Applied in Reports of Clinical Trials

In a large clinical trial, some missing data is likely to occur even when the study is well designed and well conducted. The previous section, section 1.6.2, reviewed some methods to deal with missing data. In this section, there is a review of how missing data is dealt with in practice in publications in the statistical literature. The review is based on the Journal of Clinical Oncology, which is a high-impact, widely cited oncology journal. Its impact factor reported in the 2010 Journal Citation Reports® (Thomas Reuters [2011]) was 18.970 and it ranks 4th among oncology journals in number of citations. To investigate the prevalence of missing data in clinical trials and the variety of methods for dealing with missing data in journal articles describing the results of clinical trials, 209 original reports describing the results of the clinical trials in the Journal of Clinical Oncology recently published between July and December 2006 were reviewed in 2007.

There were 136 out of 209 (65%) of articles reporting clinical trials where there was a small number of patients (<10%) with missing baseline or demographic

characteristics included in the efficacy analysis but where there was no attempt to deal with the missing data and the patients with missing data were excluded from the efficacy analysis. In addition, there were 15 articles reporting clinical trials where there was a small amount of missing demographic or baseline data which was not used in the efficacy analysis.

It was also noticeable that 19 out of 53 articles reporting clinical trials in August 2006 excluded a small number of patients who did not meet eligibility criteria or could not be assessed for the primary outcome, for example because the patient did not have a suitable tumour biopsy.

Imputation Methods Applied in Articles in the Journal of Clinical Oncology: 2006

Among the articles in Journal of Clinical Oncology reviewed, there were three articles that applied imputation to deal with missing data: Kudoh et al. (2006), Siemes et al. (2006) and Stiff et al. (2006).

In Kudoh et al. (2006), the missing quality of life assessments were assigned as unimproved (extreme imputation) in order to compare the results with the results considering available data.

In Siemes et al. (2006), a Cox proportional hazards model was built with age, sex and separate potential risk factors as covariates. Missing indictors were used to study the effect of missing observations. Missing data on covariates were imputed by a single simulated value using the EM algorithm proposed by Horton and Laird (1999).

In Stiff et al. (2006), in the primary analysis, area under the curve (AUC) was calculated over the duration of the study for all patient-reported outcomes end points. Missing assessments at the start day or end day were imputed with the nearest non-missing assessment. When the AUC could not be calculated, the AUC

was imputed by the grand mean AUC value or worst AUC value among patients with the same type of haematological disease.

Recent Use of Multiple Imputation in Journal Articles and Missing Data in Articles in the Journal of Clinical Oncology in 2011

Sterne and White (2009) consider recent use and reporting of multiple imputation in four major general medical journals: New England Journal of Medicine, Lancet, British Medical Journal and Journal of the American Medical Association. They found that the use of multiple imputation roughly doubled between 2002 and 2007. In line with these findings, the issue of missing data was addressed more often in articles reporting clinical trials in the Journal of Clinical Oncology between January and June 2011 than in 2006. This was done by describing i) the assumption about the missing data mechanism, ii) how patients with missing observations were considered, or iii) the reason why no measures to account for missing data were taken. However, in contrast to findings of Sterne and White (2009), multiple imputation was only applied in one article, Cooperberg et al. (2011).

The assumptions about missing data mechanism were described in two articles (Kornblith et al. 2011 and Syrjala et al. 2011). Two articles (Grothey et al. [2011] and Osborne et al. [2011]) described how patients with missing observations are considered. The reason why no measures to account for missing data were taken was described in three articles (Ganz et al. 2011; Kitahara et al. 2011; Hurwitz et al. 2011). In Cooperberg et al (2011), the missing percent of biopsies cores positive in prostate cancer patients was replaced by multiple imputation in order to classify patients as low or high risk when analysing the primary outcome of cancer progression. The multiple imputation method was not described.

Findings from the Review

It was common in the articles reviewed for there to be a small number of patients (<10%) with missing baseline or demographic characteristics included in the

efficacy analysis but where there was no attempt to deal with the missing data. Multiple imputation was rarely used. However, between 2006 and 2011 it had become far more common for the articles to describe missing data and the related assumptions.

1.6.4 Modelling Longitudinal Data with Missing Observations Data

Missing data in clinical trials can be from longitudinal data as assessments, such as quality life assessments, may be repeated throughout the study. Generally, when an assessment is repeated throughout the study, modelling the result of the assessment over time is likely to be clinical interest, and is described in this subsection. Repeated measures models such as the mixed effects model, are used in the analysis of longitudinal data to account for the structure of the measurements. These models accommodate missing observations by considering patients who do not have a complete set of measurements of the assessment.

Repeated Measures Models

In a repeated measures model, time is conceptualised as a categorical variable. Each measurement of an assessment is assigned to one category (Fairclough 2010, p.53). Possible models for repeated measures include the general linear model and the mixed-effects model.

As previously noted, the general linear model assumes that the outcome variable can be described as the sum of a fixed component which is a linear function of the explanatory variables and a random error component. The random error component $\boldsymbol{\varepsilon}$ is assumed to be a vector of independent identically distributed Normal random errors (Searle 1971, Chapter 3). The mixed-effects model extends the general linear model by allowing flexibility in the specification of the covariance matrix of $\boldsymbol{\varepsilon}$ (Searle 1971, Chapter 9). Its historical development is described by Henderson (1990) and Searle et al. (1992).

The mixed-effects model is written as:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Z\omega} + \boldsymbol{\varepsilon} \quad (1.8)$$

where $\boldsymbol{Y}$ is the variable of interest

$\boldsymbol{X}$ is the known matrix of explanatory variables

$\boldsymbol{\beta}$ is the unknown fixed-effects parameters vector

$\boldsymbol{Z}$ is a known design matrix

$\boldsymbol{\omega}$ is an unknown random-effects parameters and

$\boldsymbol{\omega} \sim N(\mathbf{0}, \mathbf{G})$

$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{R})$

and **G** and **R** are unknown variances of a multivariate Normal distribution.

The variance of $\boldsymbol{Y}$ is

$$\mathbf{V}=\mathbf{ZGZ'} + \mathbf{R} \quad (1.9)$$

The mixed-effects model accommodates correlation and heterogeneity in variances (Searle 1971, Chapter 9).

Fairclough (2002, Chapter 3; 2010, Chapter 3) describes an analytic approach to modelling repeated measures which incorporates both incomplete data and time-dependent covariates. While the model is not strictly a mixed-effects model, it has been associated with the term (Fairclough 2010, p.54). Pinheiro and Bates (2000) note that the model can be thought of as an extended linear mixed-effects model with no random effects. The model can be expressed as:

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \quad (1.10)$$

where $\boldsymbol{Y}_i$ is the full data vector of H planned measurements of the variable of interest, which includes the observed values and the missing values

$\boldsymbol{X}_i$ is the design matrix of fixed covariates corresponding to the complete data ($\boldsymbol{Y}_i$)

$\boldsymbol{\beta}$ is the corresponding vector of fixed-effects parameters

$\boldsymbol{\varepsilon}_i$ is the vector of residual errors

Σ_i is the covariance of the full data ($\boldsymbol{Y}_i$), which is a known function of the vector of unknown variance parameters, τ

The process for building this model involves defining a model for the means ($\boldsymbol{X}_i\boldsymbol{\beta}$) and identifying the structure of the covariance of the $\boldsymbol{Y}_i$ (Fairclough 2002, p.44). The most straightforward structure for the means is the cell mean model (Searle 1971, Chapter 7). The cell means model of interest is written as:

$$Y_{gih} = \mu_{gh} + \varepsilon_{gih} \qquad (1.11)$$

where Y_{gih} is the hth measurement of the ith patient in the gth treatment group μ_{gh} is the average value of the hth measurement in the gth treatment group.

The covariance may be unstructured or structured. In the unstructured covariance, the variance of the measurement at each time point is allowed to be different. It is the least restrictive of the covariance structures and is generally the best choice when the number of repeated measure is small (Fairclough 2010, p.68.-69).

This section considered missing observations in longitudinal data. This included i) the problem of missing data in longitudinal data, ii) methods to deal with missing data, such as imputation, and iii) its prevalence. Next, the thesis outline is presented.

1.7 Thesis Outline

Motivation

This chapter began by considering general features of breast cancer clinical trials. High-quality evidence from clinical trials on the benefits and risks are required to make clinical judgements on treatment regimens for breast cancer. The assessment of benefit in breast cancer trials is generally based on DFS and OS. Clinical trials may also include endpoints which assess the patient's perception of his or her well-being. Often such quality of life assessments are repeated throughout the study. It is common for medical assessments which monitor patient safety, such as monitoring cardiac function, to be repeated throughout the study. Generally when assessments are repeated throughout the study, some missing observations are expected.

The breast cancer clinical trials considered in this thesis were introduced. In these clinical trials, quality of life or cardiac function measured throughout the study was of importance. Therefore, quality of life and the main aspects of breast cancer trials, efficacy and safety, were described. The standard analysis for time to event endpoints, the Cox proportional hazard model, and parameter estimation was described. Explanatory variables measured throughout the study may be included as time-dependent covariates in a time-dependent Cox model analysis.

Next, missing observations in longitudinal data were considered. The potential problems associated with missing observations, such as missing quality of life assessments, include bias of the parameter estimates and loss of power to detect clinically important differences among treatment groups over time. Methods for dealing with missing observations, such as imputation techniques, have been described in the statistical literature. In this thesis, the influence of missing explanatory variables is explored, with the application to the analysis of quality of life and cardiac function in breast cancer clinical trials. Imputation techniques are applied to missing explanatory variables before the time-dependent Cox model analysis, and the performance of these techniques considered. The influence of missing observations of an outcome variable assessing cardiac function is investigated and repeated measure analysis of cardiac function performed.

Structure of Thesis

Standard imputation methods are described in Chapter 2. Missing coping scores in the IBCSG dataset are imputed by standard simple and multiple imputation methods, in Chapters 3 and 4 respectively, in order to use the coping score as a time-dependent covariate in a time-dependent Cox model for DFS. Hierarchical cluster analysis is performed in order to explore the relationship between imputed values.

Imputation methods involve assumptions about the missing data mechanism. Simulation is performed with the aim of investigating the influence of the missing data mechanism resulting from the method of artificially removing data on the performance of standard imputation methods. The context of the simulation is a positive relationship between quality of life and DFS. Standard simple and multiple imputation methods are applied to the simulated datasets and this is described in Chapters 5 and 6 respectively. Following imputation, the coping scores are used as a time-dependent covariate in a time-dependent Cox model for DFS.

The last part of the thesis, Chapter 7, considers missing observations in LVEF assessments performed throughout the study as part of the cardiac monitoring in the HERA trial. The occurrence of a noticeable drop in LVEF from baseline is used as a time-dependent covariate in a time-dependent Cox model for time to a cardiac endpoint. The further cardiac analyses also investigates the influence of missing LVEF assessments on the change in LVEF from baseline by applying multiple imputation to missing LVEF values. A mixed model is used in the repeated measures analysis of the LVEF values over time to model the patients' LVEF values over time. The conclusions from the thesis are presented in Chapter 8.

2 Standard Methods of Imputation

As noted in Chapter 1, high-quality evidence from clinical trials on the benefits and risks are required to make clinical judgements on treatment regimens for breast cancer. It is recognised that a patient's sense of well-being impacts on the patient's perception of whether the treatment is beneficial. Thus, it is becoming increasingly common for endpoints addressing quality of life to be included in clinical trials. As it is usual for quality of life assessments to be repeated throughout the study, quality of life may be used in a time-dependent covariate in a time-dependent Cox model to explore the question in breast cancer clinical trials of whether quality of life is related to prognosis.

Missing observations often present challenges in the statistical analysis of data which are collected sequentially over time, such as quality of life assessments (Fayers and Machin 2007, p.355). In some situations, imputation may be appropriate (Fairclough 2010, Chapters 8 and 9; Fayers and Machin 2007, chapter 15). A scenario where imputation of quality of life assessments is beneficial is when the patients answer most questions but not all questions on the questionnaire (Fairclough 2010, p.163). Imputation-based procedures have been proposed in the statistical literature with the aim of addressing bias in the analysis of such data where there are relatively large numbers of missing observations (e.g. Rubin 1987; Little and Rubin 2002; Molenberghs and Kenward 2007). The missing observations are replaced by a single plausible value (simple imputation) or with K ($K > 1$) simulated values (multiple imputation).

The influence of missing explanatory variables in time-dependent Cox model analysis is explored in this thesis, with the application to the analysis of quality of life and cardiac function in breast cancer clinical trials. Standard imputation methods are applied to missing explanatory variables before the time-dependent Cox model analysis. These standard imputation methods are described in detail in statistical textbooks (e.g. Little and Rubin 2002; Molenberghs and Kenward 2007). In order to apply imputation techniques in subsequent chapters, standard

imputation methods are reviewed in this chapter. Standard simple and multiple imputation methods are reviewed in sections 2.2 and 2.3 respectively. An example dataset of quality of life assessments from a breast cancer trial are used to illustrate the standard imputation methods. Future directions of multiple imputation are noted. A summary is presented in section 2.4.

2.1 Introduction

The main aim in developing imputation techniques is to try to avoid bias due to missing data. This bias can be in the estimated effect and also in the estimated standard error. Historically, simple imputation was used to replace missing observations in order to create a completed dataset for statistical analysis of longitudinal data or repeated measures (see section 1.6.4), such as multivariate analysis of variance (MANOVA) or multivariate analysis of covariance (MANCOVA). This was due to the fact that statistical software could not address incomplete cases and by default excluded them (Fairclough 2002, p.115). Simple imputation methods have limitations in that they tend to underestimate the variability in the imputed variable, but can give useful information about the sensitivity of the results to assumptions about missing data (Fairclough 2010, p.178-179).

Multiple imputation was first considered in the context of complex surveys which are used to create public-use datasets (e.g. Rubin 1987). Later, multiple imputation expanded into further areas, including observational data from public health research and clinical trials (Kenward and Carpenter 2007). Multiple imputation has been applied to missing covariates in survival data (e.g. Paik 1997; White and Royston 2009) and to quality of life assessments in breast cancer trials (e.g. Bordeleau et al. 2003; Stanton et al. 2005). Statistical software is increasingly facilitating the implementing of imputation techniques, for example the introductions to the *MI procedure* in SAS® version 9 (SAS Institute 2002-2008).

Whereas in simple imputation a single replacement value is drawn from an imputation model, in multiple imputation *K* replacement values are drawn. Many of the issues concerning multiple imputation relate to the construction and application of the imputation model. In particular, careful consideration must be given to the choice of variables in the model (Molenberghs and Kenward 2007, p.110-111). van Buuren and Groothuis-Oudshoorn (2011) note that the imputation model should:

i) Account for the process that created the missing data
ii) Preserve the relations in the data
iii) Preserve uncertainty about these relations

and note that issues that potentially could arise while imputing multivariate missing data (Y_1,..,Y_j) include:

i) Circular dependence can occur, where Y_1 depends on Y_2 and Y_2 depends on Y_1 because in general Y_1 and Y_2 are correlated
ii) The relation between Y_j and Y_{-j} could be complex, e.g. non-linear
iii) Imputation can create impossible value outside the range of the variable, can create impossible combinations of variables or destroy deterministic relations in the data (e.g. sum scores)

New forms of model selection and diagnostic procedures, for example to detect influential observations and the imputation of impossible values, may be required due to the increasing automation of multiple imputation using statistical software (Kenward and Carpenter 2007).

It is important to note that as imputation methods require untestable assumptions every effort should be made to minimise missing observations. When imputation is applied, the sensitivity of the results to the assumptions of the imputation method should be explored. This is a topic where further development is required (Kenward and Carpenter 2007).

Analyses of data following imputation are more efficient than complete case or available case analysis of clinical trials as information from patients with incomplete data is considered. The fact that the imputation model may be different from the analytic model of interest for the data being considered is an advantage in the context of clinical trials. For example, including additional covariates in the imputation model that are not necessary in the analytic model may make the MAR assumption appropriate. There is an increasing understanding of the potential of multiple imputation and its continued development remains of scientific interest (Kenward and Carpenter 2007). The chained equations method (van Buuren et al al. 1999; Raghunathan et al. 2001; Taylor et al. 2002) is prominent among the recent work on multiple imputation methods in the statistical literature. Kenward and Carpenter (2007) note a more formal justification of this method is of particular interest.

However, imputation is not always a necessary way of dealing with missing observations (Rubin 1996; Fairclough 2010, chapter 9). Standard statistical methods such as maximum likelihood estimation or regression models can be used if the missing data can be assumed to be MCAR or MAR and so imputing data is of more importance when there is informative missing data (Fairclough 2010, p.181). This is investigated in Chapter 6 where the influence of missing longitudinal assessments on major safety endpoints is discussed.

Missing observations complicate the analysis of quality of life assessments in breast cancer trials. As noted, the concern is that missing observations may result in bias in the parameter estimates. In the context of investigating the possibility that quality of life is related to prognosis, imputation techniques can be used to investigate the influence of missing observations of quality of life assessments. This is discussed in Chapter 3.

2.1.1 Example Quality of Life Assessment

As part of the quality of life assessment in Trials VI and VII, coping/perceived adjustment ("coping score") was assessed by a VAS ("How much effort does it cost you to cope with your illness?" [none – a great deal]; the Perceived Adjustment to Chronic Illness Scale [Pacis]) (Hürny et al. 1993). The coping score ranges from 0 to 100, with lower scores indicating better quality of life. The quality of life questionnaire also assessed mood and physical well-being, with a range from 0 to 100 and lower scores indicating better quality of life. The primary endpoint of the trial was DFS (see section 1.2).

A very small example dataset is used to illustrate the techniques (Table 2.1), both to demonstrate how the imputations are made and to highlight the differences among the different imputation methods. The example dataset is not meant to illustrate good estimation procedures or good imputation techniques, but to illustrate the methodology. Here, the blanks corresponding to missing observations are highlighted in blue. There are 5 patients in treatment group A and 5 patients in treatment group B. The coping score was measured at baseline (Time 1, approximately at randomisation) and the first 3 measurements at approximately 3-monthly intervals up to 9 months after randomisation were considered (Time 2 – Time 4). For 5 patients, DFS was censored (censoring indicator is 0). Potentially, other quality of life assessments such as the assessment of mood ("mood score") and physical well being ("physical score") could be used in the imputation of the coping score. Suppose that the mood scores and physical scores for the 10 patients in the example are as shown in Table 2.2.

Table 2.1 Example Coping Scores and Disease-Free Survival from a Breast Cancer Trial

Patient	Trt	Time 1	Time 2	Time 3	Time 4	DFS (days)	DFS (censored)
6	B	27	32	18	18	4031	1
47	B	69	85	49		778	1
456	B	50	50	10	20	5486	0
635	A	50	46			2193	1
828	A	9		17	2	5062	0
1099	A	19				5120	0
1304	A	50	2	21	16	4737	0
1728	B	79	59	40	17	296	1
2237	A	43	38	32		993	1
2509	B	51	50	60	56	3569	0

The blanks corresponding to missing observations are highlighted in blue;
Indicator for DFS is 0 = event, 1 = censored;
DFS= disease-free survival; Trt = Treatment

Table 2.2 Example Mood Scores and Physical Scores from a Breast Cancer Trial

Patient	Treatment	Quality of Life Domain	Time 1	Time 2	Time 3	Time 4
6	B	Mood	49	14	6	8
		Physical	27	17	4	3
47	B	Mood	66	90	43	81
		Physical	52	88	53	72
456	B	Mood	1	14	11	5
		Physical	4	15	8	3
635	A	Mood	18	32	33	40
		Physical	18	11	33	45
828	A	Mood	30	7	23	16
		Physical	7	6	24	16
1099	A	Mood	80	37	59	55
		Physical	27	31	57	25
1304	A	Mood	22	4	10	10
		Physical	23	6	16	13
1728	B	Mood	41	64	67	68
		Physical	44	57	46	26
2237	A	Mood	22	46	38	46
		Physical	18	15	40	25
2509	B	Mood	51	48	49	58
		Physical	52	31	29	34

The mood score at Time 3 for patient 635 and the physical score at Time 4 for patient 1099 shown in black are disregarded when considering a general missing data pattern in Example 2.4.

The patients showed different patterns of quality of life as measured by the coping score and there was large within- and between-patient variability. The patients' coping scores also displayed different missing data patterns. There were 4 patients (patients 47, 635, 1099, 2237) with a missing monotone data pattern (see section 1.6.1) and patient 828 had intermittent missing data. The remaining 5 patients had complete coping scores. All of the patients had complete mood and physical scores. The coping, mood and physical scores measure different aspects of quality of life and do not necessarily correspond.

2.2 Simple Imputation Methods

Simple imputation involves generating a single value from a predictive distribution for the imputation based on the observed data. In explicit modelling, the predictive distribution is based on a formal statistical model, for example multivariate Normal. In implicit modelling, imputation is based on an algorithm, which implies an underlying model, which is often not explicitly described. While the assumptions in implicit modelling are less obvious than in explicit modelling they are as important (Little and Rubin 2002, p59-60).

Simple imputation methods using explicit modelling include:

i) imputation of low or high values
ii) mean or median imputation
iii) simple imputation using linear regression models
iv) imputation of conditional means by Buck's method

while those based on implicit modelling include:

i) last observation carried forward (LOCF)
ii) hot-deck imputation
iii) nearest neighbour hot-deck imputation
iv) cold-deck imputation

Imputation of low or high values can provide a range of the variable of interest in the unknown complete data and is described first in section 2.2.1. Next, LOCF, which is the most prominent simple imputation method (Molenberghs and Kenward 2007, p.46), is described in section 2.2.2. This is followed by the remaining simple imputation methods using explicit modelling in section 2.2.3 to section 2.2.5. The remaining simple imputation methods using implicit modelling are described last in section 2.2.6 to section 2.2.8. Simple imputation is summarised in section 2.2.9. After each simple imputation method is described, it is illustrated using the example dataset described in section 2.1.1 and the imputed values are presented in Table 2.3.

Table 2.3 Example Coping Scores from a Breast Cancer Trial Following Simple Imputation

		Section							
		2.2.1	2.2.2	2.2.3	2.2.4	2.2.5	2.2.6	2.2.7	2.2.8
Patient	Time Period	High Value	LOCF	Median	LinReg	Cond Means	Hot Deck	NNHD	Cold Deck
47	4	90	49	69	72	72	17	56	35
635	3	90	46	48	33	35	21	17	25
635	4	90	46	48	48	48	16	56	25
828	2	90	9	9	24	24	46	2	10
1099	2	90	19	19	45	45	50	38	15
1099	3	90	19	19	50	50	32	40	15
1099	4	90	19	19	29	31	20	17	15
2237	4	90	32	38	29	29	20	16	25

High value = imputing high value; LOCF = last observation carried forward; median = median imputation by patient; linreg=simple imputation using linear regression models; cond means = imputing conditional means by Buck's method; NNHD = nearest neighbour hot-deck imputation

2.2.1 Imputing Low or High Values

This method uses an arbitrary high or low value for the missing quality of life assessments, such as imputing 0 or imputing a value just below the minimum observed value for all missing quality of life assessments. Imputation of low or high values is most commonly used when the missing observations are due to an adverse event such as death (e.g. Rabound et al. 1998), however it does not have a full theoretical justification (Fairclough 2010, p.174-175).

Example 2.1 Imputing High Values of Coping Score (Low Quality of Life)

There may be concern that patients with missing coping scores have a worse quality of life than patients with observed coping scores. Let the missing coping scores in the example dataset be replaced with a coping score of 90, a very poor quality of life slightly worse than the poorest coping score of 85 observed in the example. The imputed coping scores are shown in Table 2.3.

2.2.2 Last Observation Carried Forward

Last observation carried forward (LOCF) is also a common method of simple imputation. In this approach, a missing quality of life indicator is replaced by the

patient's last available value of the quality of life indicator. It is important to note that LOCF is not always a conservative analysis. In some settings, it can result in bias towards a treatment with more dropout associated with morbidity (Fairclough 2010, p.179; Molenberghs and Kenward 2007, Chapter 4).

The assumption is that if the missing quality of life assessment had been performed, the patient's quality of life would have been the same as the last available quality of life assessment. In many clinical trials, this assumption does not hold. For example, if a patient withdraws from a clinical trial because of adverse events, the patient's quality of life assessment after withdrawing is likely to be lower than before the patient withdrew. In some clinical trials, quality of life may decrease over time and imputation using last observation carried forward would lead to quality of life estimates which are too high (Fayers and Machin 2007, p.370-371; Fairclough 2010, p.172-173). The imputed coping scores in the example dataset are shown in Table 2.3.

2.2.3 Median or Mean Imputation

Imputing a median or mean value to replace missing observations is a common method of simple imputation. Median or mean imputation assumes that the missing values of the variable of interest follow the same distribution as the observed values, for example multivariate Normal. A further important assumption when the mean or median is calculated based on all observed values is that the missing observations are MCAR (Little and Rubin 2002, chapter 4). A disadvantage of median or mean imputation is that the missing values of the variable of interest Y are imputed by values at the centre of the distribution, leading to an underestimate of the variance of Y in the completed dataset (Little and Rubin 2002, p61).

It is possible to calculate the mean or median to be imputed from a group of patients with similar characteristics as the patient with the missing observation. An example is calculating the mean or median from patients suffering from an

adverse event. Calculating the median or mean from a group of patients with similar characteristics as the patient with the missing observation assumes that the impact on quality of life of the characteristics, such as suffering an adverse event, is the same among patients with an observed quality of life value as among patients with a missing quality of life value. (Fairclough 2002, p.167-168; Fayers and Machin 2007, p371-373).

Example 2.2 Median Imputation by Patient

There are several different methods of calculating the median or mean of a quality of life indicator that could be applied to the example quality of life assessments, including:

i) the median of the quality of life indictor across all patients at a particular time period
ii) the median of the quality of life indictor across all patients in a particular treatment group in a particular time period
iii) the median quality of life indictor for the patient across all time periods

As there is such a wide range of coping scores in the example dataset, imputing the median coping score is preferred to imputing the mean. For illustration, the median coping score by patient was used. It was calculated by *proc univariate* in SAS and rounded to the nearest whole number. The imputed coping scores are shown in Table 2.3.

2.2.4 Simple Imputation Using Linear Regression Models

Simple imputation using linear regression models involves identifying a regression model to predict the missing observation. This extends mean imputation by imputing conditional means given observed values (Little and Rubin 2002, chapter 4). In clinical trials involving quality of life, this approach has the advantage that the linear regression model used to impute the missing observations can include concurrent information such as adverse events or stage of disease that are not used in the analytic model of interest. The linear regression

model could also include quality of life values assessed by another person, such as a relative or nurse caring for the patient (Fairclough 2002, p.118-119).

The analytic model for the i^{th} patient (i=1,…,n) is written as:

$$Y_i = \boldsymbol{X}_i\boldsymbol{\beta} + \varepsilon_i \qquad (2.1)$$

where Y is the response variable (outcome measure)

$\boldsymbol{X}_i$ is the design matrix of covariates used in the analytic model for patient i

$\boldsymbol{\beta}$ is the vector of parameters used in the analytic model

ε_i is the residual error for patient i

X_{0i} is set to 1 in order that β_0 is the intercept

Similarly, the imputation model is for the ith patient is written as:

$$Y_i^* = \boldsymbol{X}_i^*\boldsymbol{B}^* + \varepsilon_i^* \qquad (2.2)$$

where $\boldsymbol{X}_i^*$ is the design matrix of covariates used in the imputation model for patient i

$\boldsymbol{B}^*$ is the vector of parameters used in the imputation model

ε_i^* is the residual error for patient i

Here * is used to distinguish the covariates and corresponding parameters of the imputation model from those included in the analytic model. The vector of parameters used in the imputation model is written as $\boldsymbol{B}^*$ as though it may be augmented from the vector of parameters used in the analytic model, it will have different terms and estimates compared to the analytical model.

Using a linear regression model to predict the missing observations will only be appropriate if the assumptions for the general linear model (e.g. Lang and Secic 2006) are met. The aim when carrying out imputation using linear regression models is to identify a linear regression model where the missing data mechanism depends only on the observed data and the covariates X_i^* in the imputation model. The covariates included in the linear regression model are likely to be strongly correlated with the variable being imputed and the probability that the observation

of the variable is missing. The assumption that the data are MAR should then be reasonable in the imputation model (Fairclough 2010, p.169). The covariates included in the regression model must have no missing observations, giving a univariate missing data pattern (see section 1.6.1).

In the longitudinal setting, the parameters of the imputation model (2.2) for the *i*th patient at the *h*th measurement (*h*=1,...,*H*) are estimated from the observed data with the model:

$$Y_{hi}^{obs} = \boldsymbol{X}_{hi}^{*\,obs}\boldsymbol{B}_{h}^{*} + \varepsilon_{hi}^{*} \tag{2.3}$$

The predicted values which will be imputed to replace the missing values are then calculated as follows:

$$Y_{hi}^{*miss} = \boldsymbol{X}_{hi}^{*\,miss}\hat{\boldsymbol{B}}_{h}^{*} \tag{2.4}$$

Example 2.3 Imputation Using Linear Regression with Concurrent Quality of Life Assessments

Let a linear regression model with the mood score, $X_{h\text{mood}}$, and physical score, $X_{h\text{phys}}$, (Table 2.2) be used to impute missing observations of the coping score Y_h^{miss} at the *h*th assessment (*h*=1 for Time 1, *h*=2 for Time 2, *h*=3 for Time 3 and *h*=4 for Time 4) in the example dataset (Table 2.1).

The linear regression models were calculated by *proc reg* in SAS from the patients with observed coping scores. The linear regression models were:

$$Y_{4i} = 7.813 - 0.041X_{4i\text{mood}} + 0.936X_{4i\text{phys}} \tag{2.5}$$

$$Y_{3i} = 10.457 + 0.558X_{3i\text{mood}} + 0.116X_{3i\text{phys}} \tag{2.6}$$

$$Y_{2i} = 19.469 + 0.441X_{2i\text{mood}} + 0.287X_{2i\text{phys}} \tag{2.7}$$

The imputed coping scores are shown in Table 2.3.

2.2.5 Imputing Conditional Means Using Buck's Method

Buck (1960) extended simple imputation using linear regression models to impute conditional means when the missing observations have a general missing data pattern rather than a univariate missing data pattern. Suppose the variables Y_1, …,Y_J follow a multivariate Normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Suppose c of the n units (patients in the case of clinical trials) have a complete set of J observations. Then if the ith patient has J_i^{miss} of the J observations missing, these missing values can be estimated by linear regression on the $J–J_i^{miss}$ observed variables.

In the method proposed, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are first estimated from the sample mean and covariance matrix based on the c patients with complete observations. The estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are then used to generate the linear regressions of the variables to be imputed on the observed variables. For each patient, the $J– J_i^{miss}$ observed variables are substituted into the linear regression model to give the estimate of the missing values of the J_i^{miss} variables to be imputed. In the special case where a patient has a missing value in a single variable ($J_i^{miss} = 1$), the method is equivalent to simple imputation using linear regression models (section 2.2.4).

Conditional mean imputation using Buck's method assumes that the data is MCAR and the variables Y_1, …,Y_J are not highly correlated and there is not a large proportion of missing data (Buck 1960; Little and Rubin 2002, p.63). A proportion of missing data of 10% was considered large by Gleason and Staelin (1975). Little and Rubin (2002, p.63) note that Buck's method is a valid way to estimate $\boldsymbol{\mu}$ under certain types of MAR missing data mechanisms.

Example 2.4 Imputation of Conditional Means using Buck's Method

Here, the coping score, mood score and physical score ($J = 3$ in the notation of this section) at a particular time point h are considered. As noted, imputing conditional means using Buck's method applies to a dataset with a general

missing data pattern. In order to illustrate the method, the mood score at Time 3 for patient 635 and the physical score at Time 4 for patient 1099 shown in black in Table 2.2 are disregarded.

Suppose the coping score $Y_{h\text{sp}}$, mood score $Y_{h\text{mood}}$ and physical score $Y_{h\text{phys}}$ at the hth assessment follow a multivariate Normal distribution

$$\boldsymbol{Y}_h \sim \text{N}(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h) \tag{2.8}$$

where h=1 for Time 1, h=2 for Time 2, h=3 for Time 3 and h=4 for Time 4

With 2 exceptions, the missing coping score in the example dataset (Table 2.1) is the single variable with a missing value. Therefore, the missing coping scores in the example dataset are imputed as in Example 2.3 with 2 exceptions. The exceptions are the missing coping score at Time 4 for patient 1099 and the missing coping score at Time 3 for patient 635.

When only the mood score at Time 4 was observed, a linear regression model with mood score at Time 4 is used to impute the missing coping score at Time 4 The linear regression model was calculated by *proc reg* in SAS from patients with observed coping, mood and physical scores at Time 4. The linear regression model for the missing coping score at Time 4 based on mood score is:

$$Y_{4i} = 12.376 + 0.332Y_{4i\text{mood}} \tag{2.9}$$

Similarly, the linear regression model for missing coping score at Time 3 based on physical score is:

$$Y_{3i} = 11.614 + 0.700Y_{3i\text{phys}} \tag{2.10}$$

The imputed coping scores are shown in Table 2.3.

2.2.6 Hot-deck Imputation

The concept of hot-deck procedure is to impute a value from similar patients in the trial to replace missing observations, although the precise definition of the hot-

deck procedure is not well defined (Little and Rubin 2002, chapter 4). The observed values of the variable to be imputed, *Y*, from the patients with similar characteristics to the patient with the missing observation form the hot-deck. The advantage of the hot-deck method compared to mean or median imputation is that the imputed values are not always from the centre of the distribution of *Y* and so do not distort the distribution of the variable *Y* in the completed dataset (Little and Rubin 2002, p.68). However, Rubin and Schenker (1986) note that by drawing the imputed values from a set of observed values, the hot-deck method acts as though the distribution of observed values of *Y* is the same as the population. In fact, the distribution of *Y* in the population is not precisely known and this uncertainty is not reflected. As with other simple imputation methods, the variability in *Y* is underestimated.

In the simplest version of hot-deck imputation, the hot-deck of potential values is formed from all patients with an observed value and the imputed values are drawn with an equal probability. This is referred to as hot-deck imputation by simple random sampling with replacement and assumes the data are MCAR. Generally, the hot-deck is restricted to a subset of patients with observed values defined as similar to the patient with the missing observation. Assume that there are n patients classified as belonging to the same group of similar patients as the patient with the missing value and r out of the n have an observed value of Y, where $n > r$. Using random sampling with replacement amongst the group of similar patients, the hot-deck here consists of the r patients with an observed value.

Many different methods for selecting these patients with similar characteristics to the patient with the missing observation have been used and may be based on the previous experience of the analyst (Little and Rubin 2002, p.60; Marker et al. 2002, p.329). The aim is to use covariates associated with the value of interest and the probability of the variable of interest being missing to define the hot-deck (Marker et al. 2002, p.329). For example, among the covariates used to define the hot-deck in Bordeleau et al. (2003) were i) baseline performance status and ii)

previous global quality of life scores when imputing missing global quality of life scores at i) baseline ii) follow-up assessments respectively. The covariates used to define the hot-deck must have no missing observations, giving a univariate missing data pattern. Selecting patients with similar characteristics as the patient with the missing observation assumes that the impact on quality of life of the characteristics, such as baseline performance status, is the same among patients with an observed quality of life value and patients with a missing quality of life value (Meng 2002, p.344).

Example 2.5 Hot-deck Imputation

In the example dataset (Table 2.1), the baseline coping score could potentially be used to restrict the hot-deck. Let the hot-deck of potential coping scores to replace the missing coping score be defined as: the observed coping scores at the same time period as the missing coping score among patients with baseline coping score in the same group as the patient with the missing coping score. The baseline coping score groups are patients with baseline coping score ≤ 50 and patients with baseline coping score > 50.

Proc surveyselect in SAS with the *method* option set to urs (for unrestricted random sampling) performs simple random sampling with replacement. For illustration, suppose that the coping score selected at random from the hot-deck of potential coping scores to replace the missing coping scores was as shown in Table 2.3.

2.2.7 Nearest Neighbour Hot-deck Imputation

Nearest neighbour hot-deck imputation uses information from covariates to attempt to address the issue of missing data and makes the assumption that the data are MAR. A definition of the distance between patients is specified based on the values of the covariates. A missing observation is replaced by an imputed value chosen from the observed values of patients defined as close to the patient with a missing observation (Little and Rubin 2002, p.68-69). A possible metric to

describe the distance between patient *a* with a missing observation and patient *b* with an observed value is the predictive mean:

$$d(a,b)=[\hat{y}(x_a)-\hat{y}(x_b)]^2 \qquad (2.11)$$

where $\hat{y}(x_a)$ is the predicted value of the missing Y_a^{miss} and $\hat{y}(x_b)$ is the predicted value of the observed Y_b^{obs} from the regression of *Y* on the covariates $\boldsymbol{X}$ computed from the complete patients. (Little and Rubin 2002, p.69). A linear regression model to predict the value of *Y* requires the assumptions described in section 2.2.4. The hot-deck is formed from the observed values of *Y* from complete patients such that *d(a,b)* is less than some value d_0. The value of d_0 selected controls the number of potential values in hot-deck (Little and Rubin 2002, p.69). Increasing the value of d_0 increases the number of potential values in the hot-deck. However, it will reduce the probability the potential values are similar to the missing value by allowing a larger difference between $\hat{y}(x_a)$ and $\hat{y}(x_b)$.

Example 2.6 Nearest Neighbour Hot-Deck Imputation

Let a linear regression model with the mood score, $X_{h\text{mood}}$, and physical score, $X_{h\text{phys}}$, be used to impute missing observations of the coping score Y_h^{miss} at the h^{th} assessment as in Example 2.3.

In the example dataset, the values of *d(a,b)* have a wide range, which is widest at Time 4. They are generally larger at Time 4 than at Times 2 and 3. The values of d_0 selected for each time period reflect this. The hot-deck of potential coping scores to replace the missing coping scores at Time 4 was defined by setting d_0 to 1300. The hot-deck of potential coping scores to replace the missing coping scores at Time 3 and Time 2 was defined by setting d_0 to 100. For illustration, suppose the coping scores selected at random to replace the missing coping scores were as shown in Table 2.3.

2.2.8 Cold-deck Imputation

The cold-deck procedure involves imputing a constant value from an external source, in place of missing observations. An example of an external source for a missing quality of life observation is a value from a previous study of similar patients (Fayers and Machin 2000, p.244). Fayers and Machin (2000, p.244) note that cold-deck imputation is unlikely to be useful in the context of clinical trials. The statistical analysis of the completed dataset usually proceeds as for complete data (Little and Rubin 2002, p.60-61). Little and Rubin (2002, p.61) noted that theoretical justification for such statistical analysis of completed datasets following cold-deck imputation may be lacking. Lessler and Kalsbeek (1992, p.214) describes cold-deck imputation as rarely used and of historical interest.

Example 2.7 Cold-deck Imputation

Suppose that information from a similar trial to the breast cancer trial in the example was used as the cold deck for the median coping score between 3 months (Time 2) and 9 months (Time 4). Next, suppose that the missing coping scores are replaced with the median coping score after baseline according to baseline coping score suggested by the similar trial. For example, that the median coping score after baseline among patients with baseline coping score 0 to 6 are replaced by 5. Given these suppositions, the imputed coping scores are as in Table 2.3.

2.2.9 Summary of Simple Imputation

The standard simple imputation methods were illustrated by a small example dataset. The imputed values in Table 2.3 indicate that imputing a high value was least like the other simple imputation methods. As previously noted, imputing conditional means by Buck's method (Example 2.4) extended simple imputation using linear regression to a general missing data pattern (Example 2.3). The difference between Example 2.3 and Example 2.4 is that in Example 2.4 the mood score at Time 3 for patient 635 and the physical score at Time 4 for patient 1099 are disregarded. It was thus expected that the imputed values following these imputation methods were very alike. There was no suggestion that there is a

completed dataset which will be achieved regardless of which simple imputation method is applied (Table 2.3).

There are only limited circumstances when it is appropriate to draw inferences from the parameter estimate resulting from simple imputation. Justification should be provided if the parameter estimates are considered (Molenbergs and Kenward 2007, Chapter 4). Simple imputation methods can give useful information about the sensitivity of the results from clinical trials to assumptions about missing data. However, an important disadvantage is that they lead to the underestimation of the variance of the observations. Imputing a predetermined value for the missing observations assumes that there is no variation in the missing values. The true predetermined value for the missing observations are assumed to be known when they are only estimated (Little and Rubin, 2002, chapter 4). The variance of the observations is also underestimated by the hot-deck method. This is due to the fact the hot-deck method acts as though the distribution of observed values of Y is the same as the population when drawing the imputed value from the set of potential values. Unlike simple imputation, multiple imputation, described in the next section, allows the uncertainty about the values to imputed, and the uncertainty of the model the imputed values are drawn from, to be considered (Little and Rubin, p.85-86).

2.3 Multiple Imputation

Multiple imputation has become an important technique for dealing with missing observations. It is most directly motivated from the Bayesian perspective (Little and Rubin 2002, p.87). Kenward and Carpenter (2007) note that a key feature is that it is based on two distinct models: the analytic model, the target of the analysis, and the imputation model, which can be thought of as the conditional predictive distribution. The analytic model, also known as the substantive model, is the model for analysis that would have been appropriate had the data been complete. The parameters of interest, such as the mean or hazard ratio, are derived from it (Molenbergs and Kenward 2007 p. 106). The imputed values are draws from the imputation model. Little and Rubin (2002, p.86) recommend that imputed values be drawn according to the following protocol. For each model for non-response being considered, the K imputations of Y^{miss} are K repetitions from the posterior predictive distribution of Y^{miss}, each repetition corresponding to an independent drawing of the parameters and missing observations. In practice, an implicit model rather than an explicit model (see section 2.2) can often be used as the imputation model (Little and Rubin 2002, p.86).

The purpose of applying multiple imputation is usually to obtain valid inferences from standard statistical analysis. Multiple imputation addresses the important disadvantages of simple imputation that simple imputation methods do not reflect the uncertainty of the values to be imputed and the uncertainty of the model the imputed values are drawn from. It leads to K completed datasets which are analysed by standard procedures. The results from the analysis of the K completed datasets can be combined to give valid inferences reflecting the uncertainty of the missing values drawn from the imputation model. Drawing imputed values under more than one imputation model allows the uncertainty about the correct model to be shown by the variation in valid inferences across the models (Little and Rubin 2002, p.85-87).

The procedure for combining the results from the analysis of the K completed datasets is described by Rubin (1987, p.75-76). Let $\boldsymbol{\theta}$ be the vector of parameters from the analytic model. Let $\widehat{\boldsymbol{\theta}}_k$ and $\boldsymbol{V}_k$, k=1,..,K be the estimate of $\boldsymbol{\theta}$ from the k^{th} completed dataset and the corresponding conventional estimate of the covariance matrix of $\widehat{\boldsymbol{\theta}}_k$, calculated as though the completed dataset was fully observed. The combined estimate of $\boldsymbol{\theta}$, $\bar{\boldsymbol{\theta}}_K$, is the average of the estimates $\widehat{\boldsymbol{\theta}}_k$ from each complete dataset. The variability of $\boldsymbol{\theta}$ associated with estimates from multiple imputation, V_{MI}, has two components: the average within-imputation variance component (W) and the between-imputation component (B), giving the total variance:

$$\boldsymbol{V}_{MI} = \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{V}_k + \left(1+\frac{1}{K}\right)\left(\frac{1}{K-1}\right)\sum_{k=1}^{K}\left(\widehat{\boldsymbol{\theta}}_k - \bar{\boldsymbol{\theta}}_K\right)^2 \tag{2.12}$$

or

$$\mathbf{V}_{MI} = \boldsymbol{W} + \left(1+\frac{1}{K}\right)\boldsymbol{B} \tag{2.13}$$

We assume that with complete data, inferences for $\boldsymbol{\theta}$ would be based on the Normal approximation

$$\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right) \sim N(\mathbf{0}, \boldsymbol{V}) \tag{2.14}$$

where $\widehat{\boldsymbol{\theta}}$ is a statistic estimating $\boldsymbol{\theta}$, $\boldsymbol{V}$ is a statistic providing the covariance matrix and $N(\mathbf{0}, \boldsymbol{V})$ is a multivariate Normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{V}$.

Now consider a scalar θ and large sample size. The reference distribution for tests and confidence intervals is a t distribution

$$(\theta - \bar{\theta}_K)V_{MI}^{-1/2} \sim t_v \tag{2.15}$$

where

$$v = (k-1)(1+\tfrac{1}{r})^2 \tag{2.16}$$

and the relative increase in variance r is

$$r = \frac{(1+k^{-1})B}{W} \tag{2.17}$$

The validity of such inferences depends on the imputation technique that led to the imputed values. It is important that the imputed values give reasonable predictions for the unknown missing observation and that the variability of the imputed values reflects a degree of uncertainty (Schafer 1999). Rubin (1987, chapter 3) discusses how to obtain valid inference from multiple imputation. Consider a theoretically fundamental form of multiple imputation, repeated imputation (Rubin 1987, p.75-76). Repeated imputations are draws from the posterior predictive distribution of the missing observations under a particular Bayesian model for both the data and the missing data mechanism. The repeated inference method involves analysing each of the simulated datasets by standard methods and then combining the results.

The number of repetitions of multiple imputation required for good statistical inference has been discussed in the statistical literature (e.g. Rubin 1987, chapter 4; Schafer and Olsen 1998; Graham et al. 2007). The fraction of missing information γ and the efficiency of the estimator are considered. Schafer and Oslen (1998) give the formula for γ as:

$$\gamma = \frac{r + (2/(v+3))}{r+1} \tag{2.18}$$

where r is the relative increase in variance and v is the number of degrees freedom of the t distribution. Rubin (1987, p.114) shows that the efficiency of the estimate based on K imputations is:

$$\left(1 + \frac{\gamma}{K}\right)^{-1} \tag{2.19}$$

and notes that gains in the efficiency of the estimate rapidly diminish after the first few imputations (p. 548-549). In many applications, K=3 to 5 imputations are sufficient (Rubin 1987, chapter 4; Molenbergs and Kenward 2007 p. 109). However, Graham et al. (2007) recommend using a larger number of imputations based on considering statistical power to detect a small effect sizes in the parameter estimate compared to maximum likelihood methods as well as γ.

An important issue in multiple imputation is the construction and application of the imputation model. As with simple imputation, if the imputation model does not capture the missing data mechanism, then any analysis based on the imputation is flawed. Therefore, it is important to investigate the missing data mechanism thoroughly and to give careful consideration to the choice of variables in the imputation model. Unlike simple imputation, the uncertainty about the correct imputation model can be explored by performing multiple imputations from more than one imputation model.

Both simple and multiple imputation techniques allow the imputation model to be different from the analytic model. As noted, in the clinical trial setting, it is possible that using a different imputation model from the analytic model of interest in the data being considered may make it appropriate to apply imputation methods assuming the data are MAR. Most of the techniques for multiple imputation assume that the missing observations are MAR. This assumption means that an explicit probability model for the missing observations is not required. A further assumption made by most multiple imputation methods in the longitudinal setting is that there is a montontone missing data pattern.

Methods based on an explicit parametric Bayesian model for the imputation model include:

i) explicit univariate regression
ii) Markov chain Monte Carlo methods – data augmentation
iii) Markov chain Monte Carlo methods – Gibbs' sampling
iv) pattern mixture models – Curran's analytic technique

Methods based on an implicit modelling include:

i) approximate Bayesian bootstrap
ii) nearest neighbour imputation
iii) predictive mean matching

Among these standard multiple imputation methods (based on explicit or implicit modelling), the following assume a monotone missing data pattern:

i) approximate Bayesian bootstrap
ii) explicit univariate regression
iii) nearest neighbour imputation
iv) predictive mean matching
v) pattern mixture models – Curran's analytic technique

and the following extend to data with a general missing data pattern:

i) Markov chain Monte Carlo methods – data augmentation
ii) Markov chain Monte Carlo methods – Gibbs' sampling

The Markov chain Monte Carlo (MCMC) methods that extend to data with a general missing data pattern are described in section 2.3.1 and section 2.3.2 and the multiple imputation methods assuming a monotone missing data pattern are described in section 2.3.3 to section 2.3.6. Future directions are described in section 2.3.7 and multiple imputation is summarised in section 2.3.8.

Considering the example data from a breast cancer trial (Table 2.1 and Table 2.2), the imputed coping scores following 5 repetitions of the multiple imputation methods described in this section are shown in Table 2.4. As previously noted, patient 828 had intermittent missing data. For methods requiring a monotone missing data pattern, the imputed coping scores for patient 828 at Time 2 are indicated in black. These coping scores were imputed by the MCMC method of data augmentation in order to create a monotone missing data pattern to illustrate the method.

Table 2.4 Example Coping Scores from a Breast Cancer Trial Following Multiple Imputation

		Section	2.3.1	2.3.1	2.3.2	2.3.3	2.3.4	2.3.5	2.3.5	2.3.6
		Example	2.8	2.9	2.1	2.11	2.12	2.13	2.14	2.15
k	Patient	Time Period	Data Aug (Mult)	Data Aug (Rep)	Gibbs'	ABB	Exp Uni Reg	NNI	Pred Mean Match	P Mix Models
1	828	2	12	35	12	12	12	12	12	12
2	828	2	30	19	30	30	30	30	30	30
3	828	2	9	54	9	9	9	9	9	9
4	828	2	9	0	9	9	9	9	9	9
5	828	2	35	28	35	35	35	35	35	35
1	1099	2	37	52	37	32	57	32	38	38
2	1099	2	45	51	45	32	13	32	46	61
3	1099	2	54	22	54	2	18	32	38	32
4	1099	2	79	9	79	46	63	32	46	73
5	1099	2	89	31	89	50	27	32	38	54
1	635	3	61	23	61	18	15	60	40	56
2	635	3	46	61	46	21	54	10	60	50
3	635	3	15	45	15	18	42	18	18	36
4	635	3	31	24	31	17	30	60	32	48
5	635	3	38	32	38	21	38	10	10	69
1	1099	3	55	2	55	18	12	17	17	34
2	1099	3	53	56	53	32	21	18	10	11
3	1099	3	33	12	33	21	16	10	10	50
4	1099	3	73	29	73	21	9	18	60	33
5	1099	3	69	28	69	21	16	18	32	46

k indicates the repetition of multiple imputation; Data aug (Mult) = Markov chain Monte Carlo methods – data augmentation with more than one type of assessment; Data aug (Rep) = Markov chain Monte Carlo methods – data augmentation with repeated measures; Gibbs' = Markov chain Monte Carlo methods – Gibbs' sampling; ABB = approximate Bayesian bootstrap; Exp uni reg = explicit univariate regression; NNI = nearest neighbour imputation; P mix models = pattern mixture models - Curran's analytic technique

For methods requiring a monotone missing data pattern, the imputed coping scores imputed by the MCMC method of data augmentation in order to create a monotone missing data pattern to illustrate the method are shown in black

Table 2.4 Example Coping Scores from a Breast Cancer Trial Following Multiple Imputation (continued)

		Section	2.3.1	2.3.1	2.3.2	2.3.3	2.3.4	2.3.5	2.3.5	2.3.6
		Example	2.8	2.9	2.10	2.11	2.12	2.13	2.14	2.15
k	Patient	Time Period	Data Aug (Mult)	Data Aug (Rep)	Gibbs'	ABB	Exp Uni Reg	NNI	Pred Mean Match	P Mix Models
1	47	4	76	55	76	56	96	56	56	52
2	47	4	98	54	98	56	34	2	16	52
3	47	4	67	28	67	56	48	56	18	61
4	47	4	81	40	81	56	28	56	20	12
5	47	4	83	51	83	56	23	56	17	57
1	635	4	45	53	45	20	0	56	18	40
2	635	4	59	34	59	16	81	2	56	81
3	635	4	43	45	43	20	5	20	2	39
4	635	4	46	18	46	20	11	56	20	41
5	635	4	62	54	62	2	32	18	16	39
1	1099	4	22	9	22	20	39	20	2	16
2	1099	4	32	34	32	16	39	2	18	58
3	1099	4	21	13	21	20	60	17	16	26
4	1099	4	29	23	29	20	21	18	20	4
5	1099	4	3	15	3	2	20	20	17	56
1	2237	4	23	53	23	16	11	20	16	14
2	2237	4	21	31	21	16	9	17	20	48
3	2237	4	28	21	28	16	64	17	17	46
4	2237	4	22	25	22	2	63	17	20	36
5	2237	4	30	22	30	2	21	20	16	5

k indicates the repetition of multiple imputation; Data aug (Mult) = Markov chain Monte Carlo methods – data augmentation with more than one type of assessment; Data aug (Rep) = Markov chain Monte Carlo methods – data augmentation with repeated measures; Gibbs' = Markov chain Monte Carlo methods – Gibbs' sampling; ABB = approximate Bayesian bootstrap; Exp uni reg = explicit univariate regression; NNI = nearest neighbour imputation; P mix models = pattern mixture models - Curran's analytic technique

2.3.1 Markov chain Monte Carlo Method of Data Augmentation Bayesian Inference

Generally, the application of Markov chain Monte Carlo (MCMC) is for the purpose of Bayesian inference (Gilks et al. 1996, p.2). Let $\boldsymbol{\theta}$ denote the unknown model parameters and the missing data. Information about $\boldsymbol{\theta}$ is expressed as a posterior probability distribution

$$P(\boldsymbol{\theta} \mid Y^{obs}) = \frac{P(\boldsymbol{\theta})P(Y^{obs} \mid \boldsymbol{\theta})}{\int P(\boldsymbol{\theta})P(Y^{obs} \mid \boldsymbol{\theta})\, d\boldsymbol{\theta}} \tag{2.20}$$

Any features, such as the mean or quartiles, of the posterior distribution may be considered in Bayesian inference. These quantities can be expressed in terms of the posterior expectations of functions of $\boldsymbol{\theta}$. The posterior expectation of a function:

$$E[f(\boldsymbol{\theta}) \mid Y^{obs}] = \frac{\int f(\boldsymbol{\theta})P(\boldsymbol{\theta})P(Y^{obs} \mid \boldsymbol{\theta})\ d\boldsymbol{\theta})}{\int P(\boldsymbol{\theta})P(Y^{obs} \mid \boldsymbol{\theta})\, d\boldsymbol{\theta}} \tag{2.21}$$

To consider this in general terms, let $\boldsymbol{X}$ be a vector of J random variables, with distribution $\pi(.)$. In Bayesian applications, $\boldsymbol{X}$ comprises of model parameters and missing data and $\pi(.)$ is a posterior distribution. The expectation required to be evaluated is:

$$E[f(\boldsymbol{X})] = \frac{\int f(x)\pi(x)\ dx}{\int \pi(x)\, dx} \tag{2.22}$$

for some function of interest $f(.)$ (Gilks et al. 1996, p.3-4).

In most applications, $E[f(X)]$ cannot be evaluated analytically. MCMC is a collection of methods for evaluating $E[f(X)]$ and has two components: Monte Carlo integration and Markov chains (Gilks et al. 1996, p.3-4).

Monte Carlo integration

Monte Carlo integration evaluates $E[f(X)]$ by drawing samples $\{X_t, t=1,..l\}$ from $\pi(.)$ and then approximating (Gilks et al. 1996, p.4):

$$E[f(X)] \approx \frac{1}{l}\sum_{t=1}^{l} f(X_t) \tag{2.23}$$

When the samples $\{X_t, t=1,..l\}$ are independent, the approximation can be made as accurate as required by increasing the sample size l. However, in general, drawing samples $\{X_t, t=1,..l\}$ independently from $\pi(.)$ is not feasible. Markov chains can be used to generate the samples $\{X_t\}$ (Gilks et al. 1996, p.4).

Markov Chains

Consider a sequence of random variables, $\{X_0, X_1, X_2,\ldots\}$ such that at each time $t \geq 0$ the next step X_{t+1} is sampled from a distribution $P(X_{t+1} \mid X_t)$ which depends only on the current state of the chain X_t. The sequence is a Markov chain and $P(.|.)$ is the transition kernel of the chain. Assuming that $P(.|.)$ does not depend on t means that the Markov chain is time-homogeneous (Gilks et al. 1996, p.5).

In MCMC methods, Markov chains are used to simulate random observations from non-standard distributions (Schafer 1999). A Markov chain is constructed long enough for the distribution of the elements to stabilise to a unique stationary distribution, which does not depend on t or X_0. (Gilks et al. 1996, p.5). The stationary distribution is denoted as $\varphi(.)$. This implies that the sampled points $\{X_t\}$ will look increasingly like dependent samples from $\varphi(.)$ (Gilks et al. 1996, p.5). There are many ways of constructing these chains (Gilks et al. 1996, chapter 1).

After a sufficiently long "burn-in" of d iterations, draws from the distribution of interest $\varphi(.)$ are simulated by repeatedly simulating the steps of the Markov chain. The output from the Markov chain is used to estimate the expectation $E[f(X)]$ where X has distribution $\varphi(.)$ (Gilks et al. 1996, chapter 1).

Implementing MCMC Methods

Implementing MCMC methods begins with preliminary analysis, such as plotting the raw data, to explore how to model the data. When implementing MCMC methods, several issues must be considered. These include: i) running single or multiple chains, ii) burn-in iii) run length, iv) starting values and v) examining summary statistics for evidence of lack of fit of the model (Gilks et al. 1996, chapter 1). These issues are summarised briefly below.

Number of chains

Gilks et al. (1996, p.13) note that recommendations in the statistical literature on the number of chains have been conflicting. The argument made for a single chain is that one very long run is likely to be more precise for estimating a single quantity such as a posterior mean and comparison between chains cannot prove convergence. Whereas, the argument made for multiple chains is that comparing several seemingly converged chains might reveal genuine differences if the chains have not yet approached stationarity (Gilks et al. 1996, p.13).

Starting values

In order for the distribution of X_t to converge to a stationary distribution, the chain must be irreducible. This means that, from all starting points, the Markov chain can reach any non-empty set with positive probability, in some number of iterations. Thus, the starting values will not affect the stationary distribution. (Gilks et al 1996, p.46; Gilks et al. p.13). Gilks et al (1996, p.13) note that generally starting values do not need to be chosen carefully.

Burn-in

The factors influencing the length of burn-in d are X_0, the rate of convergence of $P^{(t)}(X_t \mid X_0)$ to $\pi(X_t)$ and on how similar $P^{(t)}(. \mid .)$ and $\pi(.)$ are required to be (Gilks et al. p14). A common method for determining burn-in is the visual inspection of plots of the Monte Carlo output $\{X_t, t=1,..l\}$. Formal convergence diagnostics for determining d have been proposed (e.g. Raftery and Banfield 1991; Raftery and

Lewis 1992; Rubin 1981; Rubin 1984), though these also make use of Monte Carlo output in some way (Gilks et al 1996, p.14).

Run length

The aim is to run the chain long enough to obtain adequate precision in calculating the expectation $E[f(X)]$. Generally, an estimator $\bar{f}$ ignoring the burn-in samples is used to calculate this expectation. A common method for determining the run length l is to run several chains in parallel, with different starting values, and informally compare the estimates $\bar{f}$. If the estimates $\bar{f}$ are not sufficiently consistent, l is increased (Gilks et al. 1996, p.15). More formal methods, which aim to estimate the variance of $\bar{f}$ have been proposed (e.g. convergence diagnostics referenced above).

Assessing goodness-of-fit

When implementing MCMC, it is important to consider: i) is the given model adequate and ii) which of the potential models under consideration is the best? (Gilks et al. 1996, p.144). MCMC methods allow multi-level models with large number of parameters for which standard asymptotic likelihood theory does not apply and therefore, particular care is needed when assessing the goodness-of-fit (Gilks et al. 1996, p.34). Classical approaches to assessing model adequacy generally involve defining a measure of fit, often a deviance statistic, and complexity. Complexity is described by the number of free parameters in the model. As increasing the complexity of a model increases the fit, models are compared by trading off these two quantities (Spiegelhalter et al. 2002). Proposals are often formally based on minimising a measure of expected loss on a future replicate dataset, for example Efron (1986).

Multiple Imputation from Parametric Bayesian Models and Data Augumentation

MCMC is often used to perform multiple imputation in non-straightforward situations (Schafer, 1999). Suppose Y follows a parametric model $P(Y|\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ has a prior distribution and Y^{miss} is missing at random. Since

$$P(Y^{miss}|Y^{obs}) = \int P(Y^{miss}|Y^{obs}, \boldsymbol{\theta})\, P(\boldsymbol{\theta}|Y^{obs})\, d\boldsymbol{\theta} \quad (2.24)$$

an imputation for Y^{miss} is generated by first generating a random value of the unknown parameters from their observed-data posterior

$$\boldsymbol{\theta}^* \sim P(\boldsymbol{\theta} \mid Y^{obs}) \quad (2.25)$$

followed by generating a random value of the missing observations from their conditional predictive distribution

$$Y^{*miss} \sim P(Y^{miss}|Y^{obs}, \boldsymbol{\theta}^*) \quad (2.26)$$

For many common models the conditional predictive distribution is simple but the observed-data posterior distribution is not. Generally, the observed-data posterior distribution (2.25) is not from a standard distribution and cannot be easily simulated. In this situation, MCMC methods can be used to generate simulations from it (Schafer, 1999).

The MCMC method of Gibbs' sampling became widely used following the work of Geman and Geman (1984). Gibbs' sampling can be used to impute missing data and is described in section 2.3.2. Tanner and Wong (1987) proposed the MCMC method of data augmentation for imputing missing data and this is described in this section. Data augmentation is an iterative method of simulating the posterior distribution that can be applied to Bayesian inference with missing data.

Suppose a random vector X is divided into two subvectors, $X = (X_a, X_b)$, and let $P(X)$ be the joint distribution of X, which is the target distribution for simulation. In data augmentation, we assume that the joint distribution $P(X)$ is not easily

simulated but the conditional distributions P(X_a | X_b) and P(X_b | X_a) are (Schafer 1997, p.70).

At iteration *t*, let

$$X^{(t)} = (x_1^t, \ldots, x_m^t) = ((x_{a1}^t, x_{b1}^t), \ldots, (x_{am}^t, x_{bm}^t)) \qquad (2.27)$$

be a sample of size *m* that approximates the target distribution P(X). This sample is updated in two steps, first drawing

$$X_a^{t+1} = (x_{a1}^{t+1}, \ldots, x_{am}^{t+1}) \qquad (2.28)$$

and then drawing

$$X_b^{t+1} = (x_{b1}^{t+1}, \ldots, x_{bm}^{t+1}) \qquad (2.29)$$

(Schafer 1997, p.71). Tanner and Wang (1987) show that the distribution of $X^{(t)}$ converges to P(X) as $t \to \infty$.

One run of data augmentation iterates to a draw from the posterior predictive distribution of Y^{miss} and a draw from the posterior distribution of $\boldsymbol{\theta}$. The two steps in the iterative sampling scheme are described by Schafer (1997, p.72) and Little and Rubin (2002, p.201) as follows:

i) Imputation step (I-step)

Given the current value of $\boldsymbol{\theta}^{(t)}$ of $\boldsymbol{\theta}$ drawn at iteration *t*, draw a value for the missing data from the conditional predictive distribution of Y^{miss}

$$Y_{miss}^{(t+1)} \sim P(Y^{miss} \mid Y^{obs}, \boldsymbol{\theta}^{(t)}) \qquad (2.30)$$

ii) Posterior Step (P-step)

Conditioning on $Y_{miss}^{(t+1)}$, draw a new value of $\boldsymbol{\theta}$ from its complete-data posterior

$$\boldsymbol{\theta}^{(t+1)} \sim P(\boldsymbol{\theta} \mid Y^{obs}, Y_{miss}^{(t+1)}) \qquad (2.31)$$

After a suitably large number of iterations, $\boldsymbol{\theta}^{(t)}$ can be regarded as an approximate draw from the observed-data posterior (2.25) (Schafer 1997, p.72). Thus, one run of data augmentation approximates a draw of an imputed value to replace each

missing observation in a completed dataset from the posterior predictive distribution $P(Y^{miss} \mid Y^{obs})$ (2.24) (Schafer 1997, p.72; Little and Rubin 2002, p.202). Data augmentation is run independently *K* times to give *K* completed datasets (Little and Rubin 2002, p.202).

Statistical Software for Implementing Data Augmentation

There are S-Plus functions for applying MCMC methods for basic models for continuous, categorical and mixed multivariate data as well as models with a more complicated structure, such as repeated measures (Schafer 1999). Similar R packages for general model fitting or for specific models could also be used. This method can be applied to data from a multivariate Normal distribution by using the MCMC statement in the *MI procedure* in SAS. Iterations are run between imputations in order that is reasonable to consider that data augmentation has been run independently *K* times. The multivariate Normal distribution may describe more than one type of assessment measured at the same time or the same assessment measured repeatedly over time. Other statistical software, for example MICE in Stata® (StataCorp LP) could also be used.

Example 2.8 Example of MCMC Methods: Data Augmentation More than One Type of Assessment

Let the coping score $Y_{h\text{sp}}$, mood score $Y_{h\text{mood}}$ and physical score $Y_{h\text{phys}}$ at the h^{th} assessment follow a multivariate Normal distribution as in Example 2.4. Thus, for the h^{th} assessment, $\boldsymbol{\theta}_h = (\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)$.

For the purposes of illustrating the MCMC method of data augmentation, let the following suppositions be made:

i) a single chain is used

ii) starting values and an informative prior, supposed to come from other trials, are used (e.g. $\boldsymbol{\mu}_2 = \begin{bmatrix} 42 \\ 32 \\ 28 \end{bmatrix}, \boldsymbol{\Sigma}_2 = \begin{bmatrix} 605 & 430 & 360 \\ 430 & 780 & 510 \\ 360 & 510 & 520 \end{bmatrix}$)

iii) the burn-in length is 500 iterations before the first imputation

iv) there are 5 imputations with 300 iterations between each imputation

In the *MI procedure*, the *prior* option in the MCMC statement can be used to specify a noninformative prior, an informative prior for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ or an informative prior for $\boldsymbol{\Sigma}$ in the P-step. The *initial* option can be used to specify the initial parameter values.

When considering Time 2 (h=2), in this example, the conditional predictive distribution (2.30) from which $Y_{2\text{miss}}^{(t+1)}$ is drawn in each I-step was a multivariate Normal. After d burn-in iterations, the values of the missing coping scores drawn in the I-step are considered as an approximate draw from the posterior predictive distribution $P(Y^{\text{miss}} \mid Y^{\text{obs}})$ (2.24) and thus are used as the imputed values in the first completed dataset. The *nbiter* option in the MCMC statement can be used to specify 500 burn-in iterations and the *round* option can be used to round the imputed coping scores to the nearest whole number. Let the superscript (1, 500) refer to the first imputation and 500 burn-in iterations. Considering Time 2 for this example, given the suppositions above, the conditional predictive distribution in the I-step for the first completed dataset was:

$$Y_{2\text{miss}}^{(1,500)} \sim \text{N}\left(\begin{bmatrix}64.6\\55.3\\46.2\end{bmatrix}, \begin{bmatrix}2158.5 & 2389.9 & 2012.4\\2389.9 & 2969.4 & 2462.9\\2012.4 & 2462.9 & 2201.7\end{bmatrix}\right) \qquad (2.32)$$

The imputed coping scores at Time 2 in the first completed dataset drawn from (2.32) were as shown in Table 2.4, column "Data Aug (Mult)".

The *niter* option in the MCMC statement can be used to specify 300 iterations between each draw of imputed values and the *nimpute* option can be used to specify 5 imputations. The imputed coping scores at Time 2 in the remaining completed datasets, again rounded to the nearest whole number, drawn from the respective conditional predictive distributions were as shown in Table 2.4, column "Data Aug (Mult)".

The missing coping scores at Time 3 and Time 4 are imputed similarly to missing coping scores at Time 2 (Table 2.4, column "Data Aug (Mult)").

Example 2.9 Example of MCMC Methods: Data Augmentation Repeated measures

Let the coping scores Y_h (h=1,..,4) follow a multivariate Normal distribution

$$\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{2.33}$$

Thus, $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

In contrast to Example 2.8, the coping scores over time follow a multivariate Normal distribution. The mood score and physical score are not considered. Let the suppositions in Example 2.8 be made again for the purposes of illustration and let the imputed coping scores be again rounded to the nearest whole number.

Given the suppositions made, suppose that the conditional predictive distribution in the I-step for the first completed dataset after 500 burn-in iterations was as follows:

$$Y_{\text{miss}}^{(1,500)} \sim \mathrm{N}\left(\begin{bmatrix}44.2\\41.9\\30.4\\27.3\end{bmatrix}, \begin{bmatrix}804.9 & 292.5 & 475.9 & 504.8\\292.5 & 417.7 & 307.8 & 338.3\\475.9 & 307.8 & 576.4 & 471.9\\504.8 & 338.3 & 471.9 & 573.7\end{bmatrix}\right) \tag{2.34}$$

As there are only a small number of patients in this example, draws from the conditional predictive distribution may lead to imputed coping scores outside the range of 0-100. In order to avoid this, the range of 0-100 was specified using the *minimum* and the *maximum* option. The imputed coping scores in the first completed dataset drawn from (2.34) were as shown in Table 2.4, column "Data Aug (Rep)".

After 300 iterations between imputations, the imputed coping scores in the remaining completed datasets drawn from the respective conditional predictive distributions were as in Table 2.4, column "Data Aug (Rep)".

2.3.2 Markov chain Monte Carlo Method of Gibbs' Sampling

Sampling from Full Conditional Distributions

A popular MCMC method is Gibbs' sampling, which is closely related to data augmentation. Like data augmentation, one run of Gibbs' sampling iterates to a draw from the posterior predictive distribution of Y^{miss} and a draw from the posterior distribution of $\boldsymbol{\theta}$. Here, the random vector X is divided into J subvectors.

$$X = X_1,\ldots,X_J$$

where J is generally greater than two (Schafer 1997, p.69).

The Gibbs' sampler eventually generates a draw from the distribution $P(x_1,.,x_j)$ of a set of J random variables $X_1,..X_J$, in contexts where draws from the joint distribution are hard to compute, but draws from conditional distributions $p(x_j \mid x_1,\ldots, x_{j-1}, x_j,..., x_J)$, $j = 1,..,J$ are relatively easy to compute (Little and Rubin 2002, p.204).

Initial values $x_1^{(0)}, x_2^{(0)},\ldots, x_J^{(0)}$ are selected. Then given values $x_1^{(t)}, x_2^{(t)},\ldots, x_J^{(t)}$ at iteration t, new values are found by drawing from the following sequence of J conditional distributions:

$$x_1^{(t+1)} \sim p(x_1 \mid x_2^{(t)}, x_3^{(t)}, \ldots, x_J^{(t)}$$

$$x_2^{(t+1)} \sim p(x_2 \mid x_1^{(t+1)}, x_3^{(t)}, \ldots, x_J^{(t)})$$

$$x_3^{(t+1)} \sim p(x_3 \mid x_1^{(t+1)}, x_2^{(t+1)}, x_4^{(t)}, \ldots, x_J^{(t)})$$

$$\ldots$$

$$x_J^{(t+1)} \sim p\left(x_J \mid x_1^{(t+1)}, x_2^{(t+1)}, \ldots, x_{J-1}^{(t+1)}\right)$$

It can be shown that the sequence of iterates $x^{(t)} = \left(x_1^{(t)},..,x_J^{(t)}\right)$ converges to a draw from the joint distribution of $X_1,..X_J$ (Little and Rubin 2002, p.204). After the iterations of the Gibbs' sampler are complete, the values Y^{miss} drawn from the posterior predictive distribution are the imputed values in the completed dataset to replace missing observations. The Gibbs' sampler is run independently K times to give K completed datasets (Little and Rubin 2002, p.204).

Implementing Gibbs Sampling

Gibbs' sampling can be applied using the WinBUGS software (Lunn et al. 2000) and has also been implemented in R (Wilkinson 2011). In SAS, the *genmode*, *lifereg* and *phreg* procedures update parameters using the Gibbs' sampler. The outline of the steps required to implement Gibbs sampling is as follows (Gilks et al. 1996, p.28):

i) Starting values must be provided for all unobserved nodes (parameters and missing data)
ii) Full conditional distributions for each unobserved node must be constructed and methods for sampling from them decided
iii) The output must be monitored to decide on the length of "burn-in" and the total run length and consideration given to whether a different parameterisation or MCMC algorithm is more appropriate
iv) Summary statistic for quantities of interest must be calculated from the output, for inference about the true values of the unobserved nodes

Gilks et al. (1996, p.28) recommend adding a fifth step:

v) Examine summary statistics for evidence of fit of the model

Special case of Gibbs' Sampler

Schafer (1997) notes that data augmentation is a special case of the Gibbs sampler when $X = (X_1, X_2)$ ($J = 2$) and with m=1. Little and Rubin (2002, p.204) note that when $J = 2$, the Gibbs' sampler is essentially the same as the MCMC data augmentation method described in section 2.3.1 if $X_1 = Y^{miss}$, $X_2 = \boldsymbol{\theta}$, $\boldsymbol{\theta}$ is the unknown model parameters and the distributions condition on Y^{obs}.

However, the Gibbs' sampler can be used in more complex problems where data augmentation is difficult to compute, but partitioning the missing data into more than one subvector aids computation. The special case noted by Little and Rubin (2002, p.204) can be implemented in statistical software in the same way as data augmentation.

Example 2.10 Example of MCMC Methods: Gibbs' Sampler

Let the coping score $Y_{h\text{sp}}$, mood score $Y_{h\text{mood}}$ and physical score $Y_{h\text{phys}}$ at the h^{th} assessment be assumed to follow a multivariate Normal distribution as in Example 2.4. The imputation model is a linear regression of $Y_{h\text{sp}}$ on $Y_{h\text{mood}}$ and $Y_{h\text{phys}}$. Let the random vector X be divided into two subvectors ($J = 2$), where $X_1 = Y^{\text{miss}}$ and $X_2 = \boldsymbol{\theta}$. In this special case, the imputation proceeds as in Example 2.8.

2.3.3 Approximate Bayesian Bootstrap

The approximate Bayesian bootstrap (ABB) method is an alternative to model-based imputation. Originally the ABB method was developed for the situation where the data is missing at random from simple random samples (Rubin 1987; Rubin and Schenker 1986; Rubin and Schenker 1991). It is a method for incorporating parameter uncertainty into hot-deck imputation (section 2.2.6). The concept of the ABB is to start with a set potential responses, selected at random with replacement from a subset of subjects with the same characteristics ($\boldsymbol{X}^*$) as those subjects with missing data. Then the imputed values are drawn from the set of potential values. Unlike in hot-deck imputation, where each imputed value is drawn from a single set of potential values, the set of potential values each imputed value is drawn from will be different for each repetition of ABB. The two steps are referred to as double sampling and ensures there is sufficient variability among the imputed values. The important assumption of the ABB method is that the data are missing at random (Fairclough 2010, chapter 9).

The procedure for the ABB method was defined formally by Fairclough (2010, p.194) as follows:

Repeat steps (i) to (iii) for k=1,..,K in order to generate K completed datasets

i) Consider a set of n subjects with the same characteristics, $\boldsymbol{X}$, where a univariate variable to be imputed was observed for n^{obs} patients and missing for n^{miss} patients. The expectation is that n^{obs} will be large compared to n^{miss}.

ii) Randomly select a set of n^{obs} possible values of Y^{miss} with replacement from the n^{obs} observed values for each of the n^{miss} missing values of Y^{miss}.

iii) Then choose n^{miss} observations at random with replacement which will replace the n^{miss} missing values of Y^{miss}. Each missing value of Y^{miss} has one value from the set of n^{obs} possible values selected at random with replacement as the imputed value of Y^{miss} in the k^{th} completed dataset.

Extending ABB to Longitudinal Variables

The ABB method can be extended to apply to a longitudinal variable to be imputed with a monotone missing data pattern. The extension assumes that the values of the variable to be imputed at time h, Y_h, are missing at random conditional on the patient characteristics $\boldsymbol{X}$ and $Y_1,\ldots,Y_{h-1}$ (Fairclough 2002, p.144). K completed datasets are generated with missing values of Y_h, at each time point $Y_1,\ldots,Y_H$ imputed. First, K sets of values for the first measurement of the variable to be imputed, Y_1, are imputed conditional on $\boldsymbol{X}$. Then, in order to impute the missing values of Y_2, patients are grouped by Y_1 and $\boldsymbol{X}$. Within these groups, for each repetition of imputation, a possible sample of n^{obs} observations of Y_2 is randomly selected from the observed values. Then n^{miss} imputed values are randomly selected. This procedure is repeated for each subsequent measurement of the variable to be imputed conditional on $\boldsymbol{X}$ and $Y_1,\ldots,Y_{h-1}$ for each of the K repetitions of imputation (Fairclough 2002, p.144).

However, this extension to the ABB has important limitations. The possible number of groups conditional on $\boldsymbol{X}$ and $Y_1,\ldots,Y_{h-1}$ increases with each procedure to impute missing observations of the h^{th} measurement of the variable to be imputed Y_h and this leads to there being not enough patients in a particular group to provide an adequate sample for this extension (Fairclough 2002, p.144).

Extending ABB to Informative Missing Data

Rubin and Schenker (1991) proposed a method to extend the ABB method to informative missing data. The particular case of informative missing data

considered is when high or low values of the variable of interest are more likely to be missing compared to the remaining values of the variable of interest. The method involves taking independent samples from the set of n^{obs} possible values, where the probability of choosing an observed value is proportional to a function of the variable to be imputed Y, for example Y^2. The aim of the method is to increase the proportion of large or small values included in the sample. Another possible method for increasing the proportion of large or small values included in the sample is to randomly sample l, where l is a small number such as 2 or 3, values from the set of n^{obs} possible values and to select the maximum or minimum of these observed values (Fairclough 2002, p.146). If an extension of the ABB method is applied when informative missing data are present, a sensitivity analysis involving variations in the specifications used in the methods applied should be carried out (Fairclough 2002, p.146).

Example 2.11 Example of Approximate Bayesian Bootstrap

In the example dataset (Table 2.1), the baseline mood score and core and physical score (Table 2.2) could potentially be used as the patient characteristics $\boldsymbol{X}$ to define the strata for ABB imputation. In this example, the coping score Y_h (h=1 for Time 1, h=2 for Time 2, h=3 for Time 3 and h=4 for Time 4) is a longitudinal variable.

For illustration, let the mutually exclusive strata for imputing the missing coping score at Time 2 be defined as: i) patients with baseline (Time 1) coping score $\leq$ 50 or baseline mood score $\leq$ 40 or baseline physical score $\leq$ 40 and ii) patients with baseline (Time 1) coping score > 50 and baseline mood score and physical score > 40. In assigning patients to the strata, the 5 monotone datasets are considered (see Table 2.1 and note above Table 2.4).

Consider the missing coping score at Time 2 for patient 1099. This patient is the one patient out of 7 in strata i) with a missing coping score at Time 2. From the 6 observed coping scores at Time 2 among patients in strata i), n^{obs} = 6 potential

coping scores to replace the missing coping score for patient 1099 were selected at random with replacement by *proc surveyselect*. The coping score selected by *proc surveyselect* to replace the missing coping score at Time 2 for patient 1099 was as shown in Table 2.4.

The missing coping scores at Time 3 are imputed similarly to the missing coping scores at Time 2. Let the strata defined also consider the coping score at Time 2 and be: i) patients with baseline (Time 1) coping score $\leq$ 50 or coping score at Time 2 $\leq$ 40 or baseline mood score $\leq$ 40 or baseline physical score $\leq$ 40 and ii) patients with baseline (Time 1) coping score > 50 and coping score at Time 2 > 40 and baseline mood score and physical score > 40.

The missing coping scores at Time 4 are imputed similarly to missing coping scores at Time 2. Let the strata defined also consider the coping score at Time 3 and be: i) patients with baseline (Time 1) coping score $\leq$ 50 or coping score at Time 2 $\leq$ 40 or coping score at Time 3 $\leq$ 30 or baseline mood score $\leq$ 40 or baseline physical score $\leq$ 40 and ii) patients with baseline (Time 1) coping score > 50 and coping score at Time 2 > 40 and coping score at Time 3 > 30 and baseline mood score and physical score > 40.

The imputed coping scores at Time 3 and Time 4 were as shown in Table 2.4.

2.3.4 Multiple Imputation Using Explicit Univariate Regression

Identification of Imputation Model

Multiple imputation using explicit univariate regression involves identifying a regression model to predict the missing observations. This regression model is similar to the regression model identified in simple imputation using linear regression models (section 2.2.4). The analytic model is as in (2.1) and the imputation model is as in (2.2). However, in the multiple imputation procedure, a random error is added to the estimated parameters $\widehat{\boldsymbol{B}}^*$, as the true parameters are unknown (Fairclough 2010, p.183). The *K* sets of estimated parameters, including

the random error, $\boldsymbol{\beta}^{(k)}$ k=1,..,K, are then used to predict the average value to be imputed for a patient with characteristics defined by the covariates $\boldsymbol{X}^{*\text{miss}}$. Additional random error is then added to these values to reflect the natural variability among patients of the variable to be imputed (Fairclough 2010, p.183).

The advantage of explicit univariate regression is again that the regression model can include additional information about the variable to be imputed not used in the analysis to compare treatment groups. For example, in clinical trials involving quality of life, this could be concurrent information such as adverse events or stage of disease. As in simple imputation using regression models, the covariates included in the regression model are likely to be strongly correlated with the variable being imputed and the probability that the observation of the variable is missing. The assumption that the data are missing at random should then be reasonable under the imputation model (Fairclough 2010, p.183-184).

However, multiple imputation using explicit univariate regression models has limitations. There must be sufficient patients to give precise estimates of the parameters in the imputation model. There must also be covariates which are strong predictors of the variable to be imputed. (Fairclough 2010, p.185-186). The underlying assumption for the imputation model (2.2) is that:

$$Y_i^* \sim \mathrm{N}(\boldsymbol{X}_i^{*T}\boldsymbol{B}^*, \sigma^{*2}) \tag{2.35}$$

The outline of the procedure for generating K sets of imputed values (Rubin 1987, Crawford et al. 1995, Little and Yau 1986) is described by Fairclough (2010, p.184-185) as follows:

i) Calculate the parameter estimates $\widehat{\boldsymbol{B}}^*$ and $\hat{\sigma}^{2*}$ using the observed data

ii) Generate the k^{th} set of regression model parameters $\boldsymbol{\beta}^{(k)}$ and mean square error $\sigma^{2(k)}$ by adding random error to the estimates to reflect the imprecision of the estimates

$$\sigma^{2(k)} = \hat{\sigma}^{2*}\frac{(n^{obs} - p^*)}{C^{(k)}} \tag{2.36}$$

$$\boldsymbol{\beta}^{(k)} = \widehat{\boldsymbol{B}}^{*} + \boldsymbol{U}_{\beta}\boldsymbol{Z}_{\beta}^{(k)} \quad (2.37)$$

where p^* is the number of unknown parameters in $\boldsymbol{B}^*$

$C^{(k)}$ is randomly drawn from a χ^2 distribution with $(n^{obs} - p^*)$ degrees of freedom

$\boldsymbol{U}_{\beta}$ is the upper triangular matrix of the Cholesky decomposition of the variance of $\widehat{\boldsymbol{B}}^{*}$

$\boldsymbol{Z}_{\beta}^{(k)}$ is a vector of p^* random numbers each drawn from a standard Normal distribution

iii) Generate the imputed values of Y_i^{miss} for the k^{th} imputation by adding random error corresponding to inter- and intra-patient variability of the variable to be imputed to the predicted value

$$Y_i^{miss(k)} = \boldsymbol{X}_i^{miss}\boldsymbol{\beta}^{(k)} + \sigma^{(k)}\boldsymbol{Z}_Y^{(k)} \quad (2.38)$$

where $\boldsymbol{Z}_Y^{(k)}$ is a random number drawn from a standard Normal distribution

Here, $\boldsymbol{X}_i^{miss}\boldsymbol{\beta}^{(k)}$ is the predicted value and $\sigma^{(k)}\boldsymbol{Z}_Y^{(k)}$ is the random error.

iv) Repeat step ii) and iii) for each of the K imputations

Fairclough (2010, p.185) notes that the procedure outlined assumes the parameters $\boldsymbol{\beta}^{(k)}$have a Normal distribution with variance approximately equal to $\sigma^{*2}\left(\sum \boldsymbol{X}_i^{T*obs}\,\boldsymbol{X}_i^{*obs}\right)^{-1}$, where $\boldsymbol{X}_i^{*obs}$ is the design matrix of covariates used in the imputation model (2.2) for patient i among the patients with an observed value of the response variable Y.

Step (i) in this context can be thought of obtaining values that characterise the joint posterior distribution of $\boldsymbol{B}^*$ and σ^{*2} under a conventional non-informative prior distribution (Gelman et al. 2004, chapter 14).
Step (ii) can be thought of as drawing $\boldsymbol{\beta}^{(k)}$ and $\sigma^{2(k)}$ from their joint observed-data posterior (2.25).

Step (iii) can be thought of as obtaining a set of imputed values for $Y_i^{miss(k)}$ by drawing from N($\boldsymbol{X}_i^{T*miss}\boldsymbol{\beta}^{(k)}$, $\sigma^{2(k)}I_{n_{miss}\,\mathrm{x}\,n_{miss}}$), where N($\boldsymbol{X}_i^{*Tmiss}\boldsymbol{\beta}^{(k)}$, $\sigma^{2(k)}I_{n_{miss}\,\mathrm{x}\,n_{miss}}$) is the (approximate) posterior predictive distribution (2.24) (Carlin 2014).

Therefore, explicit univariate regression is related to the MCMC method of data augmentation (section 2.3.1).

Extension to Longitudinal Trials

Little and Yau (1996) proposed an extension to the explicit univariate regression method for when the variable to be imputed is measured longitudinally. The proposed extension is a sequential procedure and applies to a monotone missing data pattern.

The procedure involves replacing missing observations of the first measurement (h=1) of the variable to be imputed (Y_1) with imputed values, generating K sets of imputed values for the missing observations of Y_1. The next step is to replace the missing observations of the second measurement (h=2) of the variable to be imputed (Y_2) given the observed or imputed value of the first measurement Y_1. In this step, a missing observation of Y_2 is replaced by one imputed value in each of the K datasets. The missing observations of the later measurements of the variable to be imputed are imputed using all previous observed or imputed values.

However, the extension to the explicit univariate regression model method for longitudinal trials requires at least two untestable assumptions. The first assumption is that the relationship between the variable to be imputed and the variables included in the regression model is the same for patients who complete all measurements and patients who do not. The second assumption is that all the relevant covariates are included in the imputation model so the assumption that the data are missing at random under the imputation model is reasonable. The explicit univariate regression model method also requires an assumption that the

residual errors and the parameter estimates of the imputation method follow a Normal distribution (Fairclough 2010, p.187).

Implementing Explicit Univariate Regression

Explicit univariate regression can be applied by using the *reg* option in the monontone statement in the *MI procedure* in SAS. Each variable is transformed into a N(0, 1) distribution (standardised). The calculations in the *MI procedure* are based on the standardised data.

Example 2.12 Example of Explicit Univariate Regression

To simplify illustration, only the previous coping scores were included in the imputation models for missing coping scores. The imputation model for the missing coping score for patient 1099 at Time 2 contains the coping score at baseline (Time 1):

$$Y_{i2}^{*obs} = \beta_{2inter}^{*} + Y_{i1}^{(k)obs}\beta_{2|1}^{*} + \varepsilon_{i2}^{*} \qquad (2.39)$$

where β_{2inter}^{*} is the intercept.

For example, for the first monotone dataset (see Table 2.1 and note above Table 2.4), $\widehat{\boldsymbol{B}}_2^{*} = \begin{bmatrix} -0.095 \\ 0.711 \end{bmatrix}$ and $\hat{\sigma}^{2*}$= 0.607

The second step of generating the model parameters for imputation requires:

i) the upper triangular matrix of the Cholesky decomposition of the variance of $\widehat{\boldsymbol{B}}^{*}$, $\boldsymbol{U}_{\beta_2}^{(k)}$ (e.g. $\boldsymbol{U}_{\beta_2}^{(1)} = \begin{bmatrix} 0.263 & -0.042 \\ 0.000 & 0.283 \end{bmatrix}$)

ii) random draws $C_2^{(k)}$ drawn from the χ_7^2 distribution (e.g. $C_2^{(1)}$= 5.414)

iii) random draws $\boldsymbol{Z}_{\beta_2}^{(k)}$ of 2 random numbers, both drawn from a standard Normal distribution (e.g. $\boldsymbol{Z}_{\beta_2}^{(1)} = \begin{bmatrix} 0.400 \\ 0.388 \end{bmatrix}$)

The third step of generating the imputed standardised value for the missing coping score for patient 1099 at Time 2 requires random draws $Z_{Y_2}^{(k)}$ from a standard Z_{Y_2}

Normal distribution (e.g. $Z_{Y_2}^{(1)}$ = 1.805). The imputed coping score in standardised form in step iii), is converted to the original scale (e.g. 0.609 is converted to 57). The imputed coping score in the original scale, again rounded to the nearest whole number, was as shown in Table 2.4.

The imputation model for the missing coping scores at Time 3 contains the coping score at baseline (Time 1) and Time 2.

$$Y_{i3}^{*obs} = \beta_{3inter}^{*} + Y_{i1}^{(k)obs}\beta_{3|1}^{*} + Y_{i2}^{(k)obs}\beta_{3|2}^{*} + \varepsilon_{i3}^{*} \quad (2.40)$$

When imputing missing coping scores at Time 3, observed and imputed coping scores at Time 2 are considered as observed coping scores. The missing coping scores at Time 3 are imputed similarly to missing coping scores at Time 2. Again, due to the small number of patients in this example, the range of 0-100 was specified using the minimum and the maximum option. The imputed coping scores at Time 3 were as shown in Table 2.4.

The imputation model for the missing coping scores at Time 4 contains the coping score at baseline (Time 1), Time 2 and Time 3.

$$Y_{i4}^{*obs} = \beta_{4inter}^{*} + Y_{i1}^{(k)obs}\beta_{4|1}^{*} + Y_{i2}^{(k)obs}\beta_{4|2}^{*} + Y_{i3}^{(k)obs}\beta_{4|3}^{*} + \varepsilon_{i4}^{*} \quad (2.41)$$

The missing coping scores at Time 4 are imputed similarly to missing coping scores at Time 2 (Table 2.4).

2.3.5 Nearest Neighbour and Predictive Mean Matching

Nearest neighbour (Rubin 1987, chapter 5; Van Buuren et al. 1999) and predictive mean matching (Rubin 1987, p.168, Rubin and Schenker 1991; Heitjan and Landis 1994) were developed from the explicit univariate regression method. The advantage of nearest neighbour and predictive mean matching is that is not possible to impute a value out of the range for the scale of the variable to be imputed. An important assumption for both methods is that the relationship between the explanatory variables and the variable of interest is the same when

the variable of interest is observed and when missing (Fairclough 2010, p.187). The data are assumed to be missing at random. These methods may also be robust to departures from the assumption that the residual errors and the parameter estimates of the imputation model follow a Normal distribution (Fairclough 2010, p.187).

Procedure for Nearest Neighbour Imputation

The outline of the procedure for generating *K* sets of imputed values is outlined by Fairclough (2010, p.192) as follows:

The initial steps are the same as for explicit univariate regression in section 2.3.4.

i) Estimate the parameters of the regression model using the observed data
ii) Generate the k^{th} set of regression model parameters $\boldsymbol{\beta}^{(k)}$ and $\sigma^{2(k)}$ by adding random error to the estimates to reflect the imprecision of the estimates

The next steps are:

iii) Generate predicted values for patients with both observed values and missing observations
iv) For each patient with a missing observation, identify the patient with the closest predicted value and impute the observed value for this closest patient for the missing observation
v) Repeat steps (ii) to (iv) for each of the *K* imputations

Nearest neighbour imputation can be applied by using the *regpredmeanmatch* option in the *monotone* statement in the *MI procedure* in SAS. The *K* option in the *monotone* statement is set to 1 to identify the nearest neighbour. The *regpredmeanmatch* option can be shortened to *regpmm*.

Procedure for Predictive Mean Matching Imputation

The initial steps for predictive mean matching can be based on bootstrap samples (Heitjan and Landis 1991). The outline of the procedure for generating *K* sets of imputed values is described by Fairclough (2010, p.192-193) as follows:

i) Generate *K* bootstrap samples, by sampling with replacement, from patients with observed values
ii) For each of the *K* bootstrap samples, calculate the model parameters $\boldsymbol{\beta}^{(k)}$
iii) Generate predicted values for patients with both observed values and missing observations
iv) For each patient with a missing observation, identify the five patients with an observed value with the closest predicted value
v) Select one of the five patients with the closest predicted value at random and impute the observed value for this selected patient for the missing observation

Alternatively, the initial steps can follow the same procedure as for nearest neighbour imputation (Rubin 1987, p.168).

Predictive mean matching can be applied by using the *regpredmeanmatch* option in the monotone statement in the *MI procedure* in SAS. The initial steps follow the same procedure as for nearest neighbour imputation. The number of potential coping scores to be imputed considered in step (iv) is defined by the *K* option in the *monotone* statement.

Example 2.13 Example of Nearest Neighbour Imputation

The initial steps in nearest neighbour imputation are the same as explicit univariate regression in section 2.3.4. For illustration, when considering Time 2, let the initial steps for the imputation of the missing coping for patient 1099 proceed as in Example 2.12. In steps iii) and iv), the predicted values are generated and the nearest neighbour identified. For example, for the first monotone dataset, the standardised predicted value for patient 1099 was -0.99. The nearest neighbour in is then patient 6, who has a standardised predicted value of -0.683. The imputed coping score at Time 2 for patient 1099 was 32 as shown in Table 2.4.

When imputing missing coping scores at Time 3, observed and imputed coping scores at Time 2 are considered as observed coping scores. Similarly, when imputing missing coping scores at Time 4, observed and imputed coping scores at Time 2 and Time 3 are considered as observed coping scores. The missing coping scores at Time 3 and Time 4 are imputed similarly to missing coping scores at Time 2 (Table 2.4).

Example 2.14 Example of Predictive Mean Matching Imputation

Initial Steps based on a Bootstrap Sample

As noted, the first step is to generate *K* bootstrap samples, by sampling with replacement, from patients with observed value. Here, a bootstrap sample was generated from the patients with observed or imputed coping scores with replacement in each of the datasets with a monotone missing data pattern. In this example, the parameter values for the imputation model in step ii) are calculated based on the standardised data (e.g. $\boldsymbol{\beta}_2^{(1)} = \begin{bmatrix} 0.000 \\ 0.733 \end{bmatrix}$). This ensures that the predicted values in step (iii) were all within the range of the standardised value for a coping score of 0 and a coping score of 100 in the corresponding monotone dataset.

Initial Steps based on Procedure for Nearest Neighbour Imputation

See section 2.3.4 and Example 2.13. For illustration, let the parameter values in step (ii) be the same as in the initial steps based on a bootstrap sample in Example 2.14. It then follows the imputed values are identical for both sets of initial steps.

Predicted and Imputed Coping Scores at Time 2

In step (iii), the predicted coping scores are as calculated based on the parameter values for the imputation models in step ii). For example, the patients with the nearest predicted scores for patient 1099 in the first imputation were:
Patients 6, 456, 635, 828, 2237.

The coping scores selected at random to replace the missing coping scores for patient 1099 at Time 2 were as shown in Table 2.4. For example, the coping score selected in the first completed dataset was from patient 2237.

Coping Scores at Time 3 and Time 4

When imputing missing coping scores at Time 3, observed and imputed coping scores at Time 2 are again considered as observed coping scores. Similarly, when imputing missing coping scores at Time 4, observed and imputed coping scores at Time 2 and Time 3 are considered as observed coping scores. The missing coping scores at Time 3 and Time 4 are imputed similarly to missing coping scores at Time 2 (Table 2.4).

2.3.6 Pattern Mixture Models - Curran's Analytic Technique

Pattern mixtures models (Little 1993; Little 1994; Little 1995) are a modelling approach proposed to analyse informative missing data. The method stratifies incomplete data by the pattern of missing values and formulates distinct models within each stratum (Little and Wang, 1996). It is common that the parameters for many of these models can only be estimated by imposing restrictions (Fairclough 2010, p. 213). Thus, restrictions have been proposed for longitudinal data with monotone missing data. These are the complete case missing value restriction,

available case missing value restriction and neighbouring case missing value restriction. An analytic technique using multiple imputation has been proposed by Curran (2000) for applying such sets of restrictions.

Methodology of Pattern Mixture Models

The methodology of pattern mixture models is based on there being *P* different missing data patterns. The part of the pattern mixture model specifying the missing data mechanism (*f*[**M**]) is independent of the missing observations. Thus, it is not necessary to specify how the missing data mechanism for the variable to be imputed depends on the missing observations. The advantage of pattern mixture models is that only the proportion of patients with each pattern of missing data is required. However, the disadvantages of the pattern mixture models are the large number of potential patterns of missing data and the difficulties in estimating all the parameters in each model (Fairclough 2010, p. 213).

Outline of Procedure for Pattern Mixture Models

Among the *P* missing data patterns, there may be different distributions of the responses, Y_i, with different parameters, $\boldsymbol{\beta}^{\{p\}}$, and variance $\Sigma^{\{p\}}$:

$$\boldsymbol{Y}_i | M^{\{p\}} \sim N\left(\boldsymbol{X}_i \boldsymbol{\beta}^{\{p\}}, \Sigma_i^{\{p\}}\right), p = 1 \ldots, P \tag{2.42}$$

where is $\boldsymbol{X}$ is the design matrix of covariates

$M^{\{p\}}$ is the missing data mechanism for the p^{th} pattern

$\boldsymbol{R}_i \in M^{\{p\}}$ where $\mathbf{R}_i$ is a vector of indicators of the missing data pattern for the i^{th} patient and is identical for all patients in the p^{th} pattern

The true distribution of the response variable for all patients is a mixture of the *P* distributions from each group of patients. Therefore, the expected values of the parameters averaged over the *P* missing data patterns,

$$E[\boldsymbol{\beta}] = \sum_{p=1}^{P} \pi^{\{p\}} \boldsymbol{\beta}^{\{p\}} \tag{2.43}$$

where $\pi^{\{p\}}$ is the proportion of patients with the p^{th} pattern, is of interest.

The first step in determining the expected values of the parameters averaged over the P missing data patterns is to stratify the patients by the missing data pattern. Then, for each strata, the parameters ($\boldsymbol{\beta}^{\{p\}}$, $\Sigma^{\{p\}}$) are estimated. The proportion of patients in each strata estimates the weights for averaging the parameters over the P missing data patterns,

$$\hat{\pi}^{\{p\}} = n^{\{p\}}/N \tag{2.44}$$

The parameters for the true distribution of the response variable for all patients is estimated by the average of the estimates of P missing data patterns.

Underidentification of Missing Data Patterns

Unfortunately, problems with this method may arise. A possible problem is having a large number of missing data patterns corresponding to unusual scenarios with a small number of patients. Another possible problem is that for many missing data patterns the model is underidentified and so it is not possible to estimate all of the parameters $\boldsymbol{\beta}^{\{p\}}$ without making further assumptions. In this situation, explicit restrictions, such as complete case restrictions (Little 1993), have been proposed.

Complete case missing variable (CCMV) restriction for Bivariate Data

Considering an example dataset with two assessments of coping score (Y_1 and Y_2), there are 4 possible missing data patterns, including no observed coping scores. These four patterns are described below, considering both observed coping scores as pattern 1 and with 1 indicating an observed coping score and 0 indicating a missing coping score:

Table 2.5 Possible Missing Data Patterns in an Example Dataset with Two Assessments of Coping Score

Pattern	Time 1	Time 2
1	1	1
2	1	0
3	0	1
4	0	0

Let the subscript 1 and 2 refer to Time 1 and Time 2 respectively and the superscript *p* refer to the pattern being considered. Considering a cell means model (section 1.6.4), for each treatment group, in each of the four patterns, there are five possible parameters to be estimated: two means ($\hat{\mu}_1^{\{p\}}, \hat{\mu}_2^{\{p\}}$) and three parameters for the covariance ($\hat{\sigma}_{11}^{\{p\}}, \hat{\sigma}_{12}^{\{p\}}, \hat{\sigma}_{22}^{\{p\}}$). Thus, the model is underidentified as not all of these 20 parameters can be estimated from the data (Fairclough 2010, p.226). The CCMV restriction assumes that the missing value distributions are equal to the complete case distributions. Let the notation $\theta^{\{p\}}$ refer to the 5 possible parameters for the *p*th pattern. For the bivariate example (Table 2.5), the restrictions are:

$$\theta_{[2.1]}^{\{2\}} = \theta_{[2.1]}^{\{1\}}, \theta_{[1.2]}^{\{3\}} = \theta_{[1.2]}^{\{1\}} \text{ and } \theta^{\{4\}} = \theta^{\{1\}} \qquad (2.45)$$

where $\theta_{[2.1]}^{\{1\}}$ denotes the parameters from the regression of Y_2 on Y_1 using the complete cases in pattern 1

$\theta_{[2.1]}^{\{2\}}$ denotes the parameters from the regression of Y_2 on Y_1 for cases in pattern 2

$\theta_{[1.2]}^{\{1\}}$ denotes the parameters from the regression of Y_1 on Y_2 using the complete cases in pattern 1

$\theta_{[1.2]}^{\{3\}}$ denotes the parameters from the regression of Y_1 on Y_2 for cases in pattern 3. (Fairclough 2010, p.227).

Consider the special case where there is a monotone missing data pattern excluding no observed assessments (pattern 1 and 2 in Table 2.5). The CCMV restriction can be applied to estimate $\mu_2^{\{2\}}$ using the regression of the Y_2 on Y_1. In

this special case, the parameter estimates are the same as the maximum likelihood estimates using all available data (Fairclough 2012, p.227).

Longitudinal Studies with Monotone Dropout

For longitudinal trials with monotone dropout, three set of restrictions have been proposed. These restrictions are the extension to the CCMV restriction proposed by Little (1993) for the bivariate case, the available case missing value restriction and the neighbouring case missing value restriction. Curran (2000) proposed an analytic technique for these restrictions using multiple imputation by the MCMC method of data augmentation (section 2.3.1). As with MCMC method of data augmentation, Curran's analytic technique can be applied to data from a multivariate Normal distribution by using the MCMC statement in the *MI procedure* in SAS.

Considering an example dataset with four assessments of coping score ($Y_1 - Y_4$) and excluding the possibility of no observed coping scores, there are four monotone missing data patterns. These four patterns are described below in Table 2.6, considering four observed coping scores as pattern 1.

Table 2.6 Monotone Missing Data Patterns in an Example Dataset with Four Assessments of Coping Score

Pattern	Time 1	Time 2	Time 3	Time 4
1	1	1	1	1
2	1	1	1	0
3	1	1	0	0
4	1	0	0	0

Similarly to the bivariate example (Table 2.5), a cell means model with different means and covariances in each pattern is underidentified (Fairclough 2010, p.226). Extending the notation from the bivariate case, in the description of the restrictions proposed, $\theta^{\{1\}}_{[34.12]}$ 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

1. Curran's analytic technique for these restrictions can be used to create completed datasets with missing coping scores replaced.

Complete Case Missing Value Restriction (CCMV)

Under the CCMV restriction, the data for patients with complete data are used to predict the means for the missing observations in the remaining patterns. The assumption is that the missing observation distributions are equal to the complete case distributions. This restriction is only feasible when there are sufficient patients in pattern 1 to estimate these parameters reliably (Fairclough 2010, p.232). In the example (Table 2.6), the restrictions are as follows:

$$\theta_{[4.123]}^{\{2\}} = \theta_{[4.123]}^{\{1\}} \quad (2.46)$$

$$\theta_{[34.12]}^{\{3\}} = \theta_{[34.12]}^{\{1\}} \quad (2.47)$$

$$\theta_{[234.1]}^{\{4\}} = \theta_{[234.1]}^{\{1\}} \quad (2.48)$$

Available Case Missing Value Restriction

Under the available case missing value restriction, data from patients in all patterns with observed values are used to predict the means for the missing observations in the remaining patterns. The assumption is that the missing observation distributions are equal to the available case distributions. This method is more feasible than the CCMV restriction because there is more data to estimate the parameters (Fairclough 2010, p.232). In the example (Table 2.6), the restrictions are as follows:

$$\theta_{[4.123]}^{\{2\}} = \theta_{[4.123]}^{\{1\}} \quad (2.49)$$

$$\theta_{[4.123]}^{\{3\}} = \theta_{[4.123]}^{\{1\}}, \theta_{[3.12]}^{\{3\}} = \theta_{[3.12]}^{\{1,2\}} \quad (2.50)$$

$$\theta_{[4.123]}^{\{4\}} = \theta_{[4.123]}^{\{1\}}, \theta_{[3.12]}^{\{4\}} = \theta_{[3.12]}^{\{1,2\}}, \theta_{[2.1]}^{\{4\}} = \theta_{[2.1]}^{\{1,2,3\}} \quad (2.51)$$

Neighbouring Case Missing Value Restriction

Under the neighbouring case missing value restriction (NCMV), available data from patients in the neighbouring pattern are used to impute the means for the

missing observations. This may be the most useful of the three sets of restrictions proposed for longitudinal data (Fairclough 2010, p.234). The assumption is that the missing observations are MAR conditional on the nearest neighbouring pattern. This may be hard to justify for later assessments. For example, the assumption to restrict the parameters from a quality of life assessment at Time 4 to be the same among patients who dropped out after the assessment at Time 1 as for patients with 4 complete assessments may not be justifiable. It may be more realistic that patients who dropout early have worse quality of life than the patients with complete assessments. The assumption in NCMV implies that the relationship between assessments is similar for patients who have complete observations and those who dropout early in the study (Fairclough 2010, p.234).

In the example (Table 2.6), the restrictions are as follows:

$$\theta_{[4.123]}^{\{2\}} = \theta_{[4.123]}^{\{1\}} \tag{2.52}$$

$$\theta_{[4.123]}^{\{3\}} = \theta_{[4.123]}^{\{1\}}, \theta_{[3.12]}^{\{3\}} = \theta_{[3.12]}^{\{2\}} \tag{2.53}$$

$$\theta_{[4.123]}^{\{4\}} = \theta_{[4.123]}^{\{1\}}, \theta_{[3.12]}^{\{4\}} = \theta_{[3.12]}^{\{2\}}, \theta_{[2.1]}^{\{4\}} = \theta_{[2.1]}^{\{3\}} \tag{2.54}$$

Fairclough (2010, p.234) describes the procedures for Curran's analytic technique as follows for each of the *K* completed datasets:

i) Impute the missing values at Time 2 in pattern 4 based on patients in pattern 3 and 4
ii) Combine the observed and imputed values from step i) with the information for patients in pattern 2
iii) Impute the missing values at Time 3 based on patients in pattern 2, 3 and 4. For patients in pattern 4, imputed values at Time 2 are considered in the calculation (Equation 2.52).
iv) Combine the observed and imputed values from step iii) with the information for patients in pattern 1
v) Impute the missing values at Time 4 based on patients in pattern 1. For patients in pattern 4, observed and imputed values at Time 3 and Time 2

are considered in the calculation (Equation 2.52). For patients in pattern 3, imputed values at Time 3 are considered in the calculation (Equation 2.51).

Then analyse each completed dataset and finally combine estimates as for multiple imputation techniques.

Example 2.15 Example of Pattern Mixture Models – Curran's Analytic Technique for Neighbouring Case Missing Value Restriction

Let the coping scores Y_h (h=1,..,4) follow a multivariate Normal distribution as in Example 2.9 and let the suppositions in Examples 2.8 and 2.9 be made again for the purposes of illustration. Again, let the imputed coping scores be rounded to the nearest whole number and the range be specified as 0-100.

As the basis of Curran's analytic technique is a cell means model (section 1.6.4), each treatment group is considered separately. Considering four observed coping scores as pattern 1, the 5 datasets with a monotone missing data pattern have 4 missing data patterns as follows:

Table 2.7 Missing Data Patterns for Datasets with Monotone Missing Data Pattern in Example of Pattern Mixture Models

Pattern	Time 1	Time 2	Time 3	Time 4	Patients in Treatment Group A	Patients in Treatment Group B
1	1	1	1	1	828, 1304	6, 456, 1728, 2509
2	1	1	1	0	2237	47
3	1	1	0	0	635	
4	1	0	0	0	1099	

Consider the patients in patterns 3 and 4, for which only coping scores at Time 1 and Time 2 are relevant (step i). In this example, only treatment group A is relevant and there are insufficient patients in patterns 3 and 4 to illustrate step i) directly (Table 2.7). However, for the purposes of illustration, let the example patients represent a larger dataset obtained from the full IBCSG dataset, ignoring intermittent coping scores to give a monotone missing data pattern. Imputation of

missing coping scores at Time 2 then proceeds similarly to Example 2.10. For example, using similar notation to Example 2.9 and Example 2,10, the relevant conditional predictive distributions for the first completed dataset were as follows:

$$Y_{\text{A2miss}}^{(1,500)} \sim \text{N}\left(\begin{bmatrix}45.0\\40.4\end{bmatrix}, \begin{bmatrix}534.9 & 83.8\\83.8 & 628.3\end{bmatrix}\right) \quad (2.55)$$

The coping scores drawn from the conditional predictive distributions to replace the missing coping score at Time 2 for patient 1099 were as in Table 2.4.
The observed and imputed coping scores for patients in patterns 3 and 4 are then combined with information for patients in pattern 2. For these patients, coping scores at Time 1, Time 2 and Time 3 are relevant. Again, only treatment group A is relevant in this example (Table 2.7). The coping scores drawn from the respective conditional predictive distribution to replace the missing coping scores at Time 3 for patient 1099 and patient 635 were as shown in Table 2.4.

The observed and imputed coping scores for patients in patterns 2, 3 and 4 are then combined with information for patients in pattern 1. The missing coping scores at Time 4 in treatment group A and treatment group B are imputed similarly to missing coping scores at Time 3 (Table 2.4).

2.3.7 Further Development and Chained equations

It likely that, given the increasing availability of multiple imputation methods in standard statistical packages and increasing understanding of its potential advantages, the use of multiple imputation in clinical trials will increase (Kenward and Carpenter 2007). The implementation of imputation techniques in statistical software is becoming more automatic. This may require new forms of model selection and diagnostic procedures, for example to detect influential observations and the imputation of impossible values (Kenward and Carpenter 2007).

When multiple imputation is applied, the sensitivity of the results to the assumptions of the imputation method is often explored. In such sensitivity analysis, it is important to be clear what assumptions and what variations of these

assumptions are being considered. Thus, further development of models which can easily be communicated for appropriate sensitivity analysis is required (Kenward and Carpenter 2007).

As the chained equations method is of increasing importance, a more formal justification of this method is of particular interest (Kenward and Carpenter 2007). The basic concept of the method is that it uses a series of univariate conditional models in the spirit of the Gibbs' sampler (section 2.3.2). This means it is also related to the MCMC method of data augmentation (section 2.3.1), multiple imputation using explicit univariate regression (section 2.3.4), nearest neighbour imputation and predictive mean matching (section 2.3.5). It was developed to address the difficulty of constructing an appropriate imputation model for a combination of types of variables. The method avoids specifying an appropriate joint imputation distribution by replacing this by the selection of appropriate univariate conditional distributions. While the chained equations method has been used successfully, the properties of conditional distributions is an area for further research (Kenward and Carpenter 2007).

The procedure starts with an appropriate univariate conditional distribution for each variable to be imputed and imputes in turn from these distributions. At the end of one cycle, which considers all variables to be imputed, imputations will exist for all the missing observations. The whole cycle is repeated several times, taking the draws from the last cycle to form the first imputed dataset in line with the Gibbs' sampler. The contrast to genuine Gibbs' samplers is that only a comparatively few cycles are used, generally 10 or 20 (Kenward and Carpenter 2007). Then the whole process is repeated to obtain further completed datasets. This approach has been implemented by in the statistical software MICE (van Buuren et al al. 1999) and IVEWARE (Raghunathan TE et al. 2001; Taylor et al. 2002).

2.3.8 Summary of Multiple Imputation

Multiple imputation techniques have proved useful in the context of clinical trials and multiple imputation of quality of life assessments has been used successfully (e.g. Bordeleau et al. 2003; Stanton et al. 2005; Peyre et al. 2011). An example application in this context is investigating the relationship between quality of life and prognosis in breast cancer patients (see Chapter 3). Potentially, several standard imputation methods could be applied to missing quality of life assessments as part of such an investigation. However, this would not be useful if the completed datasets are similar regardless of which imputation method is applied. In this scenario, the imputation method applied would have little influence on the parameter estimates from the analytic model.

Standard multiple imputation methods were illustrated using a small example dataset. Considering the imputed values in Table 2.4 indicated that the MCMC methods (data augmentation with more than one type of assessment / MCMC methods – Gibbs' sampling and pattern mixture models - Curran's analytic technique) may be more similar to each other than the remaining methods. However, as with simple imputation, there was no suggestion that the completed datasets are similar regardless of which imputation method is applied (Table 2.4).

Explicit univariate regression and ABB were developed from standard simple imputation techniques. Nearest neighbour and predictive mean matching where then developed from explicit univariate regression and are closely related. These multiple imputation methods were not originally developed for longitudinal data but can be extended to longitudinal data. Some of the standard multiple imputation methods, for example explicit univariate regression, are based on an explicit parametric Bayesian model for the imputation model. In the case of MCMC methods and pattern mixture models - Curran's analytic technique, the assumption is that the data follow a multivariate Normal distribution. Other methods, for example nearest neighbour imputation, are based on an algorithm implying an underlying imputation model. In general the standard multiple

imputation methods assume a monotone missing data pattern, though the MCMC methods extend to a general missing data pattern. The MCMC methods of data augmentation and Gibbs sampling are closely related.

Many of the standard multiple imputation methods, for example ABB, assume the data are MAR. Rubin and Schenker (1991) have proposed an extension of ABB to a particular case of informative missing data. Pattern mixture models were proposed to address informative missing data for longitudinal data. Curran's analytic technique for implementing pattern mixture models is the only standard multiple imputation method based on the cell mean model.

2.4 Summary

This review of standard imputation methods noted that simple imputation methods, where a single value is imputed, can provide useful information as part of a sensitivity analysis in clinical trials. Multiple imputation, where K ($K > 1$) values are imputed is a more useful method, although more complex. It takes into account the variability induced by imputation.

The standard imputation methods differ in whether the imputation model is based on i) an explicit parametric Bayesian model or ii) on an algorithm implying an underlying imputation model. In general the standard multiple imputation methods assume a monotone missing data pattern, though the MCMC methods extend to a general missing data pattern. Many of the standard multiple imputation methods assume the data are MAR. In contrast, pattern mixture models were proposed to address informative missing data for longitudinal data. When applying imputation, it is important to investigate the missing data mechanism and to carry out a sensitivity analysis of the results to specific assumptions about the missing observations.

The potential advantages of multiple imputation and the increasing availability of statistical software to implement multiple imputation methods suggest that

multiple imputation will become increasingly important in breast cancer clinical trials. This particularly applies to quality of life assessments which, while becoming increasingly common in breast cancer trials, are associated with a high level of missing observations (Fairclough 2010, chapters 1 and 6).

Imputation techniques provide a potential tool for investigating the question of whether quality of life is related to prognosis in breast cancer patients. As noted, several standard imputation methods could be applied as part of such an investigation. The illustration of the standard imputation methods in a small example dataset indicated that using several standard imputation methods may provide useful information in this investigation. A specific application is replacing missing quality of life assessments before the time-dependent Cox model analysis of DFS with quality of life as a time-dependent covariate.

Thus, in the next chapters, the missing quality of life assessments as measured by coping score in the IBCSG dataset are imputed by several standard imputation methods. These imputed values are considered in the time-dependent Cox model analysis with coping score as a time-dependent covariate. The hypothesis investigated is that quality of life throughout the study is associated with prognosis. The efficacy endpoint of DFS was not considered in the imputation of coping scores. Preliminary to analysis, the assumption of whether the data are MAR is reasonable is considered. The results from the time-dependent Cox model analysis following standard imputation methods are presented and the performance of the standard imputation methods compared. The standard imputation methods are compared by i) considering the parameter estimates for quality of life and the corresponding standard errors and ii) by using simulated datasets to estimate the difference between the coping score and the missing coping score.

Simple Imputation

As described in this chapter, there are several standard methods of simple imputation. Given the limitations of the standard simple imputation methods noted, it was not of interest to apply all standard simple imputation methods to the IBCSG dataset. In particular, cold deck imputation was noted to be only of historical interest. However, as part of a sensitivity analysis investigating the question of whether quality of life is related to prognosis, it was of interest to apply selected simple imputation methods. Therefore, 3 standard imputation methods, were selected representing the main approaches to simple imputation in the statistical literature to form part of sensitivity analysis. These 3 methods include LOCF as a simple imputation method commonly applied and are:

i) last observation carried forward
ii) mean or median imputation: median by patient, by time period and by time period and treatment group
iii) imputation using regression models: linear regression with previous coping score(s) and with concurrent variables

Extreme imputation of highest coping score of 100 (lowest quality of life) and lowest coping score of 0 (highest quality of life) was applied to illustrate the worst possible and best possible complete set of quality of life assessments in the IBCSG dataset. The remaining simple imputation methods described in this chapter are not applied to the IBCSG dataset in Chapter 3.

Multiple Imputation

There are two standard multiple imputation methods described in this chapter that are not applied to the IBCSG dataset in Chapter 4:

i) explicit univariate regression
ii) MCMC methods

MCMC methods are more computationally complicated than other standard multiple imputation methods described in this chapter and are not applied to the

IBCSG dataset. Simple imputation using linear regression models is applied to the IBCSG dataset and it did not appear that explicit univariate regression would add any additional information about the question of the whether quality of life is related to the prognosis.

The IBCSG dataset had a general missing data pattern which was not close to monotonic. However, it is possible to ignore the non-monotone missing data pattern when identifying the subgroups of patients considered in the bootstrapping method. This allowed the question of whether quality of life is related to prognosis to be investigated by a version of bootstrapping in the IBCSG dataset without first creating a monotone missing data pattern. In order to apply nearest neighbour imputation, predictive mean matching (initial steps as for nearest neighbour imputation) and pattern mixture models – Curran's analytic technique to the IBCSG dataset, non-monotone missing coping scores were imputed by LOCF.

Therefore, the four standard multiple imputation methods applied to the IBCSG dataset in Chapter 4 are:

i) bootstrapping: subgroups defined by baseline coping score and subgroups defined by previous coping score
ii) predictive mean matching
iii) nearest neighbour imputation
iv) pattern mixture models – Curran's analytical technique

3 Investigation of the Effects of Using Simple Imputation Methods to Estimate the Effect of Quality of Life on Disease-Free Survival in IBCSG Trials VI and VII

3.1 Introduction

In the previous chapter, several methods of imputation were reviewed and illustrated using a small example dataset. In this chapter the association between quality of life and DFS in IBCSG Trials VI and VII is investigated. Many studies have shown that a diagnosis of breast cancer can negatively affect a woman's quality of life, but whether the resulting quality of life is associated with survival remains under debate (Epplein et al. 2011). Studies have reported that social well-being in the first year after cancer recurrence is a significant prognostic factor for recurrence or mortality (Epplein et al. 2011) and that changes in quality of life during adjuvant therapy may be associated with recurrence (Keene Sarenmalm et al. 2009).

Previous work by Coates et al. (2000) indicated that in the IBCSG dataset DFS was not significantly predicted by quality of life scores at baseline or month 18, or by changes in quality of life score between baseline and months 3 or 18. However, Herring et al. (2004) indicated that quality of life at baseline was associated with prognosis in postmenopausal patients. As noted, this may reflect the toxicity of chemotherapy treatment. Here, previous work is extended by considering quality of life as a time-dependent effect. The hypothesis investigated is that poorer quality of life throughout the study is associated with poorer DFS and conversely better quality of life throughout the study is associated with better DFS.

In section 3.2, IBCSG Trials VI and VII are described together with a brief summary of the main published analyses on efficacy and quality of life of these trials. The reason for imputing missing values before the time-dependent Cox

model analysis is discussed in section 3.3. The parameters included in the time-dependent Cox model are also described in section 3.3. In section 3.3, the time-dependent Cox model analysis is performed on the IBCSG dataset with no imputation for illustrative purposes.

The selection of standard imputation methods that are applied is described in section 2.4. This chapter describes applying standard simple imputation methods to the IBCSF dataset. Applying multiple imputation methods is described in Chapter 4. The simple imputation methods that will be applied are:

i) extreme imputation
ii) last observation carried forward
iii) mean or median imputation: median by patient, by time period and by time period and treatment group
iv) imputation using regression models: linear regression with previous coping score(s) and with concurrent variables

The technical details of applying the simple imputation methods are described in section 3.4. The results from the time-dependent Cox model analysis following standard simple imputation methods are presented in section 3.5. The standard simple imputation methods are compared in two main ways:

i) comparing the estimated effect of the coefficient associated with quality of life from the time-dependent Cox model
ii) comparing the imputed quality of life value with the observed value artificially removed in a simulated dataset based on the patients with complete quality of life assessments in the IBCSG dataset

The relationship between the imputed values is explored together with multiple imputation in the next chapter, section 4.4. At the end of this chapter, a summary of the chapter is presented in section 3.6.

3.2 Description of IBCSG Trials VI and VII

The description of IBCSG Trials VI and VII in this section begins with the background (section 3.2.1) and patients and methods (section 3.2.2). The published efficacy analysis is summarised in section 3.2.3. Then quality of life and follow-up assessments are described in section 3.2.4. The published analysis of quality of life is summarised in section 3.2.5. The author did not contribute to these published analyses. Lastly, the further analyses of quality of life described in this chapter and Chapter 4 is outlined in section 3.2.6.

3.2.1 Background to IBCSG Trials VI and VII

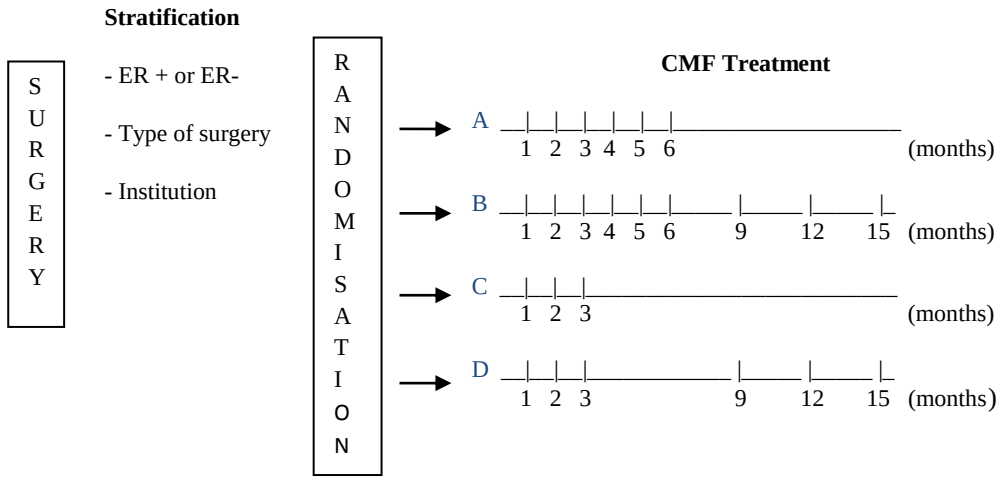
Adjuvant treatments with chemotherapy, endocrine therapy and combinations of both have been shown to increase DFS and OS in patients with node-positive breast cancer. However, adjuvant treatments can have substantial adverse effects, such as nausea which may impact on quality of life (Hürny et al. 1996) and increased risk of cardiac failure which may impact on survival (Piccart-Gebhart et al. 2005). An important clinical question is whether the DFS and OS benefits outweigh the adverse effects of adjuvant treatment on quality of life, though this is not the focus of this thesis.

At the time Trials VI and VII were started in 1986, adjuvant chemotherapy for breast cancer was generally administered for 6 to 12 months after surgery. Several trials had previously directly investigated the duration of chemotherapy. Since cancer cells are more likely to be sensitive to chemotherapy during phases of fast growth, there was a hypothesis that the timing of chemotherapy influences treatment efficacy and this was investigated in these trials.

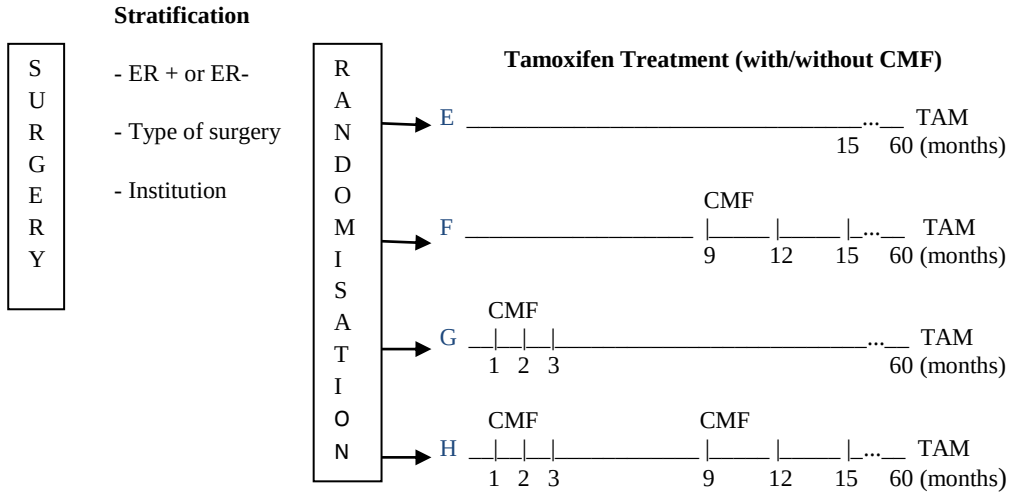
Trials VI and VII both examined two different durations and two timings of adjuvant chemotherapy in premenopausal and perimenopausal patients (VI) and compared tamoxifen with different durations and timing of chemotherapy in postmenopausal patients (VII). Therefore, each trial had 4 treatment groups, labelled A-D in Trial VI and E-H in Trial VII. The adjuvant chemotherapy was

cyclophosphamide, methotrexate and 5-fluorouracil (CMF). A secondary aim was to assess the effects of adjuvant treatments on self-assessed quality of life which was obtained prospectively on 9 occasions during the first 2 years on study. In the main findings from statistical analysis of quality of life in the IBCSG dataset (Hürny et al. 1996), Trials VI and VII were analysed together as they have similar protocols. A schema for Trials VI and VII, after protocol amendments, is shown in Figure 3.1.

TRIAL VI: PRE- AND PERIMENOPAUSAL



TRIAL VII: POSTMENOPAUSAL



CMF = cyclophosphamide, methotrexate, 5-fluorouracil; TAM = tamoxifen;
ER = oestrogen receptor

Figure 3.1 Schema of IBCSG Trials VI and VII

3.2.2 Patients and Methods

Premenopausal and perimenopausal patients with node-positive operable breast cancer were randomised in a 2x2 factorial design in Trial VI to receive three or six initial cycles of oral CMF, with or without delayed chemotherapy reintroducing three single cycles of CMF after 3-month intervals (labelled arms A-D in Figure 3.1). Postmenopausal patients with node-positive operable breast cancer were randomised in Trial VII to tamoxifen (20mg daily) for 5 years, alone or with chemotherapy (labelled arms E-H in Figure 3.1). The chemotherapy consisted of three early cycles of CMF, three delayed cycles of CMF, spaced as in Trial VI, or both early and delayed CMF. Randomisation in Trials VI and VII was stratified by participating institution, type of surgery (mastectomy vs breast conserving procedure with breast irraditation) and oestrogen receptor (ER) status (negative vs positive). The randomisation schedule was produced using pseudorandom numbers generated by a convergence method and was conducted centrally.

The primary endpoint was DFS, defined as the length of time from the date of randomisation to any relapse (including ipsilateral breast cancer), the appearance of a second primary cancer (including contralateral breast cancer), or death, whichever occurred first. Date of relapse was defined as the time when recurrent disease was diagnosed or, if confirmed later, when it was first suspected. Secondary endpoints included OS, defined as the time from randomisation to death due to any cause.

All patients had a histologically proven unilateral breast cancer with either ER-positive or negative primary tumours. Surgery of the primary tumour was either total mastectomy with axillary clearance or a lesser procedure (quadrantectomy or lumpectomy) with axillary lymph node dissection. For women treated with breast-conserving surgery, radiotherapy was mandatory and had to be postponed until the end of the initial phase of chemotherapy (three or six courses).

The protocol required that adjuvant chemotherapy began within 6 weeks of surgery and consisted of CMF (cyclophosphamide 100mg/m^2 orally on days 1 to 14, methotrexate 40mg/m^2 intravenously [IV] on days 1 and 8 and flurouracil 600mg/m^2 IV on days 1 and 8, repeated every 28 days). Eligible patients must have no evidence of metastatic spread and have acceptable baseline hepatic and renal function.

In the sample size calculations, the baseline 5-year DFS rate for patients entering Trial VI and receiving 6 months of CMF chemotherapy was assumed to be approximately 60%. Randomisation of 1400 patients to Trial VI was planned. The baseline 5-year DFS rate for patients entering Trial VII and receiving at least 2 years of tamoxifen was assumed to be approximately 50%. Randomisation of 1200 patients to Trial VII was planned.

Twenty-four institutions from nine countries took part in Trials VI and VII. Between July 1986 and April 1993, 1554 premenopausal and perimenopausal patients were randomised to Trial VI and out of the 1554 patients 1475 (95%) were eligible and assessable. During the same time period 1266 postmenopausal patients were randomised to Trial VII and out of the 1266 patients 1212 (96%) were eligible and assessable.

3.2.3 Published Statistical Analysis of Efficacy

The section summaries the main statistical analysis of efficacy in Trials VI and VII (The International Breast Cancer Group 1996; The International Breast Cancer Group 1997). DFS and OS percentages were estimated with the Kaplan-Meier method (Kaplan and Meier 1958). Greenwood's formula for the calculation of standard error (Greenwood 1926) and log-rank tests for the comparison of treatment effects were used. Cox proportional hazard regression models were used to control for prognostic factors (type of surgery, number of lymph nodes involved and ER status). Cox proportional hazard regression models were also used to estimate hazard ratios and confidence intervals for the treatment

comparisons and to test for interactions between potential prognostic factors and treatment effects.

With a median follow-up of 60 months, Trial VI showed that three courses of adjuvant chemotherapy are not sufficient compared with longer duration of CMF chemotherapy, especially in younger women and in patients with ER-negative tumours. Patients who received 3 cycles of CMF without reintroduction of chemotherapy had a 5-year DFS rate of 53% compared with 58% for the other three treatment groups (hazard ratio 1.20; 95% confidence interval [CI], 1.00 to 1.45). Delayed chemotherapy showed some evidence of additional therapeutic benefit (hazard ratio 0.86%; 95% CI 0.73 to 1.01) but remained investigational. With a median follow-up of 60 months, Trial VII found that early chemotherapy with three courses of CMF added to tamoxifen given for 5 years is beneficial for postmenopausal patients when compared with tamoxifen alone given for the same duration. The 598 patients who received early chemotherapy had a 5-year DFS rate of 64% compared with 57% for the 614 patients without early chemotherapy (hazard ratio 0.79; 95% CI 0.66 to 0.95). The use of tamoxifen before CMF chemotherapy might be detrimental, especially for patients with ER- primary tumours.

3.2.4 Quality of Life and Follow-Up Assessments

A secondary aim of IBCSG Trials VI and VII was to assess the effects of adjuvant treatments on self-assessed quality of life. The quality of life objectives were to evaluate the hypotheses:

i) The level of early coping/well-being of the patient at diagnosis can be used as a prognostic indictor of outcome
ii) The coping/well-being of the patient is different for different treatment groups

The quality of life questionnaire was designed to allow use in busy clinics and included five indicators of health-related quality of life that are particularly relevant in breast cancer patients. Physical wellbeing (good-lousy), mood (happy-miserable), appetite (good-none) and perceived adjustment/coping ("How much effort does it cost you to cope with your illness?" [none – a great deal]; Pacis) were assessed with single-item linear analogue self-assessment scales (100mm), which had previously been validated in several cancer populations (Coates et al. 1983; Coates et al. 1990). The fifth indicator is quality of life assessed with the Befindlichkeits-Skala (BfS) checklist for emotional wellbeing (von Zerssen 1986). The Bf-S scores have been compared with the mood scores (Hürny et al. 1996a). The quality of life forms were translated into the ten required languages.

The baseline quality of life was assessed on, or as close as possible to, the first day of adjuvant therapy. Quality of life was recorded approximately 3 months after randomisation, then every 3 months until 24 months, and also at 1 and 6 months after recurrence when applicable. The assessment of quality of life was intended to occur during a clinic visit when other clinical heamatologic and biochemical assessments would also take place. A clinical heamatologic and biochemical assessment of each patient was required every 3 months for 2 years, every 6 months from the third to the fifth year and yearly thereafter. After instruction from clinic staff, patients were asked to complete the quality of life questionnaire before receiving any scheduled chemotherapy.

3.2.5 Published Statistical Analysis of Quality of Life

Main Findings from Statistical Analysis of Quality of Life (Hürny et al. 1996)

The main findings from statistical analysis of quality of life in the IBCSG Trials VI and VII were published by Hürny et al. 1996 and are summarised in this section. In this statistical analysis, at a median follow-up of 60 months, the scores were treated as numbers from 0 to 100, and transformed so that higher numbers indicated a better quality of life.

ANOVA was used to test for associations between quality of life measures and biomedical and sociodemographic factors. Heterogeneity among treatment groups at each time of assessment was analysed by ANOVA, after controlling for language/culture. Tests for differences in quality of life scores between any two assessment times used within-patient changes in an ANOVA model that included assigned treatment and language/culture.

The baseline analysis and the tests for heterogeneity among treatment groups at each time point used the square root of the quality of life scores, because this transformation approximated a Normal distribution and was effective in stabilising the variances for quality of life scales.

Baseline prognostic factors were significantly associated with quality of life scores. In Trial VI, there was a poorer quality of life as the number of involved axillary nodes increased. In Trial VII, older patients (>60) reported better quality of life than younger postmenopausal patients. Patients who did not have all 9 quality of life assessments had systematically poorer quality of life than those for whom all 9 assessments were available.

Heterogeneity among the treatment groups was significant at each time point from 6 to 15 months in Trial VI and at each time point from 3 to 15 months in Trial VII. While the patient was still having chemotherapy the quality of life was poorer. By month 18, after all treatment groups had completed chemotherapy, there was no significant difference among treatment groups for any quality of life measures.

Between baseline and 18 months, there was a significant improvement of quality of life. There was a significant adverse impact of delayed chemotherapy on all quality of life measures.

Exploring Quality of Life as a Prognostic Factor of DFS

One of the quality of life objectives in Trial VI and VII was to evaluate the hypothesis that the level of early coping/well-being of the patient at diagnosis can be used as a prognostic indictor of outcome. This objective has been addressed by Coates et al. (2000) and Herring et al. (2004). As with Hürny et al. 1996, the transformation of square root of the quality of life scores was used and higher scores indicated a better quality of life.

In the analysis by Coates et al. (2000), Cox model analyses were used to test the relationship between quality of life scores and DFS. Coping score at baseline, Month 3 and Month 18 was considered. Patients with the relevant coping scores missing were not considered in the analyses. All models were stratified by language/country group and included other factors related to quality of life and/or outcome. The analysis indicated that DFS was not significantly predicted by quality of life scores or by changes in quality of life scores from baseline.

The goal of the analysis by Herring et al. (2004) of Swiss postmenopausal patients in Trial VII was to evaluate the effect of quality of life on outcome while taking into consideration that the data may be informative missing data. Here, the quality of life assessment considered was coping/perceived adjustment ("coping score"). They proposed a method for estimating parameters in the Cox proportional hazards model when missing covariates may be non-ignorable. The square root of the baseline coping scores was among the covariates considered, along with other prognostic factors. The analysis indicated that poor baseline coping scores were associated with improved relapse-free survival. Unlike the endpoint of DFS, occurrence of second primary cancer or death without prior event was not an event for relapse-free survival.

In the rest of this chapter and in Chapter 4, previous work is extended by considering quality of life as a time-dependent effect. As with Herring et al. (2004), coping score is the quality of life assessment considered. Of note, the

original coping score is considered and thus lower numbers indicated a better quality of life. Standard imputation methods are applied to impute the missing coping scores in the IBCSG dataset. The square root of the coping score (S_Pacis) is used in a time-dependent Cox model for DFS (see section 1.5) and therefore coping score is an example of a missing explanatory variable.

3.2.6 Further Analyses of Quality of Life as a Prognostic Factor of Disease-Free Survival

In the rest of this chapter and in Chapter 4, the IBCSG dataset is used to investigate the influence of missing quality of life values, as assessed by coping score, when exploring the relationship between quality of life and DFS. The 9 coping scores up to Month 24 are considered, which focus on the impact of adjuvant treatment in the early stage of the trial. Previous work is extended by considering quality of life as a time-dependent effect in a time-dependent Cox model. Standard imputation methods are applied to impute the missing coping scores before analysis and the performance of the standard imputation methods compared. In this section, preliminary work and the further analyses are described.

Preliminary Work

Quality of life assessments were repeated throughout IBCSG Trials VI and VII. Preliminary to the further analysis of quality of life as a prognostic factor of DFS, the status of coping scores over time was summarised in Table 3.1 and Figure 3.2.

As shown in Figure 3.2A and Table 3.1, considering only non-compliance where the quality of life assessment was expected, there was a high proportion of missing data in the IBCSG dataset, 17.0% [456/2687] at baseline (Time 1, approximately at randomisation) and 30.7% [666/2168] at 24 months (Time 9). The trend in the proportion of missing quality of life assessments was that initially the proportion was higher in Trial VII than in Trial VI, then was nearly identical

at 6 months (Time 3) and then was higher in Trial VI than Trial VII (Figure 3.2B). This high proportion indicates that the time-dependent Cox model analysis using only observed data will suffer from a lack of precision compared to the time-dependent Cox model analysis if all quality of life assessments were available.

Table 3.1 Summary of Status of Coping Score in IBCSG Trials VI and VII
Baseline (Time 1) – 24 Months (Time 9)

	Observed	Missing	Post-recurrence	Lost to follow-up	Dead
Baseline (Time 1)	2231 (83.0)	456 (17.0)	0 (0.0)	0 (0.0)	0 (0.0)
Month 3 (Time 2)	1918 (71.4)	744 (27.7)	25 (0.9)	0 (0.0)	0 (0.0)
Month 6 (Time 3)	1871 (69.6)	751 (27.9)	56 (2.1)	1 (0.0)	8 (0.3)
Month 9 (Time 4)	1817 (67.6)	745 (27.7)	103 (3.8)	1 (0.0)	21 (0.8)
Month 12 (Time 5)	1812 (67.4)	662 (24.6)	173 (6.4)	3 (0.1)	37 (1.4)
Month 15 (Time 6)	1692 (63.0)	711 (26.5)	215 (8.0)	3 (0.1)	66 (2.5)
Month 18 (Time 7)	1616 (60.1)	707 (26.3)	266 (9.9)	5 (0.2)	93 (3.5)
Month 21 (Time 8)	1556 (57.9)	693 (25.8)	294 (10.9)	5 (0.2)	139 (5.2)
Month 24 (Time 9)	1502 (55.9)	666 (24.8)	339 (12.6)	5 (0.2)	175 (6.5)

Data are n (%)

Figure 3.2 Bar Graph of Status of Coping Score in IBCSG Trials VI and VII
Baseline (Time 1) – 24 Months (Time 9)
A: Status of Coping Score By Time Period

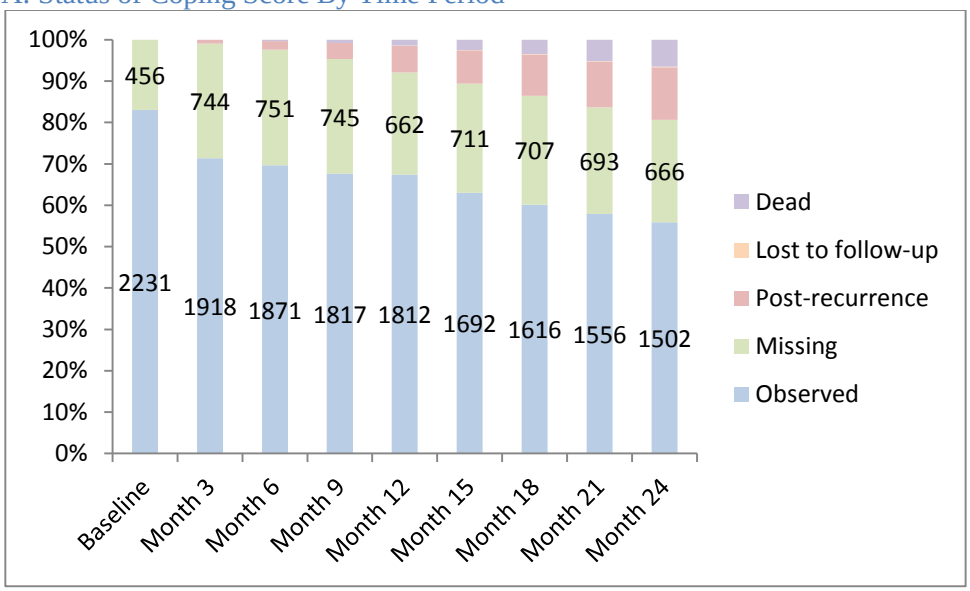
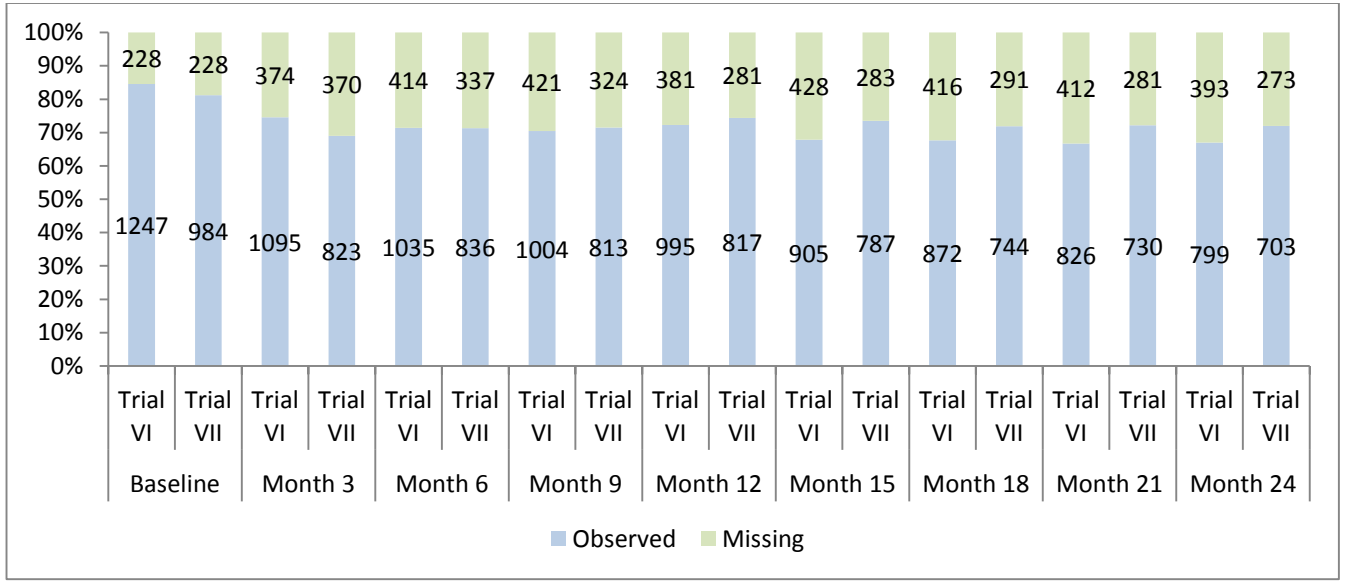


Figure 3.2 Bar Graph of Status of Coping Score in IBCSG Trials VI and VII
Baseline (Time 1) – 24 Months (Time 9)
B: Observed and Missing Coping Scores by Time Period and Trial



Next, the assumption that the coping scores are missing at random was considered. Statistical analysis of quality of life by Hürny et al. (1996) suggested that the patient's quality of life described by the coping score may be related to anticipation of future chemotherapy. In Trial VI, patients with early chemotherapy scheduled for 3 months, particularly the patients who were not scheduled to have delayed chemotherapy, had a better quality of life at 3 months (Time 2) than the patients with early chemotherapy scheduled for 6 months. In Trial VII, patients who were randomised to tamoxifen only reported a greater improvement in quality of life at 3 months (Time 2) compared to groups of patients who were randomised to receive delayed chemotherapy. This implies that the missing coping scores are likely to be informative missing data (see section 1.6.1). In addition, Hürny et al. (1996) found that patients who did not have all 9 quality of life assessments had systematically poorer quality of life than those for whom all 9 assessments were available. This fact also suggests informative missing data.

The high proportion of missing coping scores and the indication that the coping scores are informative missing data together lead to the conclusion that imputation is appropriate for the IBCSG dataset. Standard imputation methods were used, in order to include S_Pacis as a covariate in a time-dependent Cox model for DFS (see section 3.4 and 3.5). The standard imputation methods are used as part of an investigation into the possible relationship between quality of life and DFS and do not necessarily represent good estimation techniques for the IBCSG dataset.

Comparison of Standard Imputation Methods

The simple imputation methods (see section 3.4 and 3.5) are compared by:

i) Investigation of the estimated effects of the coefficient associated with S_Pacis and its standard error in the time-dependent Cox model fitted to the completed datasets following imputation by standard imputation methods. The estimates are compared to each other and to the estimated effects from the complete case analysis to consider the impact of the imputation method.

ii) From 585 patients with a complete history of 9 observed coping scores, 150 simulated datasets were created by removing coping scores to imitate the missing data pattern in the full IBCSG dataset. The missing data were imputed in each of the 150 simulated datasets, using the same methods as in i) above, and the mean and standard deviation of the difference between the imputed coping score and the real coping score, which had been artificially removed, were calculated. This was to investigate how well the imputed values reflected the real coping score. Imputation methods with a mean closer to zero and a small standard deviation are better predictors of the missing observations.

Description of Time-Dependent Cox Model

When investigating the relationship between quality of life and DFS, the changing values of quality of life, as measured by coping score throughout the study were considered. This means that a time-dependent Cox model was appropriate. Low values of coping scores, and therefore S_Pacis, indicated good quality of life. The treatment effect included in the time-dependent Cox model as well as S_Pacis was the indicator for delayed chemotherapy. The indicator was defined according to randomised arm (intent-to-treat). This treatment effect was used as statistical analysis of quality of life by Hürny et al. (1996) suggested that quality of life as described by coping score may be related to anticipation of future chemotherapy. The time-dependent Cox model analysis was stratified by trial, Trial VI (premenopausal) or Trial VII (postmenopausal).

Age was not considered as a covariate due to the association with menopausal status. Statistical analysis of efficacy (The International Breast Cancer Group 1996; The International Breast Cancer Group 1997) (see section 3.2.3) and previous statistical analysis of quality of life by Herring et al. (2004) suggested also including an indicator for sufficient early chemotherapy and oestrogen positive receptor status in the time-dependent Cox model. Sufficient early chemotherapy was defined as 6 initial cycles of CMF in Trial VI and 3 initial

cycles of CMF in Trial VII. This was not done in the main investigation of this chapter to keep the model parsimonious. However, the results from the extended model are also presented (see section 3.5).

Consider the notation for the time-dependent Cox model for the main investigation of this chapter. Let $X_{sp}(t)$ denote S_Pacis at time t and let

$$D_{sp(i)}(t) = \{D_{sp(i)}(u);\ 0 \leq u < t\} \tag{3.1}$$

denote the covariate history of S_Pacis up to time t for the i^{th} patient.

While S_Pacis changes during the study, the indicator for delayed chemotherapy, X_{del}, remains 1 for patients with delayed chemotherapy and 0 for patients with no delayed chemotherapy throughout.

Under the assumption of proportional hazards (see section 1.5):

$$\log\left(\frac{\lambda(t|\boldsymbol{Z}_i(t))}{\lambda_0}\right) = \beta_{sp}X_{sp(i)}(t) + \beta_{del}X_{del(i)} \tag{3.2}$$

where the derived covariates $\mathbf{Z}_i(t)$ for the i^{th} patient considered are S_Pacis at time t and the indicator for delayed chemotherapy

the derived covariates $\mathbf{Z}_i(t)$ are functions of the covariate history $D_{sp(i)}(t)$ and the time t

$\boldsymbol{\beta}$ is a vector of regression coefficients for S_Pacis and the indicator for delayed chemotherapy, written as β_{sp} and β_{del} respectively

λ_0 is an unspecified baseline hazard function

$\lambda(t|\mathbf{Z}_i(t))$ is the hazard function at time t for the i^{th} patient given the derived covariates $\mathbf{Z}_i(t)$

Time-Dependent Cox Model Analyses With No Imputation

For illustrative purposes, time-dependent Cox model analysis was performed on the IBCSG dataset with no imputation (see section 3.3). First, an initial analysis based upon 585 patients with observed coping scores at each of the 9 time periods was performed. Next, a dataset with a monotone missing data pattern created

from the available coping scores was considered. Thirdly, a dataset with all available coping scores was considered.

3.3 Available Patients, Complete Patient and Available Patients with a Monotone Missing Data Pattern Analysis

The hypothesis investigated in this chapter and Chapter 4 is that poorer quality of life throughout the study is associated with poorer DFS and conversely better quality of life throughout the study is associated with better DFS. Standard imputation methods are applied to impute the missing coping scores before analysis in a time-dependent Cox model. The influence of the standard simple imputation methods and the standard multiple imputation methods on the parameter estimates for S_Pacis and delayed chemotherapy is considered in Chapters 3 and 4 respectively. Time-dependent Cox model analyses without imputation was carried out to provide parameter estimates for reference and illustrative purposes. The analyses of coping scores for reference and illustrative purposes were: i) complete case ii) available monotone and iii) all available.

The complete case analysis considered the 585 patients with observed coping scores at each of the 9 time periods. Of note, a patient must have been alive and disease-free for 24 months to have 9 observed coping scores. The dataset for the available monotone coping score analysis was extracted from the dataset for all available coping scores. This was done by removing the relevant observed coping scores from the patients with a non-monotone missing data pattern. As many patients in the IBCSG dataset have a non-monotone missing data pattern, a large number of observed coping scores had to be removed. The time-dependent Cox model analysis of available monotone coping scores and all available coping scores considered 2214 and 2544 patients respectively.

Technical Details of Time-Dependent Cox Model Analysis Using Available Monotone and All Available Coping Scores

The following patients were not considered in the time-dependent Cox model analysis of available monotone coping scores:

i) Patients with a missing baseline coping score (Time 1, approximately at randomisation)

ii) One patient who had the date of quality of life assessment at 21 (Time 8) and 24 months (Time 9) both on 25th June 1992

iii) One patient who was considered disease-free in the efficacy analysis but has a date of recurrence reported at 16.6 months after randomisation

iv) Fifteen patients where using the expected dates of assessment for missing coping scores led to intervals of less or equal to 0 for the time-dependent Cox model were not considered.

There were 103 patients where the baseline coping score (approximately at randomisation) was assessed before the date of randomisation, ranging from 30 days to 1 day before randomisation.

The dataset for the analysis of all available coping scores contains patients with intermittent missing coping scores. Here, the intervals for quality of life began on the date of the first quality of life assessment and patients with a missing baseline coping score could be considered. The intervals ended at the next quality of life assessment, regardless of any missing intermittent coping scores. Again, the patient who had the date of quality of life assessment at 21 (Time 8) and 24 months (Time 9) on the same date was not considered. One other patient was excluded because the date of the quality of life assessments at 6 months (Time 3) of 21st June 1994 was after the date of the quality of life assessment at 15 months (Time 6) of 15th March 1994.

In order to carry out the analysis of coping scores with a missing data pattern, a large number of observed coping scores were removed. The status of coping

scores considered for available patient analysis using a monotone missing data pattern was as shown in Figure 3.3 and summarised in Table 3.2

Figure 3.3 Bar Graph of Status of Coping Scores by Time Period for Time-Dependent Cox Model Analysis of Available Patients with Monotone Missing Data Pattern

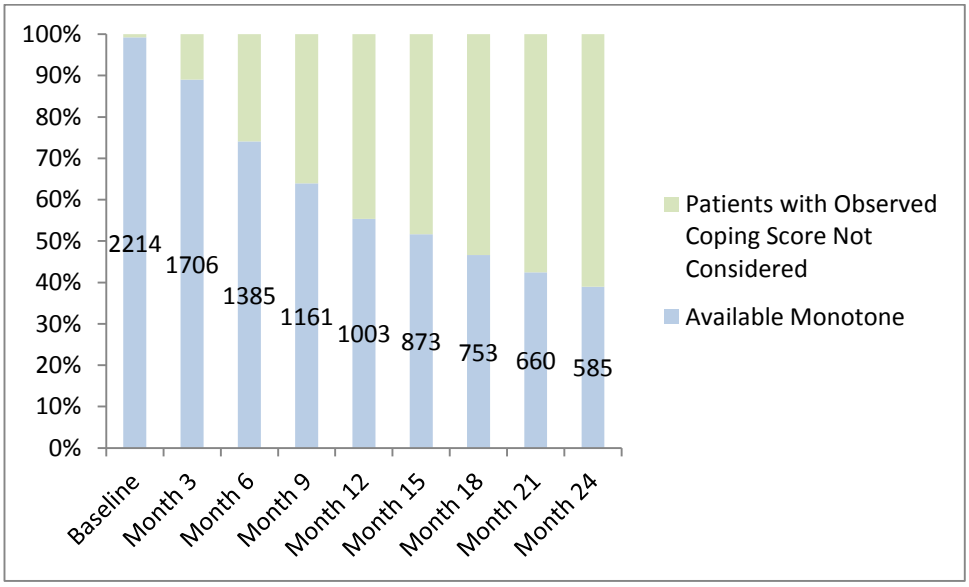


Table 3.2 Summary of Status of Coping Scores in Time-Dependent Cox Model Analysis Stratified by Trial: Available Patients with a Monotone Missing Data Pattern

	Available Monotone	Patients with Observed Coping Score Not Considered
Baseline (Time 1)	2214	16
Month 3 (Time 2)	1706	211
Month 6 (Time 3)	1385	484
Month 9 (Time 4)	1161	655
Month 12 (Time 5)	1003	808
Month 15 (Time 6)	873	817
Month 18 (Time 7)	753	862
Month 21 (Time 8)	660	895
Month 24 (Time 9)	585	916

Assumption of Proportional Hazards

The plots of Schoenfeld residuals (Schoenfeld 1982) against time for the S_Pacis and delayed chemotherapy from the time-dependent Cox model analysis of all available coping scores is shown in Figure 3.4.

Figure 3.4 A

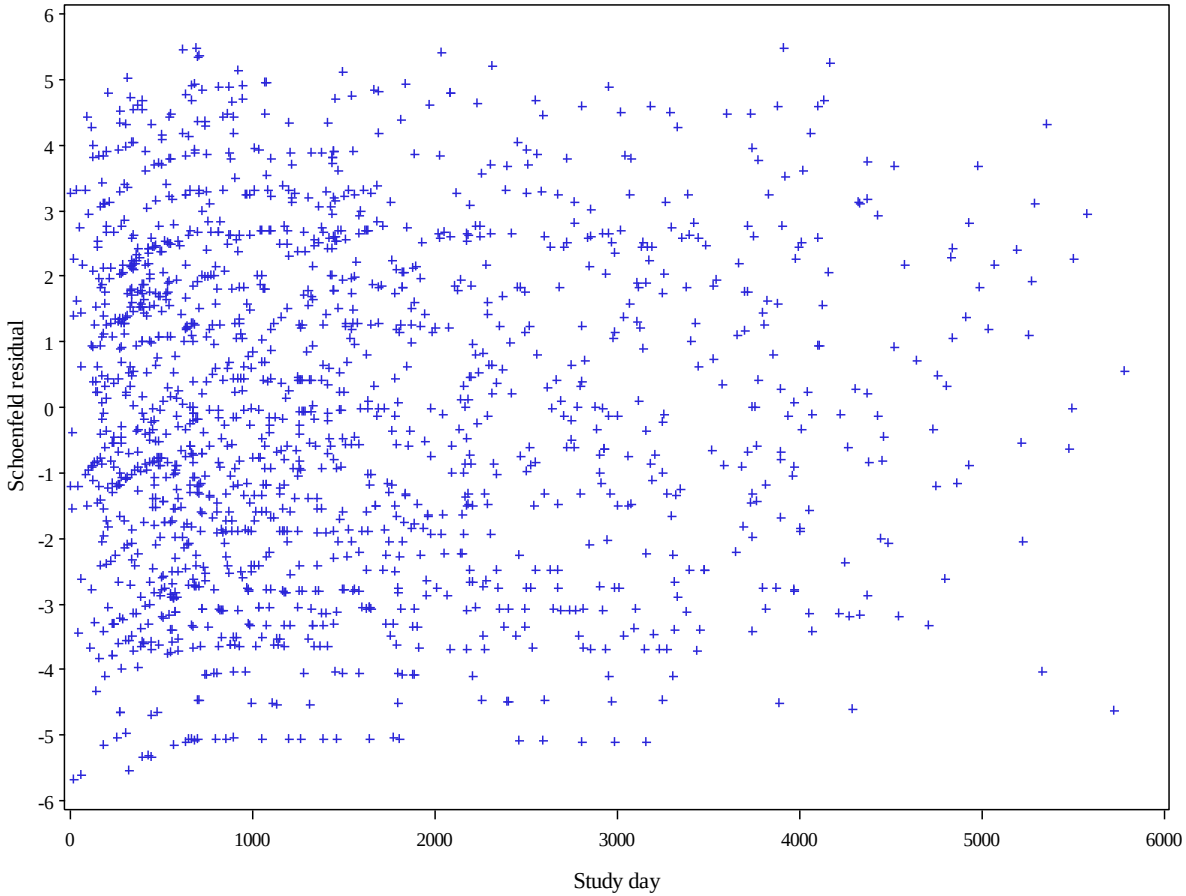

Schoenfeld residual
6
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
0
1000
2000
3000
4000
5000
6000
Study day

Figure 3.4 B

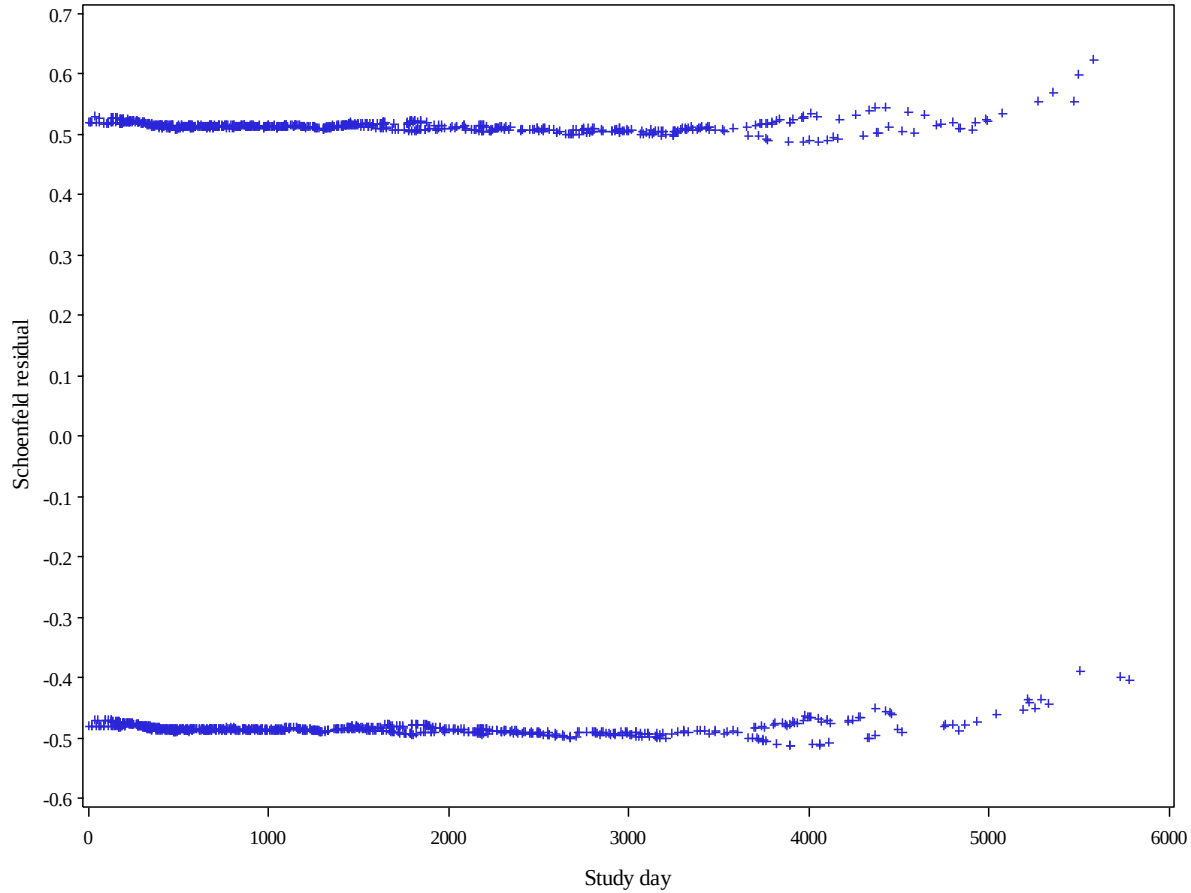


Figure 3.4 Schoenfeld residuals against time for the square root of the coping score (S_Pacis) (A) and delayed chemotherapy (B) from the time-dependent Cox model analysis of all available coping scores

A zero slope indicates that the assumption of proportional hazards is reasonable (Figure 3.4). Beyond approximately 11 years (~ 4000 days), the plots no longer indicated a zero slope for delayed chemotherapy (Figure 3.4B). However, this did not raise concerns time about the time-dependent Cox mode, as approximately 5% DFS events occurred after this time.

Results of the Time-Dependent Cox Model Analyses of Disease-Free Survival with no Imputation

The results of the time-dependent Cox model analyses with no imputation are shown below in Table 3.3. The bias from the selection of patients considered in these analyses makes it difficult to interpret the values of the parameter estimates from the time-dependent Cox model. The parameter estimates from the available monotone and all available analyses were similar and much more like each other than the parameter estimates from the complete case analysis. This difference between these parameter estimates illustrate the importance of carefully investigating the missing data mechanism and considering the assumptions of analysis methods in order to avoid biased parameter estimates. The standard error of the parameter estimates from the complete case analysis was approximately twice as large as the standard errors from the available monotone and all available analyses. There was little increase in the standard error of the parameter estimates from removing the relevant coping scores to create a monotone missing data pattern.

Table 3.3 Summary of Time-Dependent Cox Model Analysis Stratified by Trial: Complete Patients, All Available Patients and Available Patients with a Monotone Missing Data Pattern

Parameter	Dataset	Parameter estimate	Standard error	Hazard ratio	95% CI for hazard ratio
S_Pacis	Complete	-0.0200	0.0259	0.9802	(0.9317, 1.0311)
	Avail Monotone	0.0096	0.0110	1.0096	(0.9881, 1.0316)
	All Available	0.0085	0.0104	1.0086	(0.9881, 1.0294)
Delayed	Complete	-0.0415	0.1207	0.9594	(0.7572, 1.2155)
Chemotherapy	Avail Monotone	-0.0933	0.0555	0.9109	(0.8170, 1.0155)
	All Available	-0.1079	0.0519	0.8977	(0.8109, 0.9938)

S_Pacis = square root of the coping score; CI = confidence interval; Avail = available

3.4 Technical Details of Application of Simple Imputation Methods to the IBCSG Breast Cancer Trial Data

In this section, the technical details of applying the standard simple imputation methods to the IBCSG dataset and of the subsequent time-dependent Cox model analysis are described. In particular, addressing the general missing data pattern in the IBCSG dataset and the patients included in the dataset for the time-dependent Cox model analysis are described.

3.4.1 Introduction

Standard simple imputation methods were used to impute missing coping scores before analysis of DFS. The aim of the imputation was to use the S_Pacis as a covariate in a time-dependent Cox model for DFS. The time-dependent Cox model analysis investigates the hypothesis that poorer quality of life throughout the study is associated with poorer DFS and conversely better quality of life throughout the study is associated with better DFS. The standard simple imputation methods considered are listed in section 3.1. Extreme imputation is considered for illustration. The results from the time-dependent Cox model analysis after applying standard simple imputation methods to the IBCSG dataset are shown in section 3.5.

Missing Data Pattern

As noted (section 2.4), the IBCSG dataset had a general missing data pattern which was not close to monotonic. Among the 456 patients with a missing baseline coping score (Table 3.1), 312 (68%) had at least one observed coping score at a later time point. The intermittent observed coping scores among these patients did not have a consistent missing data pattern. Some of the standard simple imputation methods require a baseline coping score. In this scenario, there was no obvious way to replace the missing coping scores. It was more reasonable to exclude the patients with a missing baseline coping score from the time-dependent Cox model analysis.

Calculating Imputed Values for Coping Score

For the following standard imputation methods, S_Pacis was considered during the imputation. From these imputed values, the imputed value of the missing coping score was calculated, rounded to the nearest whole number.

i) linear regression with previous coping score(s) (section 3.4.5)

ii) linear regression with concurrent variables (section 3.4.6)

Quality of Life Assessments Considered in the Time-Dependent Cox Model Analysis Following Imputation Methods

There were 2687 patients randomised to Trials VI and VII. However, some patients could not be considered in the time-dependent Cox model analysis following standard imputation methods. These patients were:

i) one patient who had the date of quality of life assessment at 21 and 24 months (Time 8 and Time 9) both on 25th June 1992

ii) one patient who was considered disease-free in the efficacy analysis but has a date of recurrence reported at 16.6 months after randomisation

iii) patients where using the expected dates of assessment for missing coping scores led to intervals of less or equal to 1 day for the time-dependent Cox model

The 456 patients with a missing baseline coping score (approximately at randomisation) were not considered in the time-dependent Cox model analysis after applying the following standard imputation methods:

i) LOCF (section 3.4.3)

ii) linear regression with previous coping score(s) (section 3.4.5)

When considering these standard imputation methods, there were 15 patients where using the expected dates of assessment for missing coping scores together with actual dates of observed assessments led to intervals of less than 1. These 15 patients were not considered in the time-dependent Cox model analysis. Thus, for

the 6 standard imputation methods listed above, 2214 patients were considered in the time-dependent Cox model analysis.

The number of patients in the time-dependent Cox model analysis is described as part of the technical details for the remaining standard imputation methods:

i) extreme imputation (section 3.4.2)
ii) median imputation (section 3.4.4)
iii) linear regression with previous coping scores (section 3.4.6)

The status of the coping scores considered for time-dependent Cox model analysis at each time points considered was as shown in Table 3.4.

The methods which allow patients with a missing baseline coping score to be considered, such as median imputation, result in the largest number of patients and the largest number of DFS events being considered in the time-dependent Cox model analysis. Linear regression using concurrent variable results in the lowest number of patients and DFS events being considered postbaseline in the time-dependent Cox model (Table 3.4).

Table 3.4 Status of Coping Scores for Time-Dependent Cox Model Analysis

	Time	Main Group	Extreme	Median by Patient	Median by Time, Time and Trt	Lin Reg Concurr
Observed	1	2214	2214	2214	2214	2214
	2	1706	1899	1899	1899	1783
	3	1657	1855	1855	1855	1540
	4	1611	1807	1807	1807	1319
	5	1585	1795	1795	1795	1177
	6	1495	1676	1676	1676	1046
	7	1415	1604	1604	1604	910
	8	1380	1544	1544	1544	819
	9	1329	1488	1488	1488	728
Imputed	1*	0	446	308	446	157
	2	492	739	607	739	192
	3	502	741	612	741	96
	4	498	730	607	730	81
	5	451	654	537	654	72
	6	484	704	591	704	57
	7	499	698	587	699	52
	8	478	683	576	683	47
	9	470	663	560	663	39
Total	1	2214	2660	2522	2660	2371
	2	2198	2638	2506	2638	1975
	3	2159	2596	2467	2596	1636
	4	2109	2537	2414	2537	1400
	5	2036	2449	2332	2449	1249
	6	1979	2380	2267	2380	1103
	7	1914	2302	2190	2302	962
	8	1858	2227	2120	2227	866
	9	1799	2151	2048	2151	767

*see note on number of patients with an imputed baseline coping score (Time 1) in extreme imputation and median imputation by time period or by time period and treatment group in section 3.4.2 and section 3.4.4 respectively

Imputation methods in column Main group = i) last observation carried forward, ii) bootstrapping, subgroups defined by baseline coping score and subgroups defined by previous coping score, iii) linear regression with previous coping score(s), iv) nearest neighbour imputation, v) predictive mean matching, and vi) pattern mixture models – Curran's analytic technique; extreme = extreme imputation; median by patient = median imputation by patient; median by time, time and trt = median by time period/median by time period and treatment group; lin reg concur= linear regression with concurrent variables; imputed without LOCF= imputed coping scores excluding coping scores imputed by last observation carried forward in order to create a monotone missing data pattern for i) nearest neighbour imputation, ii) predictive mean matching and iii) pattern mixture models – Curran's analytic method

Table 3.4 Status of Coping Scores for Time-Dependent Cox Model Analysis (continued)

	Time	Main Group	Extreme	Median by Patient	Median by Time, Time and Trt	Lin Reg Concurr
Recurrence/	1	N/A	N/A	N/A	N/A	N/A
Death*	2	16	22	16	22	16
	3	38	41	38	41	28
	4	50	59	52	59	32
	5	72	86	80	86	42
	6	57	69	65	69	34
	7	63	76	74	76	32
	8	56	75	71	75	26
	9	59	76	72	76	29
Lost to	1	N/A	N/A	N/A	N/A	N/A
follow-up*	2	0	0	0	0	0
	3	1	1	1	1	1
	4	0	0	0	0	0
	5	1	2	2	2	0
	6	0	0	0	0	0
	7	2	2	2	2	1
	8	0	0	0	0	0
	9	0	0	0	0	0

*since last assessment

Imputation methods in column Main group = i) last observation carried forward, ii) bootstrapping, subgroups defined by baseline coping score and subgroups defined by previous coping score, iii) linear regression with previous coping score(s), iv) nearest neighbour imputation, v) predictive mean matching and vi) pattern mixture models – Curran's analytic technique; extreme = extreme imputation; median by patient = median imputation by patient; median by time, time and trt = median by time period/median by time period and treatment group; lin reg concur= linear regression with concurrent variables

Dates of Quality of Life Assessments Considered in Time-Dependent Cox Model Analysis Following the Main Group of Standard Imputation Methods

For 2 patients with a date of assessment taken from the non-compliance form but a missing coping score, the expected date of quality of life assessment was used to prevent an interval of less than 1 day in the time-dependent Cox model analysis. There were 103 patients where the baseline coping score (approximately at randomisation) was assessed before the date of randomisation, ranging from 30 days to 1 day before randomisation. Among the 1799 patients with a coping score considered at 24 months (Time 9), there were 4 patients where the date of quality of life assessment at 24 months (Time 9) was more than 2.5 years after randomisation. Two of these 4 patients had an observed coping score. One of

these 4 patients had a partial quality of life assessment at 24 months (Time 9). The remaining patient had a quality of life assessment scheduled 1244 days (41 months) after randomisation but did not complete the quality of life assessment and has a missing coping score.

Comparing Imputation Methods

To compare simple imputation methods, 585 patients with a complete history of 9 observed coping scores were identified and some values were removed to imitate the missing data pattern in the full dataset. Baseline coping scores (Time 1, approximately at randomisation) were not removed in order that all patients could be considered when applying standard imputation methods to the simulated dataset. One hundred and fifty simulated datasets with coping scores artificially removed were generated. For each of the 150 simulated datasets with coping scores artificially removed, the difference between the imputed coping score and the real coping score originally observed and artificially removed was calculated. To further investigate imputed coping scores following median imputation, the distribution of the coping scores artificially removed and the distribution of imputed coping scores was summarised.

3.4.2 Extreme Imputation

Missing coping scores were replaced by 100 (lowest quality of life) and then by 0 (highest quality of life). When considering extreme imputation, 22 patients where using the expected dates of assessment for missing coping scores led to intervals of less or equal to 0 for the time-dependent Cox model were not considered. Out of these 22 patients, 7 had a missing baseline coping score. In addition to the exclusions described in section 3.4.1, the 3 patients with a recurrence on the date of randomisation were not considered. These 3 patients had a missing baseline coping score. Thus, 2660 patients were considered in the time-dependent Cox model analysis (Table 3.4). This included 446 patients with a missing baseline coping score. (Table 3.4; Time 1 in section Imputed).

For 3 patients with a date of assessment taken from the non-compliance form but a missing coping score, the expected date of quality of life assessment is used to prevent an interval of less than 1 day for the time-dependent Cox model. There were 112 patients where the date of baseline coping score was before the date of randomisation. There were 4 patients where the date of quality of life assessment at 24 months (Time 9) was more than 2.5 years after randomisation.

3.4.3 Last Observation Carried Forward

Missing coping scores were replaced with the last observed coping score for the patient. There were 2214 patients considered in the time-dependent Cox model analysis (Table 3.4).

3.4.4 Median Imputation

Three types of median were considered. Median of the patient's observed coping scores, median of the time period and median of treatment group and time period. Where necessary, the median to be imputed was rounded to the nearest whole number.

Median Imputation by Patient

Missing coping scores were replaced by the median of the patient's observed coping scores, calculated by *proc univariate* in SAS. When considering median imputation by patient, 18 patients where using the expected dates of assessment for missing coping scores led to intervals of less or equal to 0 for the time-dependent Cox model were not considered. In addition to the exclusions described in section 3.4.1, the 141 patients with no observed coping scores were not considered. Thus, 2522 patients were considered in the time-dependent Cox model analysis (Table 3.4).

For 3 patients with a date of assessment taken from the non-compliance form but a missing coping score, the expected date of quality of life assessment is used to

prevent an interval of less than 1 day for the time-dependent Cox model. There were 109 patients where the date of baseline coping score was before the date of randomisation. There were 4 patients where the date of quality of life assessment at 24 months (Time 9) was more than 2.5 years after randomisation.

Median Imputation by Time Period and Median Imputation by Time Period and Treatment Group

Missing coping scores were replaced by the median of the observed coping scores by time period or the median of the observed coping scores by time period and treatment group. These medians were calculated by *proc univariate* in SAS. As with extreme imputation (section 3.4.2), 2660 patients were considered in the time-dependent Cox model analysis (Table 3.4) and this included 446 patients with a missing baseline coping score (Table 3.4; Time 1 in section Imputed).

3.4.5 Linear Regression with Previous Coping Score(s)

S_Pacis was modelled with a linear regression model using the square root of all previous observed or imputed coping score(s) as explanatory variables. The linear regression model was calculated by *proc reg* in SAS. There were 2214 patients considered in the time-dependent Cox model analysis (Table 3.4). The R^2 values from the linear regression model with previous coping score(s) were as in Table 3.5. There was a trend for the R^2 values to increase as the number of explanatory in the model increases up until S_Pacis at Time 6 is modeled.

Table 3.5 R^2 Value from Linear Regression Model for Square Root of Coping Score (S_Pacis) Based on the Square Root of Previous Coping Score(s)

Time	2	3	4	5	6	7	8	9
R^2 value	0.274	0.406	0.475	0.511	0.565	0.558	0.631	0.583

3.4.6 Linear Regression with Concurrent Variables

S_Pacis was modeled with a linear regression model using the concurrent categorical variables assessed at approximately the same time as the quality of life assessment. These concurrent variables refer to adverse events, performance status and menstrual status.

Concurrent variables were considered from the Chemotherapy form (Form D) which records the highest intensity experienced by the patient of the following adverse events:

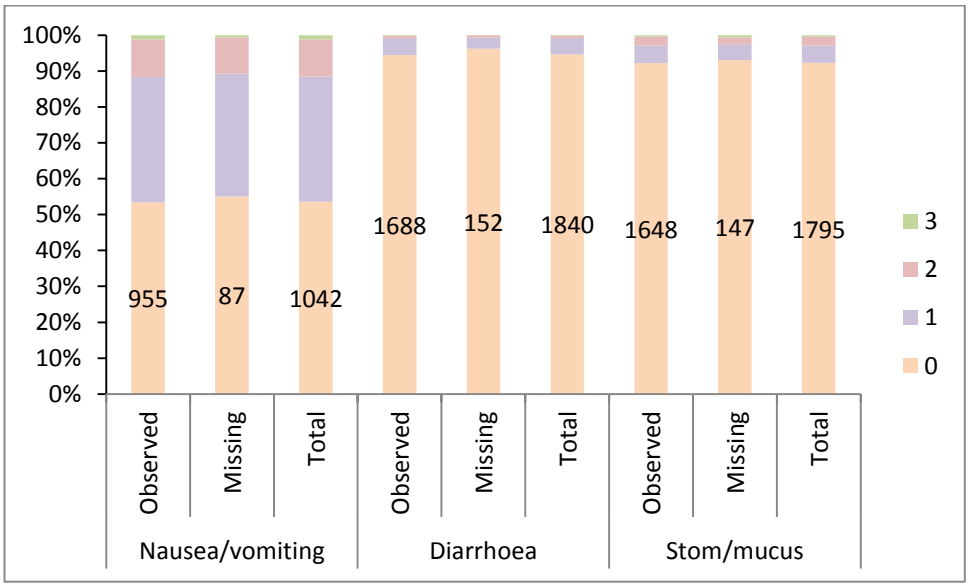
Nausea / vomiting
Diarrhoea
Stomatis / mucus membrane

The categories for the intensity for adverse events, with the reference category in bold, are:

0	**patient did not experience the adverse event**
1	mild
2	moderate
3	severe
4	life-threatening

The intensity of adverse events reported at baseline were as shown in Figure 3.5:

Figure 3.5 Intensity of Adverse Events at Baseline by Status of Coping Score



Stom / mucus = stomatis / mucus membrane

Concurrent variables were considered from the Follow Up form (Form E) which records the performance status and menstrual status. There were 6 categories for performance status and 3 categories for the menstrual status.

The categories for performance status, with the reference category in bold, are:

0 **fully active without restriction or aid of analgesics**

1 restricted in strenuous activity, but ambulatory and able to carry out work of light and sedentary nature

2 ambulatory and capable of self care, but unable to work. Up and about more than 50% of waking hours

3 capable of only limited self care, confined to bed or chair more than 50% of waking hours

4 completely disabled. Unable to carry out any self care and totally confined to bed or chair

5 dead

The categories for menstrual status, with the reference category in bold, are:

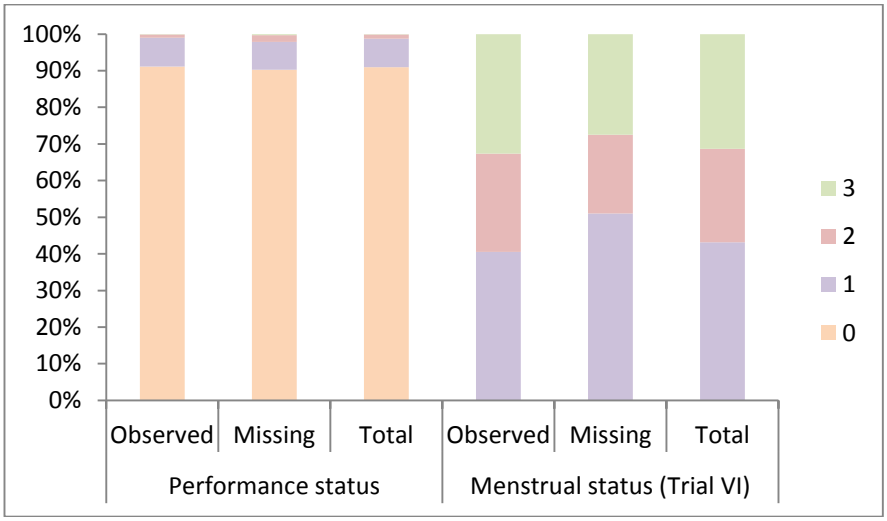
1 normal and regular during the last 3 months

2 irregular and/or scanty

3 no period

No quality of life assessment should exist after the patient's death. Thus, no patient with a quality of life assessment has a performance status of 5 Dead. The menstrual status for patients in Trial VII, who are postmenopausal, was 3 no period, with one exception. The performance status and menstrual status reported at Week 13 was as shown below in Figure 3.6:

Figure 3.6 Performance Status and Menstrual Status at Week 13 by Status of Coping Score



The schedule for completing Chemotherapy forms was not the same in all treatment groups. It was not expected that a Chemotherapy form could be matched to a quality of life assessment in all treatment groups at all time periods. Only Chemotherapy forms with information for all 3 types of adverse event were considered when matching the forms to the quality of life assessments. There were 12 Chemotherapy forms which could not be considered because of an incomplete or invalid date. The quality of life assessments were matched to

information on the Chemotherapy form with the closest date up to 30 days before the date or expected date of the quality of life assessment. It was not possible to match a Chemotherapy form to every quality of life assessment where this was expected.

When matching Follow Up forms to the quality of life assessments, only forms with information for both performance status and menstrual status were considered. Follow Up forms were not expected at baseline. Quality of life assessments after baseline (3 months – 24 months, Time 2 – Time 9) were matched to information on the Follow Up form with the closest date up to 30 days before the date or expected date of the quality of life assessment. It was not possible to match a Follow Up form to every quality of life assessment where this was expected.

The concurrent information expected to be matched to the quality of life assessments is summarised below in Table 3.6.

Table 3.6 Concurrent Information Matched to Quality of Life Assessments for Linear Regression Based on Concurrent Variables

	Trial VI (Premenopausal)				Trial VII (Postmenopausal) All Treatment Groups
	A	B	C	D	
Baseline (Time 1)	Chemo	Chemo	Chemo	Chemo	Chemo
Month 3 (Time 2)	Chemo, FU	Chemo, FU	Chemo, FU	Chemo, FU	Chemo, FU
Month 6 (Time 3)	Chemo, FU	Chemo, FU	FU	Chemo, FU	Chemo, FU
Month 9 (Time 4)	FU	Chemo, FU	FU	Chemo, FU	Chemo, FU
Month 12 (Time 5)	FU	Chemo, FU	FU	Chemo, FU	Chemo, FU
Month 15 (Time 6)	FU	Chemo, FU	FU	FU	Chemo, FU
Month 18 (Time 7)	FU	FU	FU	FU	Chemo, FU
Month 21 (Time 8)	FU	FU	FU	FU	Chemo, FU
Month 24 (Time 9)	FU	FU	FU	FU	Chemo, FU

Chemo = Chemotherapy form (Form D); FU = Follow Up form (Form E)

The linear regression models for missing coping scores at Time 2 to Time 9 were calculated by *proc reg* in SAS. Both a Chemotherapy form and Follow Up form must be matched to a quality of life assessment where indicated for the quality of

life assessment to be considered. The parameters included in the linear regression models reflect the concurrent information expected to be matched to the quality of life assessment. When considering only the patients in Trial VII (postmenopausal patients), menopausal status was not included in the linear regression model. As higher intensities of adverse events and higher categories of performance status were unusual, all potential categories were not always included in the linear regression model. The parameters in the respective linear regression models are shown in Table 3.7.

Table 3.7 Parameters in Linear Regression Model for Square Root of Coping Score (S_Pacis) Based on Concurrent Variables

	Treatment Group	Adverse Event Grade from Chemotherapy From			Status Category from Follow Up Form	
		Nausea/ vomitting	Diarrhoea	Stom / mucus	Performance Status	Menstrual Status
Baseline (Time 1)	All	1, 2, 3	1, 2, 3	1, 2, 3		
Month 3 (Time 2)	All	1, 2, 3	1, 2, 3	1, 2, 3	1, 2	2, 3
Month 6 (Time 3)	A, B, D, E, F, G, H	1, 2, 3	1, 2	1, 2	1, 2	2, 3
Month 6 (Time 3)	C				1, 2, 3	2, 3
Month 9 (Time 4)	B, D, E, F, G, H	1, 2, 3	1, 2	1, 2	1, 2	2, 3
Month 9 (Time 4)	A, C				1, 2	2, 3
Month 12 (Time 5)	B, D, E, F, G, H	1, 2, 3	1, 2, 3,	1, 2	1, 2, 3	2, 3
Month 12 (Time 5)	A, C				1, 3	2, 3
Month 15 (Time 6)	B, E, F, G, H	1, 2, 3	1, 2	1, 2	1, 2, 3	2, 3
Month 15 (Time 6)	A, C, D				1	2, 3
Month 18 (Time 7)	E, F, G, H	1, 2			1, 2, 3	2, 3
Month 18 (Time 7)	A, B, C, D				1, 3	2, 3
Month 21 (Time 8)	E, F, G, H	1		1	1	
Month 21 (Time 8)	A, B, C, D				1	2, 3
Month 24 (Time 9)	E, F, G, H	1		1	1, 3	
Month 24 (Time 9)	A, B, C, D				1, 2, 3	2, 3

Stom / mucus = stomatis / mucus membrane

Category of adverse events:
0 = patient did not experience the adverse event; 1 = mild; 2 = moderate; 3 = severe

Category of performance status:

0	**fully active without restriction or aid of analgesics**
1	restricted in strenuous activity, but ambulatory and able to carry out work of light and sedentary nature
2	ambulatory and capable of self care, but unable to work. Up and about more than 50% of waking hours
3	capable of only limited self care, confined to bed or chair more than 50% of waking hours

Category of menstrual status:

1	**normal and regular during the last 3 months**
2	irregular and/or scanty
3	no period

Reference categories shown in bold
Treatment groups A-D: Trial VI (premenopausal); treatment groups E-H: Trial VII (postmenopausal)

3.4.7 Reason for 150 Simulated Datasets to Estimate Difference Between the Imputed Coping Score and the Missing Coping Score for Simple Imputation Methods

Hauser and Walsh (2008) noted that small changes (less than 5mm) on a VAS may not be clinically relevant. A clinically significant change has been suggested as 50% of the scale's standard deviation, which is 8-10mm on a 100mm VAS such as PACIS. Sloan and Dueck (2004) consider detecting a treatment effect between two treatment groups. They note that around 400 and 55 patients in each treatment group would be needed to achieve an 80% power to detect a small (3mm) or medium (8mm) effect size respectively. This is based on a two-sample *t*-test with a 5% type I error rate (see section 1.1.4). As noted in section 3.2.6, the estimated mean and standard deviation of the difference between the imputed coping score and missing coping score in the full IBCSG dataset was calculated from 150 simulated datasets. In each of the 150 simulated datasets, there are approximately 1000 artificially missing coping scores among approximately 500 patients. Given this large number, it is reasonable to assume, that even a small difference (~ 3 out of a range 0-100) that is not clinically meaningful between the imputed coping score and the real coping score artificially removed would be found in each of the simulated completed datasets.

Each simulated completed dataset following simple imputation is a potential representation of the completed dataset if the missing data pattern from the IBCSG dataset had applied to the patients with a complete history of 9 observed coping scores. Considering LOCF, the range of the estimated mean difference between the imputed coping score and the missing coping score from each individual dataset was constant after 100 simulated datasets (Appendix B, Figure B1.1). This was approximately similar for all standard simple imputation methods (Table B1.1). These considerations on power and completed datasets indicate that there would be no benefit from considering more than 150 simulated datasets

when estimating the mean and standard deviation of the difference between the imputed coping score and the missing coping score in the IBCSG dataset.

3.5 Results from Applying Simple Imputation Methods to the IBCSG Dataset

In this section, the contents of summary tables of results from applying simple imputation methods to the IBCSG dataset are described (section 3.5.1) and then results are summarised (section 3.5.2).

3.5.1 Description of Contents of Tables of Results from Applying Standard Imputation Methods to the IBCSG Dataset

The results from the time-dependent Cox model analysis following standard simple imputation methods are shown in Table 3.8. The estimated mean difference between the imputed coping score and the missing coping score is the estimate of the real value of the missing coping score – the imputed coping score. The standard error shown is the square root of the variance of the parameter estimate. The standard error and the parameter estimate are used to calculate the 95% confidence interval for the parameter estimate. The exponential of the lower and upper 95% confidence limits for the parameter estimate gives the lower and upper 95% confidence limits for the hazard ratio. Results from the extended time-dependent Cox model are also presented in Table 3.9 are calculated similarly.

Table 3.8 Summary of Time-Dependent Cox Model Analysis Considering Square Root of Coping Score (S_Pacis) and Delayed Chemotherapy Stratified by Trial

Square root of coping score (S_Pacis)					
Method	Detail	Parameter estimate	Standard error	*t* statistic	95% CI for hazard ratio
Extreme	100	0.0239	0.0080	2.99	(1.008, 1.040)
	0	-0.0148	0.0087	-1.70	(0.969, 1.002)
LOCF		0.0047	0.0112	0.42	(0.983, 1.027)
Median	by patient	0.0136	0.0107	1.27	(0.993, 1.035)
	time period	0.0233	0.0117	1.99	(1.000, 1.047)
	time period and trt group	0.0235	0.0117	2.01	(1.000, 1.048)
Linear regression	previous coping scores	0.0069	0.0124	0.56	(0.983, 1.032)
	concurrent variables	0.0112	0.0112	1.00	(0.989, 1.034)
Delayed Chemotherapy					
Method	Detail	Parameter estimate	Standard error	*t* statistic	95% CI for hazard ratio
Extreme	100	-0.1040	0.0504	-2.06	(0.817, 0.995)
	0	-0.0927	0.0504	-1.84	(0.826, 1.006)
LOCF		-0.0929	0.0555	-1.67	(0.817, 1.016)
Median	by patient	-0.1124	0.0521	-2.16	(0.807, 0.990)
	time period	-0.1050	0.0505	-2.08	(0.816, 0.994)
	time period and trt arm	-0.1095	0.0507	-2.16	(0.811, 0.990)
Linear regression	previous coping scores	-0.0937	0.0556	-1.69	(0.817, 1.015)
	concurrent variables	-0.1030	0.0536	-1.92	(0.812 1.002)

Table 3.9 Summary of Time-Dependent Cox Model Analysis Considering Extended Model Stratified by Trial

Square root of coping score (S_Pacis)					
Method	Detail	Parameter estimate	Standard error	*t* statistic	95% CI for hazard ratio
Extreme	100	0.0240	0.0080	3.00	(1.008, 1.040)
	0	-0.0144	0.0087	-1.66	(0.969, 1.003)
LOCF		0.0050	0.0112	0.45	(0.983, 1.027)
Median	by patient	0.0136	0.0107	1.27	(0.993, 1.035)
	time period	0.0236	0.0117	2.02	(1.001, 1.048)
	time period and trt group	0.0251	0.0118	2.13	(1.002, 1.050)
Linear regression	previous coping scores	0.0076	0.0124	0.61	(0.983, 1.032)
	concurrent variables	0.0117	0.0112	1.04	(0.990, 1.034)
Delayed Chemotherapy					
Method	Detail	Parameter estimate	Standard error	*t* statistic	95% CI for hazard ratio
Extreme	100	-0.1058	0.0504	-2.10	(0.815, 0.993)
	0	-0.0947	0.0505	-1.88	(0.824, 1.004)
LOCF		-0.0949	0.0555	-1.71	(0.816, 1.014)
Median	by patient	-0.1149	0.0521	-2.21	(0.805, 0.987)
	time period	-0.1072	0.0505	-2.12	(0.814, 0.992)
	time period and trt arm	-0.1127	0.0507	-2.22	(0.809, 0.987)
Linear regression	previous coping scores	-0.0958	0.0556	-1.72	(0.815, 1.013)
	concurrent variables	-0.1045	0.0536	-1.95	(0.811, 1.001)
Sufficient Early Chemotherapy					
Method	Detail	Parameter estimate	Standard error	*t* statistic	95% CI for hazard ratio
Extreme	100	-0.0968	0.0504	-1.92	(0.822, 1.002)
	0	-0.0980	0.0503	-1.95	(0.827, 1.008)
LOCF		-0.0994	0.0555	-1.79	(0.812, 1.009)
Median	by patient	-0.0898	0.0520	-1.73	(0.826, 1.012)
	time period	-0.0950	0.0504	-1.88	(0.824, 1.004)
	time period and trt group	-0.1004	0.0505	-1.99	(0.819, 0.999)
Linear regression	previous coping scores	-0.0997	0.0555	-1.80	(0.812, 1.009)
	concurrent variables	-0.1116	0.0536	-2.08	(0.805, 0.993)

Table 3.9 Summary of Time-Dependent Cox Model Analysis Considering Extended Model Stratified by Trial (continued)

Oestrogen Receptor Positive Status					
Method	Detail	Parameter estimate	Standard error	*t* statistic	95% CI for hazard ratio
Extreme	100	-0.1570	0.0569	-2.76	(0.764, 0.956)
	0	-0.1590	0.0569	-2.79	(0.763, 0.954)
LOCF		-0.1144	0.0630	-1.82	(0.788, 1.009)
Median	by patient	-0.1278	0.0591	-2.16	(0.784, 0.988)
	time period	-0.1589	0.0569	-2.79	(0.763, 0.954)
	time period and trt arm	-0.1591	0.0569	-2.80	(0.763, 0.954)
Linear regression	previous coping scores	-0.1148	0.0630	-1.82	(0.788, 1.009)
	concurrent variables	-0.1188	0.0608	-1.95	(0.788, 1.000)

The estimated mean difference is shown in Table 3.10. The estimated mean difference is calculated by the sum of the mean difference in each simulated dataset with coping scores artificially removed divided by the number of simulated datasets. The estimated standard deviation of the difference is calculated by the sum of the standard deviation of the difference in each simulated dataset divided by the number of simulated datasets.

Table 3.10 Estimated Mean Difference Between Imputed Coping Score and Missing Coping Score Following Imputation in Simulated Datasets with Coping Scores Artificially Removed

Method	Detail	Estimated Mean Diff	SD of Estimate Mean Diff	Estimated SD of Diff	Range of Estimated SD of Diff
LOCF		-0.73		20.64	
Median	by patient	2.09		18.01	
	time period	11.22		25.17	
	time period and trt group	10.19		25.00	
Linear regression	previous coping scores	5.36		18.35	
	concurrent variables	9.83		26.03	

diff = difference; sd=standard deviation; trt = treatment group

Estimated difference between the imputed coping score and the missing coping score is the estimate of the real value of the missing coping score – the imputed coping score

The distribution of coping scores artificially removed and the distribution of imputed coping scores following median imputation is summarised in Table 3.11. The mean of the median of coping scores artificially removed is calculated by the sum of the median coping score in each simulated dataset with coping scores artificially removed divided by the total number of simulated datasets. The mean of the median of imputed coping scores is calculated similarly. The mean of the interquartile range of coping scores artificially removed is calculated by the sum of the interquartile range in each simulated dataset with coping scores artificially removed divided by the total number of simulated datasets. The mean of the interquartile range of imputed coping scores is calculated similarly.

Table 3.11 Summary of Distribution of Coping Scores Artificially Removed and Distribution of Imputed Coping Score Following Median Imputation in Simulated Datasets with Coping Scores Artificially Removed

Method	Detail	Mean of Median of Coping Scores Artifically Removed	Mean Interquartile Range of Coping Scores Artifically Removed	Mean of Median of Imputed Coping Scores	Mean Interquartile Range of Imputed Coping Scores
Median	by patient	21.94	39.51	21.46	30.53
	time period	21.94	39.51	19.74	5.31
	time period and trt group	21.94	39.51	19.92	6.06

trt = treatment group

3.5.2 Summary of Simple Imputation Methods

Assumptions in Applying Standard Simple Imputation Methods to the IBCSG Dataset

Standard simple imputation methods were applied to the IBCSG dataset in order to investigate the relationship between quality of life and DFS in a time-dependent Cox model. It is not plausible that the missing coping scores were all approximately 100 (lowest quality of life) or all approximately 0 (highest quality of life). Extreme imputation illustrates the lowest possible collection of quality of life scores or highest possible collection of quality of life scores (Table 3.8), but is not a suitable method of imputation for the IBCSG dataset. Linear regression

using concurrent variables is also not a suitable method of imputation for the IBCSG dataset because of the lack of suitable concurrent variables to model the coping score.

Given the indication that the missing coping scores in the IBCSG dataset are informative missing data, it is unlikely that the assumptions for the standard simple imputation methods hold. However, the purpose of applying simple imputation methods is generally as part of a sensitivity analysis into the sensitivity of results to the assumptions about the missing data.

Lack of Accuracy in the Imputed Coping Score

The estimated standard deviation of the difference between the imputed coping score and the missing coping score suggests a lack of accuracy when imputing the missing coping score by both standard simple and standard multiple imputation methods. It is reasonable to consider that a difference of 8 points between the coping score and the imputed coping score is clinically significant (see section 3.4.7). The estimated standard deviation of this difference was more than twice this for all standard imputation methods. Therefore an individual imputed coping score may represent a quality of life clinically significantly different from the missing coping score. The estimated standard deviation of the difference between the imputed and the missing coping score was similar for all the standard simple methods, around 20-25 (Table 3.10).

Suggested Bias in the Imputed Coping Score

For the standard simple imputation methods except LOCF the estimated mean difference between the imputed coping score and the missing coping score was greater than 2, out of a range 0 -100 (Table 3.10). This suggests that the imputed coping score is generally lower than the real value of the missing coping score and that these standard imputation methods may be systematically underestimating the missing coping scores. In this case, the completed dataset(s) then represent a poorer quality of life than was experienced by the patients.

Further, for median imputation by i) time period and ii) by time period and treatment group, the estimated difference between the between the imputed and the missing coping score represents a clinically significant difference. Together with the lack of accuracy in the imputed coping scores, this indicates that these would not be suitable imputation methods for the IBCSG dataset. This also applies to linear regression using concurrent coping scores, which has already been noted as not being a suitable method of imputation for the IBCSG dataset.

There was no suggestion that imputation by LOCF is systematically over- or underestimating the missing coping scores (Table 3.10).

Parameter Estimate for Square Root of Coping Score

The parameter estimate for S_Pacis was positive, favouring a positive relationship between quality of life and DFS, for all standard simple imputation methods (Table 3.8), except extreme imputation where missing coping scores were replaced with 0 (the highest quality of life). S_Pacis was a significant parameter in the time-dependent Cox model analysis for DFS following median imputation by time period and median imputation by time period and treatment group. However, coping scores imputed using median imputation by time period and by time period and treatment group may not follow the same distribution as the missing coping scores (Table 3.11). The imputed coping scores may be clinically significantly different from the missing coping score. As median imputation by i) time period and ii) by time period and treatment group are not suitable methods of imputation for the IBCSG dataset, the statistical significance of β_{sp} is discounted. Therefore, there was no evidence from the standard imputation methods of a statistically significant or clinically important relationship between quality of life and DFS in the IBCSG dataset.

Parameter Estimate for Delayed Chemotherapy

The parameter estimate for delayed chemotherapy was negative, favouring a positive relationship between further treatment with delayed chemotherapy and

DFS, for all standard imputation methods (Table 3.8). Delayed chemotherapy was a significant parameter following extreme imputation where missing coping scores were replaced with 100 (the worst quality of life) and median imputation. However, this should be interpreted cautiously because of the assumptions of the imputation methods. The remaining standard simple imputation methods indicated a trend towards a positive relationship. The time-dependent Cox model analysis following imputation is consistent with the finding from the main efficacy analysis that there may be a therapeutic benefit from delayed chemotherapy (The International Breast Cancer Group 1997).

The estimate of β_{del} was similar for all standard simple imputation methods, around -0.1 (Table 3.8). The estimates of β_{sp} from the time-dependent Cox model analysis without imputation for the available monotone and all available analyses were similarly around -0.1 (Table 3.3). The parameter estimate for the complete case analysis, though smaller in magnitude, also indicated a trend towards a positive relationship between delayed chemotherapy and DFS (Table 3.3). However, again the bias from the selection of patients considered in the analyses with no imputation makes it difficult to interpret the values of the parameter estimates.

3.6 Conclusions

This chapter investigated the influence of missing quality of life values, as assessed by coping score, when exploring the relationship between quality of life and DFS in a time-dependent Cox model. As the missing coping scores are informative missing data, the all available analysis (Table 3.3) is not appropriate. Preliminary investigations indicated that imputation is appropriate for the IBCSG dataset. While recognising the limitation of simple imputation methods, they may give useful information about the sensitivity of the results to assumptions about missing data. The standard simple imputation methods noted in section 2.4 were applied to the IBCSG dataset according to the technical details in section 3.4.

Assumptions when Applying Simple Imputation

Extreme imputation and linear regression using concurrent variables are not a suitable method of imputation for the IBCSG dataset. Given the indication that the missing coping scores in the IBCSG dataset are informative missing data, it is unlikely that the assumptions for the standard simple imputation methods hold. However, potentially the simple imputation methods could provide information as part of a sensitivity analysis into the sensitivity of results to the assumptions about the missing data.

Parameter Estimates from Time-Dependent Cox Model

With one exception, the parameter estimate for S_Pacis was positive, favouring a positive relationship between quality of life and DFS, for all standard simple imputation methods. The estimate of β_{sp} was close to 0 for all standard simple imputation methods (Table 3.8). There was no evidence of a statistically significant or clinically relationship between quality of life and DFS.

The parameter estimates for delayed chemotherapy showed a trend towards a positive relationship between delayed chemotherapy and DFS. The magnitude of the parameter estimate, around -0.1, was little influenced by the imputation method (Table 3.8) or by not applying imputation in the available monotone and all available analyses (Table 3.3). The trend towards a positive relationship is consistent with the finding from the main efficacy analysis that there may be a therapeutic benefit from delayed chemotherapy.

Performance of Standard Imputation Methods

The investigation of the performance of the standard simple imputation methods (section 3.5) found that:

i) there was a suggestion of a lack of accuracy when imputing the missing coping score

ii) the standard simple imputation methods except LOCF may be systematically underestimating the missing coping scores

Limitations of Simple Imputation Methods

Limitations of simple imputation, including underestimation of the variance of observations, have been noted. Multiple imputation methods should also be considered as part of an investigation of the relationship between quality of life and DFS in the IBCSG dataset. Standard multiple methods are applied to the IBCSG dataset in Chapter 4.

4 Investigation of the Effects of Using Multiple Imputation Methods to Estimate the Effect of Quality of Life on Disease-Free Survival in IBCSG Trials VI and VII

4.1 Introduction

In the previous chapter, standard simple imputation methods were applied to the IBCSG dataset before analysis as part of an investigation of the relationship between quality of life and DFS. This chapter describes applying multiple imputation methods to the IBCSG dataset. IBCSG Trials VI and VII are described in section 3.2.

The selection of the standard multiple imputation methods that are applied is described in section 2.4. As noted, MCMC methods are not applied as they are more computational complex than other standard multiple imputation methods. The multiple imputation methods that will be applied are:

i) bootstrapping: subgroups defined by baseline coping score and subgroups defined by previous coping score
ii) nearest neighbour imputation
iii) predictive mean matching
iv) pattern mixture models – Curran's analytical technique

The technical details of applying the multiple imputation methods are described in section 4.2. The results from the time-dependent Cox model analysis following standard multiple imputation methods are presented in section 4.3. The standard multiple imputation methods are compared in the same way as the standard simple imputation methods. The relationship between the imputed values is explored in section 4.4. A summary of the chapter and its implications is presented in section 4.5.

4.2 Technical Details of Application of Multiple Imputation Methods Applied to the IBCSG Breast Cancer Trial Data

In this section, the technical details of applying the standard multiple imputation methods to the IBCSG dataset and of the subsequent time-dependent Cox model analysis are described. Again, addressing the general missing data pattern in the IBCSG dataset and the patients included in the dataset for the time-dependent Cox model analysis is described in particular.

4.2.1 Introduction

Similarly to standard simple imputation methods, standard multiple imputation methods were used to impute missing coping scores before analysis of DFS. The standard multiple imputation methods considered are listed in section 4.1. The results from the time-dependent Cox model analysis after applying standard imputation methods to the IBCSG dataset are shown in section 4.3. Fifty repetitions of multiple imputation were performed leading to 50 completed datasets. The reason for number of repetitions is described in section 4.2.2.

Missing Baseline Coping Scores and Missing Data Pattern

The standard multiple methods applied require a baseline coping score. As with the standard simple imputation methods in this scenario, the patients with a missing baseline coping score were excluded from the time-dependent Cox model analysis. In contrast, given the large number of patients with a non-monotone missing data pattern (Figure 3.3; Table 3.2), excluding these patients from the time-dependent Cox model analysis would not be a reasonable approach in this chapter. A monotone missing data pattern was created by imputing non-monotone missing coping scores by LOCF before applying selected multiple imputation methods. These coping scores were replaced by LOCF as there was no suggestion of bias in the imputed coping scores from LOCF in the IBCSG dataset (Table 3.10, column Estimated Mean Diff) and it is a commonly used simple imputation

method (e.g. Vogel et al. 2002). The standard multiple imputation methods concerned were:

i) nearest neighbour imputation (section 4.2.4)

ii) predictive mean matching (section 4.2.5)

iii) pattern mixture models – Curran's analytic method (section 4.2.6)

Calculating Imputed Values for Coping Score

For the following standard multiple imputation methods, S_Pacis was considered during the imputation. From these imputed values, the imputed value of the missing coping score was calculated, rounded to the nearest whole number.

i) nearest neighbour imputation (section 4.2.4)

ii) predictive mean matching (section 4.2.5)

iii) pattern mixture models – Curran's analytic method (section 4.2.6)

Calculations during the imputation by the *MI procedure* in SAS were based on standardised values of S_Pacis.

Quality of Life Assessments Considered in the Time-Dependent Cox Model Analysis Following Imputation Methods

As noted, there were 2687 patients randomised to Trials VI and VII. Some patients could not be considered in the time-dependent Cox model analysis following standard imputation methods and for 2 patients the expected date of quality of life assessment was used (see section 3.4.1). The 456 patients with a missing baseline coping score (approximately at randomisation) were not considered in the time-dependent Cox model analysis after applying the following standard multiple imputation methods. Thus, for the standard multiple imputation methods listed above, 2214 patients were considered in the time-dependent Cox model analysis. The status of the coping scores considered for time-dependent Cox model analysis at each time points considered was as shown in Table 3.4 (column "Main Group").

The fact that there were a large number of non-monotonic missing coping scores can be seen below in Table 4.1. Among the 3 multiple imputation methods where a monotone missing data pattern was created by LOCF, the proportion of non-monotone missing coping scores imputed by LOCF decreases from 78.5% at Month 3 (Time 2) to 34.3% at Month 21 (Time 8) and 141 then was 0 at Month 24 (Time 9).

Table 4.1 Status of Coping Scores for Time-Dependent Cox Model Analysis

	Time	Main Group	Imputed without LOCF
Imputed	1	0	0
	2	492	106
	3	502	139
	4	498	175
	5	451	194
	6	484	226
	7	499	260
	8	478	314
	9	470	470

Main group = i) last observation carried forward, ii) bootstrapping, subgroups defined by baseline coping score and subgroups defined by previous coping score, iii) linear regression with previous coping score(s), iv) nearest neighbour imputation, v) predictive mean matching, and vi) pattern mixture models – Curran's analytic technique; imputed without LOCF= imputed coping scores excluding coping scores imputed by last observation carried forward in order to create a monotone missing data pattern for i) nearest neighbour imputation, ii) predictive mean matching and iii) pattern mixture models – Curran's analytic method

Comparing Imputation Methods

The standard multiple imputation methods were compared using simulated datasets in the same was the standard simple imputation methods (see section 3.4.1). For each of the first 100 simulated datasets with coping scores artificially removed, 10 repetitions of multiple imputation was performed and the difference between the imputed coping score and the real coping score originally observed and artificially removed was calculated. From these differences, the mean and

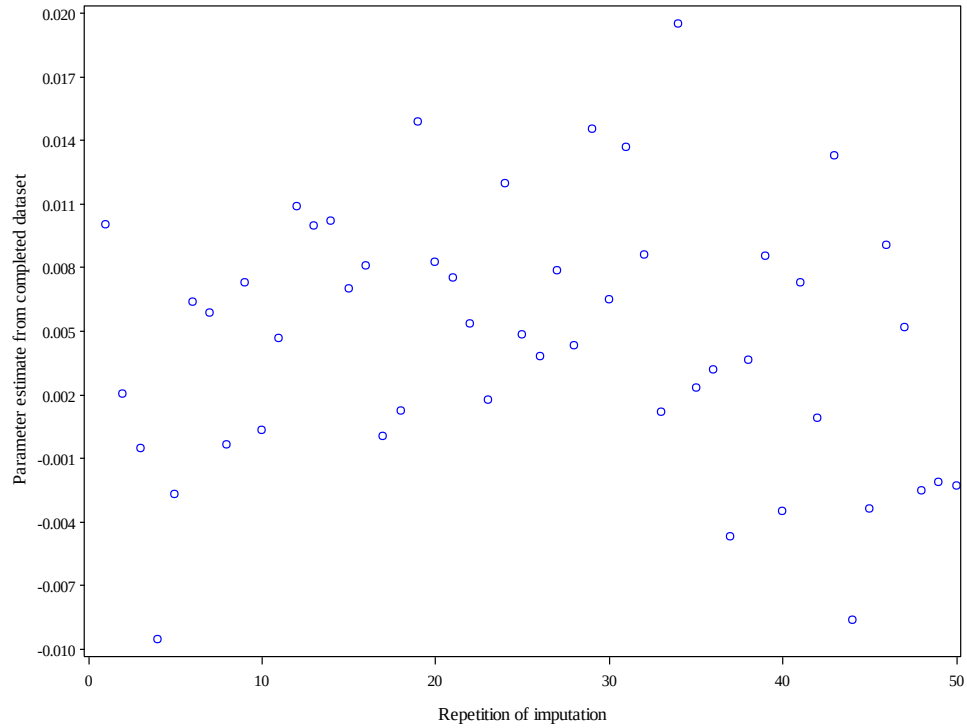
standard deviation of the difference between the imputed coping score and the unknown real value of the missing coping score was estimated for the imputation method. The reason for the number of repetitions is described in section 4.2.2 and the results are shown in section 4.3.

4.2.2 Reason for Number of Repetitions of Multiple Imputation and Simulated Datasets with Coping Scores Artificially Removed

Reason for 50 Repetitions of Multiple Imputation

As noted, Graham et al. recommends that two issues are considered in determining the appropriate number of repetitions of multiple imputation: i) the fraction of missing information γ and ii) the statistical power to detect a small effect size in the parameter estimate compared to maximum likelihood methods (see section 2.3). The issues related to determining the appropriate number of repetitions are the same for all of the standard multiple imputation methods and are illustrated here using bootstrapping, subgroups defined by baseline coping score. Figure 4.1 shows the parameter estimate from the completed dataset (A) and the cumulative mean parameter estimate (B) for S_Pacis from the time-dependent Cox model analysis versus the number of repetitions of imputation. The estimates displayed in the y-axis for the cumulative mean parameter estimate are calculated by combining the information from the number of completed datasets displayed in the x-axis. The estimates of β_{sp} from the completed datasets are randomly scattered around a mean. The cumulative mean estimate of β_{sp} converges after approximately 25 repetitions of multiple imputation. Figure 4.2 shows the same information for the parameter for delayed chemotherapy, with similar patterns.

A



B

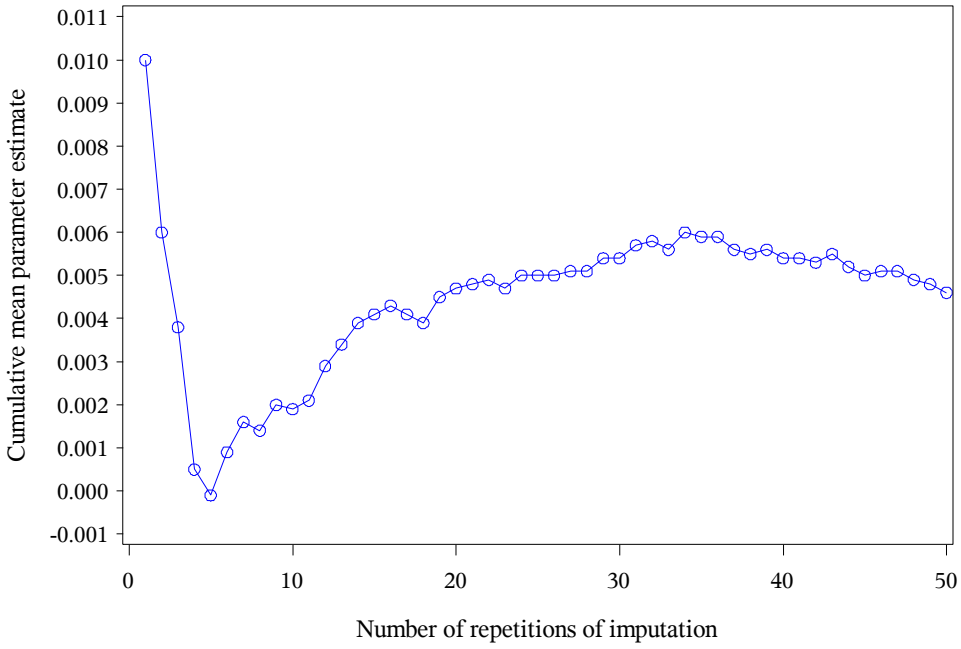
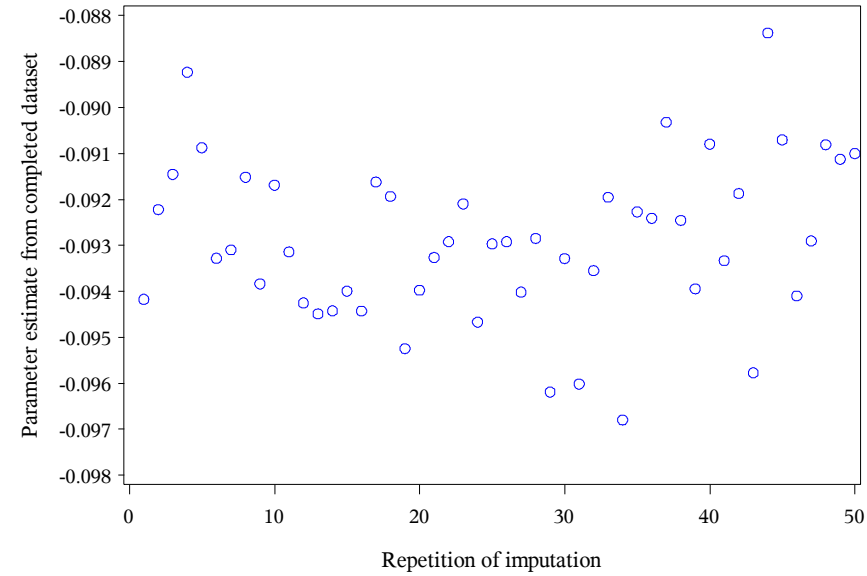


Figure 4.1 Parameter estimate from the completed dataset (A) and cumulative mean parameter estimate (B) for the square root of the coping score (S_Pacis) from the time-dependent Cox model analysis following bootstrap imputation, subgroups defined by baseline coping score by the number of repetitions of imputation

A



B

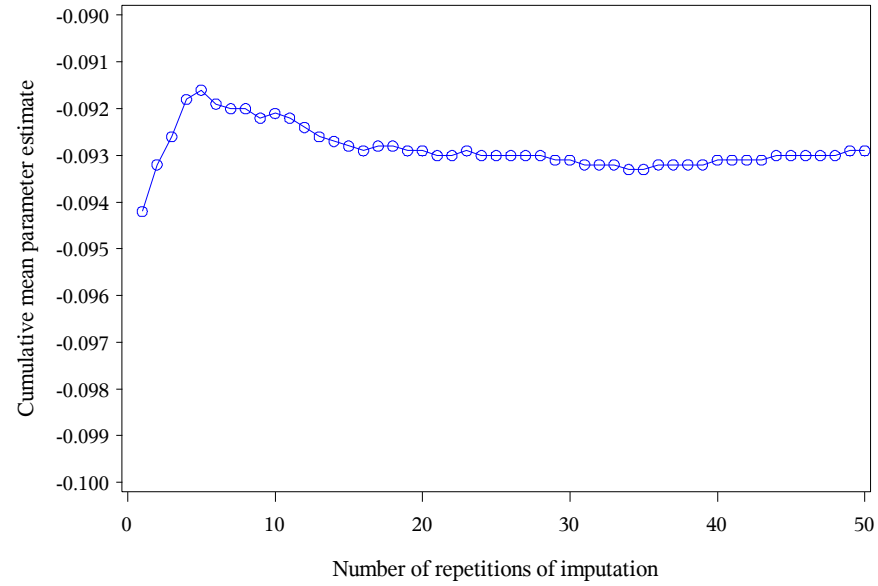


Figure 4.2 Parameter estimate from the completed dataset (A) and cumulative mean parameter estimate (B) for delayed chemotherapy from the time-dependent Cox model analysis following bootstrap imputation, subgroups defined by baseline coping score by the number of repetitions of imputation

As noted, the indicator for delayed chemotherapy, X_{del}, remains constant throughout the study. Therefore, the between-imputation component of the variance of β_{del} following 50 repetitions of imputation (0.000003) and the increase in variance (0.0010) is negligible. The variance decomposition for β_{sp} is shown below in Table 4.2:

Table 4.2 Variance Decomposition of the Parameter for Square Root of Coping Score (β_{sp}) from Time-Dependent Cox Model Analysis Following Bootstrap Imputation, Subgroups Defined by Baseline Coping Score

	K =5	K =40	K =50
Within-imputation (W)	0.000132	0.000132	0.000132
Between-imputation (B)	0.000051	0.000034	0.000037
Total variance	0.000193	0.000166	0.000170
Fraction of missing information	0.3052	0.2065	0.2251
Efficiency of estimate	0.9425	0.9949	0.9956

The efficiency of the estimate (see section 2.3) is very high (> 99%) after 40 imputations and little increased by increasing the number of repetitions of imputation to 50 (Table 4.2). The work by Graham et al. (2007) indicates that the power of the time-dependent Cox model analysis would only be negligibly increased by increasing the number of repetitions of imputation from 50 to 100. This indicates that there would be no benefit from performing more than 50 repetitions of imputation.

The figures for the parameter estimates from the completed dataset, the cumulative mean parameter estimate, and the decomposition of the variance of β_{sp} for the remaining standard multiple imputation methods are shown in Appendix A. There would be no benefit from performing more than 50 repetitions of imputation for any of the standard multiple imputation methods.

Reason for 10 Repetitions of 100 Simulated Datasets to Estimate Mean Difference Between the Imputed Coping Score and the Missing Coping Score For Multiple Imputation Methods

In the estimation of the difference between the imputed coping score and the missing coping score following multiple imputation, two components had to be considered i) the number of simulated datasets ii) the number of repetitions of multiple imputation applied to each of the simulated datasets. Performing 50 repetitions of multiple imputation on 150 simulated datasets is impractical due to computational time and so considering lower numbers was explored.

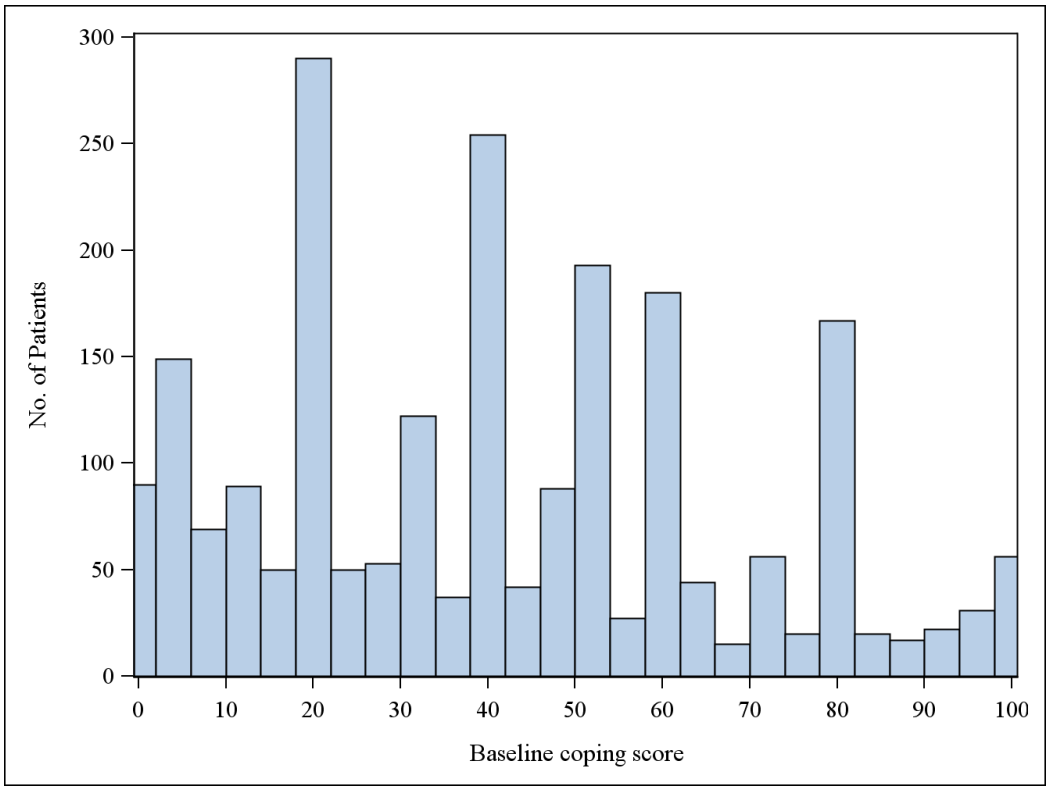
Considering the first 5 simulated datasets and imputation by bootstrapping, subgroups defined by baseline coping score, the estimated mean difference between the imputed coping score and the missing coping score from each simulated dataset was little influenced by increasing the number of repetitions from 10 to 30 (Table B2.1). Based on 10 repetitions, the range of the estimated mean difference from each simulated dataset was little influenced by considering more than 75 simulated datasets (Table B2.2). This suggests that it is reasonable to consider 10 repetitions of multiple imputation for 100 simulated datasets in the estimation of the mean and standard deviation of the difference between the imputed coping score and the missing coping score in the IBCSG dataset.

The considerations relating to the power to detect a small effect size are similar to those when applying simple imputation (section 3.4.7). It is reasonable to assume that even a small difference (~ 3) that is not clinically meaningful between the imputed coping score and the real coping score artificially removed would be found in each of the simulated completed datasets.

4.2.3 Bootstrapping from Subgroup of Patients

Two methods of defining the subgroups for bootstrapping were considered: i) baseline coping score and ii) previous coping score. The histogram of the baseline coping score is shown in Figure 4.3:

Figure 4.3 Histogram of Baseline Coping Scores



The histogram (Figure 4.3) suggests that the patients were more likely to judge how far along the line applies to her in round values, as noted may happen in Fairclough 2010. The subgroups selected for the baseline coping scores were influenced by this fact and that high coping scores (> 60) indicating poor quality of life were less common than lower quality of life scores. The same subgroups were used for baseline coping score and previous coping scores.

These subgroups were:

0 – 6

7 – 17

18 – 22

23 – 32

33 – 42

43 – 52

53 – 62

63 - 82

83 – 100

For each missing coping score, the set of potential imputed values is all the observed coping scores at the same time period among the patients in the same subgroup as the patient with the missing coping score. The imputed coping score is selected at random from the set of potential imputed values by *proc surveyselect* in SAS with the method option set to urs (for unrestricted random sampling).

4.2.4 Nearest Neighbour Imputation

A monotone missing monotone pattern was created by imputing non-monotone missing coping scores by LOCF. Nearest neighbour imputation was then implemented using the *MI procedure* in SAS (see section 2.3.5). During the imputation, the linear regression model for S_Pacis at each time point based on the previous coping score(s) was considered.

4.2.5 Predictive Mean Matching

Predictive mean matching, initial steps as for nearest neighbour imputation was implemented similarly to nearest neighbour imputation (section 4.2.3).
For completeness, time-dependent Cox model analysis following predictive mean matching, initial steps as for bootstrapping was also considered (Table 4.5). The outline of the procedure in SAS was as follows:

i) Generate *K* bootstrap samples, by sampling with replacement, of patients with observed values. The subgroups for bootstrapping are defined by the previous coping score (section 4.2.2).
ii) For each of the *K* samples, estimate the model parameters. The linear regression model for S_Pacis at each time point based on the previous coping score(s) is considered.

iii) Generate predicted values of S_Pacis for patients with both observed values and missing observations.

iv) For each patient with a missing observation, identify the patients with an observed coping score with the 5 closest predicted values to the patient with the missing coping score. The absolute difference between the predicted value for the patient with observed coping score and the patient with the missing coping score is considered. There may be 2 predicted values for observed coping scores giving the same absolute difference. Precedence is given to the lower of 2 such predicted values in order to ensure 5 closest predicted values are identified.

v) Select one of the patients with the 5 closest predicted values at random and impute the observed value for this selected patient for the missing observation using *proc surveyselect*.

4.2.6 Pattern Mixture Models - Curran's Analytic Technique

Curran's analytic technique for complete case missing value restriction was used to replace missing coping scores in completed datasets (see section 2.3.6). A monotone missing data pattern was created by imputing non-monotone missing coping scores by LOCF. Curran's analytic technique was then implemented using the *MI procedure* in SAS. During the imputation, S_Pacis at time h, Y_h (h=1,..,9), were assumed to follow a multivariate Normal distribution. The range for imputed values of S_Pacis was specified as 0-10 using the minimum and the maximum option.

Issues when implementing MCMC methods are summarised in section 2.3.1. Default settings for the MCMC statement applying MCMC methods in the *MI procedure* in SAS were used. Therefore, during the imputation i) a single chain was used, ii) the initial parameter estimates from the EM algorithm were used as starting values, iii) the burn-in length was 200 and iv) the run length was 200.

4.3 Results from Applying Multiple Imputation Methods to the IBCSG Dataset

In this section, the contents of summary tables of results from applying multiple imputation methods to the IBCSG dataset are described (section 4.3.1) and then results are summarised (section 4.3.2).

4.3.1 Description of Contents of Tables of Results from Applying Standard Imputation Methods to the IBCSG Dataset

The results from the time-dependent Cox model analysis following standard imputation multiple methods are shown in Table 4.3. The parameter estimate is the mean of the parameter estimate from each of the 50 completed datasets. The variance of the parameter estimate for each of the multiple imputation methods is calculated based on the 50 repetitions of multiple imputation according to (2.13). The standard error shown is the square root of the variance of the parameter estimate. The standard error and the parameter estimate are again used to calculate the 95% confidence interval for the parameter estimate. Results from the extended time-dependent Cox model are also presented in Table 4.4. The results from predictive mean matching, initial steps based on bootstrapping are shown in Table 4.5. These further results are calculated similarly to Table 4.3.

Table 4.3 Summary of Time-Dependent Cox Model Analysis Considering Square Root of Coping Score (S_Pacis) and Delayed Chemotherapy Stratified by Trial

Square root of coping score (S_Pacis)						
Method	Detail	Parameter estimate	Range	Standard error	*t* statistic	95% CI for hazard ratio
Bootstrap	baseline coping score	0.0046	-0.0096 to 0.0195	0.0130	0.35	(0.979, 1.030)
	previous coping score	0.0070	-0.0027 to 0.0194	0.0125	0.56	(0.983, 1.032)
Nearest neighbour		0.0030	-0.0037 to 0.0108	0.0118	0.25	(0.980, 1.026)
Predictive mean matching	initial steps as described for NNI	0.0043	-0.0026 to 0.0149	0.0121	0.36	(0.981, 1.028)
Pattern mixture models		0.0127	0.0061 to 0.0231	0.0123	1.03	(0.989, 1.037)
Delayed Chemotherapy						
Method	Detail	Parameter estimate	Range	Standard error	*t* statistic	95% CI for hazard ratio
Bootstrap	baseline coping score	-0.0929	-0.0968 to -0.0884	0.0556	-1.67	(0.802, 1.020)
	previous coping score	-0.0937	-0.0976 to -0.0907	0.0556	-1.69	(0.802, 1.020)
Nearest neighbour		-0.0926	-0.0948 to -0.0905	0.0560	-1.65	(0.802, 1.021)
Predictive mean matching	initial steps as described for NNI	-0.0928	-0.0963 to -0.0909	0.0555	-1.67	(0.803, 1.020)
Pattern mixture models		-0.0956	-0.1009 to -0.0933	0.0556	-1.72	(0.800, 1.018)

Table 4.4 Summary of Time-Dependent Cox Model Analysis Considering Extended Model Stratified by Trial

Square root of coping score (S_Pacis)						
Method	Detail	Parameter estimate	Range	Standard error	*t* statistic	95% CI for hazard ratio
Bootstrap	baseline coping score	0.0050	-0.0089 to 0.0199	0.0130	0.38	(0.980, 1.030)
	previous coping score	0.0076	-0.0017 to 0.0202	0.0116	0.66	(0.985, 1.030)
Nearest neighbour		0.0036	-0.0030 to 0.0114	0.0118	0.31	(0.980, 1.027)
Predictive mean matching	initial steps as described for NNI	0.0048	-0.0020 to 0.0153	0.0121	0.40	(0.981, 1.029)
Pattern mixture models		0.0133	0.0067 to 0.0238	0.0123	1.08	(0.989, 1.037)
Delayed Chemotherapy						
Method	Detail	Parameter estimate	Range	Standard error	*t* statistic	95% CI for hazard ratio
Bootstrap	baseline coping score	-0.0949	-0.0989 to -0.0904	0.0556	-1.71	(0.800, 1.018)
	previous coping score	-0.0958	-0.1001 to -0.0929	0.0556	-1.72	(0.800, 1.018)
Nearest neighbour		-0.0946	-0.0969 to -0.0925	0.0556	-1.70	(0.801, 1.019)
Predicted mean matching	initial steps as described for NNI	-0.0948	-0.0985 to -0.0929	0.0556	-1.71	(0.801, 1.019)
Pattern mixture models		-0.0978	-0.1035 to -0.0953	0.0556	-1.76	(0.798, 1.016)
Sufficient Early Chemotherapy						
Method	Detail	Parameter estimate	Range	Standard error	*t* statistic	95% CI for hazard ratio
Bootstrap	baseline coping score	-0.0990	-0.1010 to -0.0969	0.0555	-1.78	(0.797, 1.015)
	previous coping score	-0.0996	-0.1038 to -0.0980	0.0555	-1.79	(0.796, 1.014)
Nearest neighbour		-0.0990	-0.1006 to -0.0980	0.0555	-1.78	(0.797, 1.015)
Predictive mean matching	initial steps as described for NNI	-0.0992	-0.1013 to -0.0981	0.0555	-1.79	(0.797, 1.014)
Pattern mixture models		-0.1008	-0.1028 to -0.0995	0.0555	-1.82	(0.795, 1.013)

Table 4.4 Summary of Time-Dependent Cox Model Analysis Considering Extended Model Stratified by Trial (continued)

Oestrogen Receptor Positive Status						
Method	Detail	Parameter estimate	Range	Standard error	*t* statistic	95% CI for hazard ratio
Bootstrap	baseline coping score	-0.1149	-0.1160 to -0.1140	0.0630	-1.82	(0.768, 1.016)
	previous coping score	-0.1151	-0.1168 to -0.1128	0.0630	-1.83	(0.768, 1.015)
Nearest neighbour		-0.1150	-0.1153 to -0.1144	0.0630	-1.83	(0.768, 1.015)
Predicted mean matching	initial steps as described for NNI	-0.1148	-0.1152 to -0.1134	0.0630	-1.82	(0.768, 1.015)
Pattern mixture models		-0.1145	-0.1158 to -0.1136	0.0630	-1.82	(0.768, 1.015)

Table 4.5 Summary of Time-Dependent Cox Model Analysis Stratified by Trial Following Imputation by Predictive Mean Matching, Initial Steps Based on Bootstrap

Model Considering Square Root of Coping Score and Delayed Chemotherapy					
Parameter	Parameter estimate	Range	Standard error	*t* statistic	95% CI for hazard ratio
Square root of coping score (S_Pacis)	0.0064	-0.0039 to 0.0171	0.0125	0.51	(0.982, 1.031)
Delayed Chemotherapy	-0.0935	-0.0976 to -0.0902	0.0556	-1.68	(0.802, 1.020)
Extended Model					
Parameter	Parameter estimate	Range	Standard error	*t* statistic	95% CI for hazard ratio
Square root of coping score (S_Pacis)	0.0071	-0.0033 to 0.0181	0.0125	0.57	(0.983, 1.032)
Delayed Chemotherapy	-0.0957	-0.1002 to -0.0923	0.0556	-1.72	(0.800, 1.018)
Suff Early Chemotherapy	-0.0996	-0.1016 to -0.0979	0.0555	-1.79	(0.796, 1.014)
Oestrogen Receptor Positive Status	-0.1151	-0.1163 to -0.1144	0.0630	-1.83	(0.768, 1.015)

The estimated mean difference between the imputed coping score and the missing coping score is again the estimate of the real value of the missing coping score – the imputed coping score. The estimated mean difference is shown in Table 4.6. It is calculated by the sum of the mean difference from the completed dataset following each repetition of multiple imputation in each simulated dataset with coping scores artificially removed divided by the total number of completed datasets. The standard deviation of the estimated mean difference is defined as the standard deviation of the mean difference from the completed dataset following each repetition of multiple imputation in each simulated dataset divided by the total number of completed datasets. The mean of the estimated standard deviation is calculated by the sum of the standard deviation of the difference from the completed dataset following each repetition of multiple imputation in each simulated dataset divided by the total number of completed datasets. The mean *t*-statistic is the mean parameter estimate divided by the mean standard error.

Table 4.6 Estimated Mean Difference Between Imputed Coping Score and Missing Coping Score Following Imputation in Simulated Datasets with Coping Scores Artificially Removed

Method	Detail	Estimated Mean Diff	SD of Estimate Mean Diff	Estimated SD of Diff	Range of Estimated SD of Diff
Bootstrap	baseline coping score	3.60	1.04	30.63	28.55 to 33.02
	previous coping score	2.88	0.99	27.75	25.20 to 30.20
Nearest neighbour		-0.55	0.77	21.43	19.16 to 23.41
Predicted mean matching	initial steps as described for NNI	-0.22	0.77	21.28	19.90 to 22.70
Pattern mixture models		-0.52	0.76	21.52	19.49 to 23.74

diff = difference; sd=standard deviation; trt = treatment group

Estimated difference between the imputed coping score and the missing coping score is the estimate of the real value of the missing coping score – the imputed coping score

4.3.2 Summary of Multiple Imputation Methods

Standard multiple imputation methods were applied to the IBCSG dataset in order to investigate the relationship between quality of life and DFS in a time-dependent Cox model. The standard multiple imputation methods are not necessarily good estimation techniques in this context due to the assumptions relating to the missing data mechanism and the missing data pattern. The performance of imputation methods has also been considered in the statistical literature (see section 4.5).

Creating a Monotone Missing Data Pattern

Many multiple imputation methods assume a monotone missing data pattern, which is not the case the IBCSG dataset. In the standard multiple imputation methods except for bootstrapping, the non-monotone missing data patterns were imputed by LOCF. Between Month 3 (Time 2) and Month 21 (Time 8) at least 30% of missing coping scores were imputed by LOCF (Table 3.4). Multiple imputation then proceeds based on the dataset with a monotone missing data pattern. This gives the advantage compared to LOCF imputation of generating multiple completed datasets.

Lack of Accuracy in the Imputed Coping Score

Similarly to the standard simple imputation methods, the estimated standard deviation of the difference between the imputed coping score and the missing coping score suggests a lack of accuracy when imputing the missing coping score. The estimated standard deviation of the difference between the imputed and the missing coping score was similar for the standard simple methods and the more complex standard multiple imputation methods, around 20-25 (Table 3.10 and Table 4.6).

Suggested Bias in the Imputed Coping Score

For the standard multiple imputation method of bootstrapping, the estimated mean difference between the imputed coping score and the missing coping score was

greater than 2, out of a range 0 -100 (Table 4.6). This suggests that the imputed coping score is generally lower than the real value of the missing coping score and that these standard imputation methods may be systematically underestimating the missing coping scores. In this case, the completed dataset(s) then represent a poorer quality of life than was experienced by the patients. As with LOCF, there was no suggestion that the remaining standard multiple imputation methods systematically over- or under-estimate the missing coping scores (Table 4.6). This may be influenced by the fact that the non-monotone missing data patterns were imputed by LOCF.

Parameter Estimate for Square Root of Coping Score

The parameter estimate for S_Pacis was positive, favouring a positive relationship between quality of life and DFS, for all standard multiple imputation methods (Table 4.3). Similarly to the standard simple imputation methods, there was no evidence from the standard multiple imputation methods of a statistically significant or clinically important relationship between quality of life and DFS in the IBCSG dataset.

The multiple imputation methods showed parameter estimates of S_Pacis which were similar for each repetition (Figure 4.1A). The estimate of β_{sp} was close to 0 for all standard multiple imputation methods, with absolute magnitude less than 0.013 (Table 4.3). This similar to the estimates of β_{sp} from the time-dependent Cox model analyses without imputation carried out for reference and illustrative purposes (Table 3.3) and following simple imputation (Table 3.8).

Parameter Estimate for Delayed Chemotherapy

The parameter estimate for delayed chemotherapy was negative, favouring a positive relationship between further treatment with delayed chemotherapy and DFS, for all standard multiple imputation methods (Table 4.3). As with simple imputation, this is consistent with the finding from the main efficacy analysis that

there may be a therapeutic benefit from delayed chemotherapy (The International Breast Cancer Group 1997).

As with the parameter estimate for S_Pacis, the multiple imputation methods showed parameter estimates of delayed chemotherapy which were similar for each repetition (Figure 4.2A). The estimate of β_{del} was similar for all standard imputation methods, around -0.1 (Table 4.3). The estimates of β_{sp} from the time-dependent Cox model analysis without imputation for the available monotone and all available analyses (Table 3.3) and following simple imputation (Table 3.8) were similarly around -0.1.

Standard Error of Parameter Estimate for Square Root of Coping Score and Delayed Chemotherapy

Simple imputation does not address the issue of underestimation of the variance of observations (see section 2.2.9). Ignoring extreme imputation, the standard error of the parameter estimate for S_Pacis (~0.011) and delayed chemotherapy (~0.054) following simple imputation methods (Table 3.8) is approximately equal to the standard error considering the available monotone or all available analyses (Table 3.3). The standard error of the parameter estimates following simple imputation have not increased to reflect the uncertainty in the imputed values. As every imputed value was the same, imputing the extreme values for illustrative purposes led to a decrease in the standard error of the parameter estimates compared to the available monotone or all available analyses. In contrast, the standard error of the parameter estimates following multiple imputation showed a small increase to reflect this uncertainty. The standard error of the parameter estimate for S_Pacis and delayed chemotherapy increased by ~14% to ~0.0125 (Table 4.3) compared to ~0.011 (Table 3.3) and by ~4% to ~0.056 (Table 4.3) compared to ~0.054 (Table 3.3) respectively. Due to the smaller sample size, the standard error of the parameter estimates from the complete case analysis (Table 3.3) were approximately twice as large as the standard errors from the available monotone and all available analyses and following standard imputation methods.

4.4 Cluster Analysis of Imputed Values in the IBCSG Dataset

This section investigates whether or not it is reasonable to assume the imputed values in the IBCSG dataset following the standard imputation methods reflect the imputation method applied. Under the scenario where the imputed values are similar regardless of the imputation method applied, the parameter estimates from the analytic model would be little influenced by the imputation method applied. The similarities/differences between the imputed values following the standard imputation methods is explored in hierarchical cluster analysis.

Suppose that the imputed values following imputation in the IBCSG dataset form natural clusters. It might then be reasonable to consider that among these clusters the completed dataset will be similar regardless of which imputation method is applied. Assuming the imputed values are influenced by the imputation method applied, we do not expect the imputed values to form natural clusters. Hierarchical cluster analysis was performed on the 3874 imputed values (Table 3.4, "Main group" column) following simple imputation and following multiple imputation. In the cluster analysis, the Euclidean distance between the imputed values was considered and three linkage methods were considered: average linkage, centroid method and single linkage (Everitt 1993, chapter 4).

4.4.1 Distance Measures and Linkage Methods

Hierarchical cluster analysis involves ordering observations to construct a dendrogram to explore the relationship among the observations. The starting point is a distance matrix describing the distance between observations. Generally, this distance matrix is based on the Euclidean distance (Everitt 1993, p.46). Let **X** be a matrix columns giving the *Q* variable values for each of the *n* observations being considered. The Euclidean distance is then (Everitt 1993, p.46):

$$d_{ij}^{Eucild} = \left(\sum_{q=1}^{Q} (x_{iq} - x_{jq})^2\right)^{1/2} \quad (4.1)$$

From the distance matrix, clusters of observations are identified. It is more common to determine these clusters by joining similar observations (agglomerative method) rather than removing dissimilar observations (divisive method). Popular methods for joining similar observations (linkage methods) include: i) average linkage, ii) centroid method and iii) single linkage (Everitt 1993, chapter 4).

These linkage methods share the same basic algorithm outlined below (Everitt 1993, chapter 4):

i) Create a cluster containing the two closest observations, in this case the two observations with the smallest Euclidean distance between them

ii) A third observation, which is next closest, is added to the two-observation cluster from step i) or a new two-observation cluster is formed

iii) Continue to agglomerate one additional observation until all the observations are in one cluster

However, the definition of the difference between clusters varies among these linkage methods (Everitt 1993, chapter 4).

i) Average linkage

the average distance from observations in one cluster to any observation in the second cluster

Let C_A and C_B be clusters of size N_A ($N_A \geq 1$) and N_B ($N_B \geq 1$) respectively. The distance between the two clusters, D_{AB}, is defined by:

$$D_{AB} = \frac{1}{N_A N_B} \sum_{i \in C_A} \sum_{j \in C_B} d_{ij} \tag{4.2}$$

ii) Centroid linkage

the (squared) Euclidean distance measured between the centroids of the two clusters

In this method, once a cluster is formed, it is represented by its mean vector as the centroid. The mean vector describes the mean of the *Q* variables of the observations in the cluster.

iii) Single linkage

the shortest distance from any observation in one cluster to any observation in second cluster

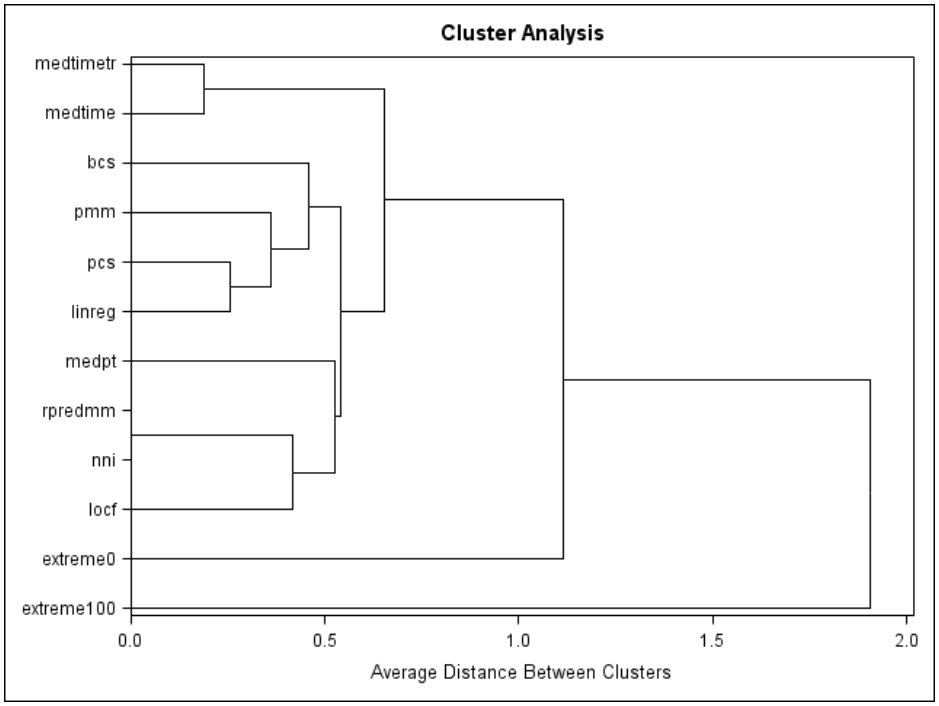
Displaying Distance Between Clusters

The calculation of the distance between clusters may be based on the distance matrix in the original scale or on squared distances. The distance between clusters displayed in the dendrogram may be in the original scale or as a Normalised value.

4.4.2 Imputed Values following Imputation

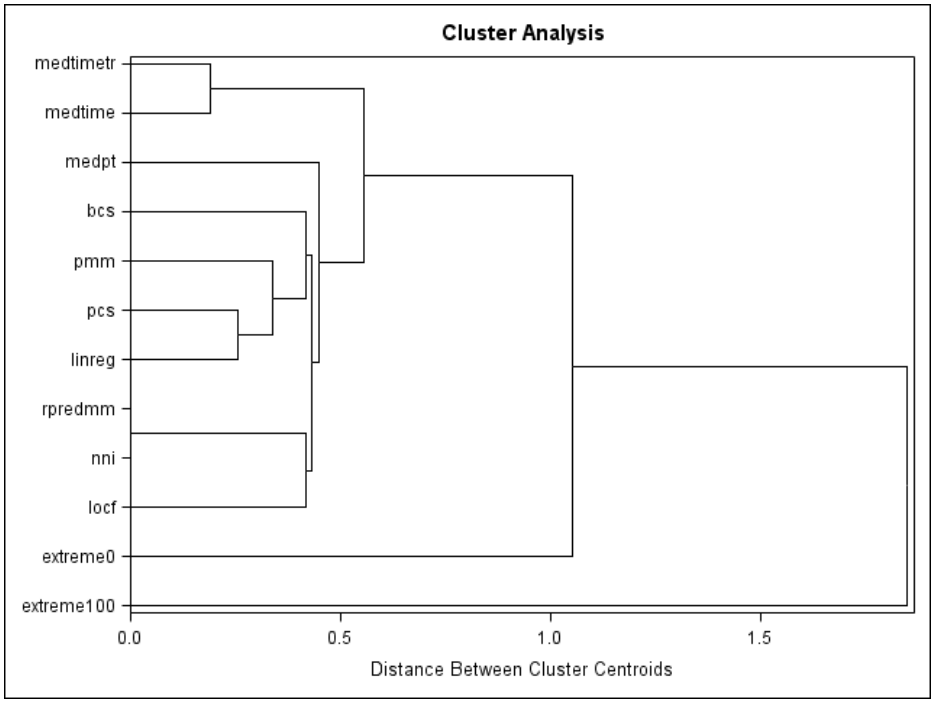
The imputed values of the missing coping scores were considered as observations for 7 of the standard simple imputation methods. The simple imputation method of linear regression using concurrent variables was not considered due to the lack of suitable concurrent variables. The median imputed value from 50 repetitions of the 5 standard multiple imputation methods was considered. This included missing coping scores imputed by LOCF in order to create a monotone missing data pattern. The dendrogram for single linkage was similar to the dendrogram for average linkage and is not shown in Figure 4.4. Here, the calculations of the distance between clusters was based on the squared Euclidean distance and the distance is displayed as the Normalised value.

Figure 4.4 A



medtimetr = median imputation by time period and treatment group; medtime = median imputation by time period; bcs = bootstrapping, subgroups defined by baseline coping score; pmm = pattern mixture models – Curran's analytic technique; pcs = baseline, subgroups defined by previous coping score; linreg = linear regression with previous coping score(s); medpt = median imputation by patient ; rpredmm = predictive mean matching; nni = nearest neighbour imputation; locf = last observation carried forward; extreme0 = extreme imputation of 0; extreme100 = extreme imputation of 100

Figure 4.4 B



medtimetr = median imputation by time period and treatment group; medtime = median imputation by time period; medpt = median imputation by patient; bcs = bootstrapping, subgroups defined by baseline coping score; pmm = pattern mixture models – Curran's analytic technique; pcs = baseline, subgroups defined by previous coping score; linreg = linear regression with previous coping score(s); rpredmm = predictive mean matching; nni = nearest neighbour imputation; locf = last observation carried forward; extreme0 = extreme imputation of 0; extreme100 = extreme imputation of 100

Figure 4.4 Dendrogram for imputed values in the IBCSG dataset following standard imputation methods considering two different linkage methods: average linkage (A) and centroid linkage (B)

The cluster analysis indicated that imputation of a high value was least like the other standard imputation methods. Among the simple imputation methods, median imputation by time period and by time period and treatment group were most similar (Figure 4.4). Considering the multiple imputation methods, nearest neighbour imputation and predicted mean matching were most similar. The imputed values from bootstrapping, subgroups defined by previous coping score were more like the imputed values from pattern mixture models – Curran's

analytical technique than the imputed values from bootstrapping, subgroups defined by baseline coping score (Figure 4.4). As with simple imputation, there was no suggestion that the multiple imputation methods form natural clusters (Figure 4.4).

4.4.3 Summary of Cluster Analysis of Imputed Values

Imputation of a high value replaced missing the coping scores was the worst score for the quality of life, whereas in the IBCSG dataset, high coping scores (> 60, indicating poor quality of life) were less common than lower quality of life scores (Figure 4.3). This influenced the fact that that imputation of high values was least like the other standard imputation methods. Among the simple imputation methods, the median coping score by time period was similar to the median coping score by time period and treatment group. It was thus expected that the imputed values following these imputation methods are the most similar. Given the initial steps of both methods are the same, it was also expected that the imputed values of nearest neighbour imputation and predicted mean matching were most similar amongst the multiple imputation methods. Apart from bootstrapping, subgroups defined by baseline coping score, the implementation of the standard multiple imputation methods involved the previous coping score (see section 3.4). This influenced the fact that the imputed values from bootstrapping, subgroups defined by previous coping score were more like the imputed values from pattern mixture models – Curran's analytical technique than the imputed values from bootstrapping, subgroups defined by baseline coping score.

The fact that the imputed values do not appear to form natural clusters indicated that the imputed values reflect the imputation method applied. There was no suggestion from the cluster analysis that the completed datasets are all similar regardless of which imputation method is applied. The parameter estimates from the analytic model therefore also reflect the imputation method. This raises the possibility that the performance of the imputation methods in this setting is influenced by the actual relationship between quality of life and DFS.

4.5 Conclusions

This chapter continued the investigation of the influence of missing quality of life values, as assessed by coping score, when exploring the relationship between quality of life and DFS in a time-dependent Cox model. Standard simple imputation methods were applied to the IBCSG dataset in Chapter 3. Here, the standard multiple imputation methods noted in section 2.4 were applied to the IBCSG dataset according to the technical details in section 4.2.

Type of Missing Data and Missing Data Pattern

Many multiple imputation methods assume the data are MAR and a monotone missing data pattern, which is not the case in the IBCSG dataset. In order to address the missing data pattern, in the standard multiple imputation methods except bootstrapping, the non-monotone missing data patterns were imputed by LOCF.

Parameter Estimates from Time-Dependent Cox Model

The parameter estimate for S_Pacis was positive, favouring a positive relationship between quality of life and DFS, for all standard multiple imputation methods. As with standard simple imputation methods (Table 3.8), the estimate of β_{sp} was close to 0 for all standard multiple imputation methods (Table 4.3). This is similar to the fact that the parameter estimates for baseline coping score in Herring et al. (2004) were little influenced by the different models accounting for missing data. As noted, in Herring et al. (2004) poor baseline coping score was associated with improved relapse-free survival in postmenopausal patients. However, considering coping scores throughout the study in a time-dependent Cox model led to parameter estimates in the opposite direction and of a smaller magnitude. There was no evidence from the standard simple or standard multiple imputation methods of a statistically significant or clinically important relationship between quality of life and DFS.

As noted (section 4.3.2), the standard error of the parameter estimates following simple imputation did not increase compared to considering all available coping scores and thus did not reflect the uncertainty in the imputed values. In contrast, there was a small increase in the standard error of the parameter estimates following the standard multiple imputation methods (section 4.3.2).

The parameter estimates for delayed chemotherapy again showed a trend towards a positive relationship between delayed chemotherapy and DFS. The magnitude of the parameter estimate, around -0.1, was little influenced by the simple (Table 3.8) or multiple (Table 4.3) imputation method or by not applying imputation in the available monotone and all available analyses (Table 3.3). The trend towards a positive relationship is consistent with the finding from the main efficacy analysis that there may be a therapeutic benefit from delayed chemotherapy.

Performance of Standard Imputation Methods

As noted, the standard multiple imputation methods are not necessarily good estimation techniques in this context due to the assumptions relating to the missing data mechanism and the missing data pattern. The investigation of the performance of the standard multiple imputation methods (section 4.3) found that:

i) there was a suggestion of a lack of accuracy when imputing the missing coping score, similarly to the standard simple imputation methods (section 3.5).

ii) the standard multiple imputation method of bootstrapping may be systematically underestimating the missing coping scores, as with the standard simple imputation methods except LOCF (section 3.5),

There was no suggestion that the remaining multiple imputation methods systematically over- or underestimated the missing coping scores (section 4.3). This may be influenced by the fact that the non-monotone missing data patterns were imputed by LOCF.

The performance of imputation methods has also been considered in the statistical literature, for example Peyre et al. (2011), Ranstam et al. (2012) and Marshall et al. (2012). Peyre et al. (2011) and Ranstam et al. (2012) were among the references identified from a PubMed search. They are described here as they consider simple and multiple imputation of quality of life scores. While both of these examples considered missing quality of life scores, unlike in this chapter, quality of life was the outcome variable. A further difference was that in both these examples, only one multiple imputation method was considered. Also of note was that the number of repetitions of multiple imputation in these examples was less than in this chapter, with 20 and 30 repetitions respectively. Marshall et al. (2012) is described as it considers missing explanatory variables in a breast cancer dataset. It was identified from the citation of Herring et al. (2004). Unlike in this chapter, the explanatory variables were standard prognostic factors in breast cancer rather than quality of life and were not time-dependent. Again, the number of repetitions of multiple imputation, 20, was less than in this chapter.

Performance of imputation in the 2003 French Decennial Health Study (Peyre et al. [2011])

The quality of life assessment was the medical outcome study 36-item short-form health survey (SF-36). Samples of 300 and 1000 subjects were randomly drawn from the 2003 French Decennial Health Survey. Various patterns of missing data were generated according to three different item non-response rates (3, 6, and 9%) and three types of missing data: i) missing completely at random, ii) missing at random, and iii) informative missing data.

The multiple imputation method used a set of external covariates in a standard multiple regression model. Imputation by personal mean score, which is similar to median imputation by patient, was also considered. Personal mean score appeared appropriate for dealing with missing items from completed SF-36 questionnaires in most routine scenarios. However, the use of personal mean score was associated with small bias (relative bias <2%) in all studied situations. This is

similar to the imputation in the IBCSG dataset where the coping score may be systematically underestimated. In contrast to imputation in the IBCSG dataset, Peyre et al. (2011) found that multiple imputation improved accuracy and precision compared to personal mean score.

Performance of Imputation in the FREE Trial in Acute Painful Vertebral Fractures (Ramstam et al. [2012])

The FREE trial was a randomised, non-blinded study comparing balloon kyphoplasty with non-surgical care for the treatment of patients with acute painful vertebral fractures. The primary endpoint was the change from baseline to 1 month in quality of life assessed using the SF-36 physical component summary (PCS) scale. Five secondary endpoints were also considered.

The multiple imputation method used chained equations. The imputation model included all six outcomes at baseline and follow-up visits, all stratification factors, age, treatment centre and number of fractures at baseline (≥ 1), in addition to treatment 'as received'. LOCF and mixed-effect models were also considered.

Similarly to the IBCSG dataset, the amount of missing data increased during follow-up (1 month: 14.5%; 24 months: 28%). Overall patterns of missing response across time were similar for each treatment group. As with imputation in the IBCSG dataset, the alternative imputation methods used for substituting missing data produced similar results. Mixed-effect model analyses, rather than imputation, appeared to be the most appropriate method for analysing the FREE trial data.

Performance of Imputation in the Simulation Study Based on the German Breast Cancer Study Group Dataset (Marshall et al. [2010])

Datasets were generated to resemble the skewed distributions seen in a motivating breast cancer example. The motivational dataset assessed the prognostic ability of eight covariates on recurrence-free survival. Multivariate missing data were

imposed on four covariates using four different mechanisms: i) missing completely at random, ii) missing at random, iii) informative missing data and iv) a combination of these 3 mechanisms. Five amounts of patients with incomplete data from 5% to 75% were considered. The scenario considering 25% of patients with incomplete data was approximately in line with the percentage of missing coping scores in the IBCSG dataset. Similarly to the IBCSG dataset, transformation of continuous covariates was used to make the assumption of Normality more applicable.

A single imputation by predictive mean matching was considered, whereas multiple repetitions were performed in the IBCSG dataset. Multiple imputation was performed by i) two data augmentation techniques, ii) regression switching imputation, iii) regression switching with predictive mean matching (MICE-PMM) and iv) flexible additive imputation models. The last imputation method fitted separate flexible additive imputation models to each incomplete covariate. The imputation model included eight standard prognostic factors in addition to the survival time and event status. The results of the single imputation by predictive mean matching and multiple imputation by data augmentation were similar considering the scenario considering 25% of patients with missing data and informative missing data. This is similar to the imputation in the IBCSG dataset. The simulation study found that performing MICE-PMM may be the preferred approach provided that less than 50% of the patients have missing data and the missing data are not informative missing data.

Relationship Among the Imputed Values

The relationship among the imputed values from the standard imputation methods was investigated by hierarchical cluster analysis (section 4.4). Imputing a high value was least like the other standard imputation methods. As expected, median imputation by time period was similar to median imputation by time period and treatment group. Among the multiple imputation methods, nearest neighbour imputation was similar to predicted mean matching. The imputed values from

bootstrapping, subgroups defined by previous coping score were more like the imputed values from pattern mixture models – Curran's analytical technique than the imputed values from bootstrapping, subgroups defined by baseline coping score. As noted, the cluster analysis indicated that the imputed values following multiple imputation reflect the imputation methods (Figure 4.4).

Possible Influence of Relationship Between Quality of Life and DFS on Performance of Imputation Methods

As the imputed values reflect the imputation method, it is possible that the performance of the imputation methods in this setting is influenced by the relationship between quality of life and DFS. The performance of the standard imputation methods may not be the same in the context of a strong positive relationship between quality of life and DFS compared to a weak positive relationship or no relationship. As noted, the assumption in many of the standard imputation methods that the data are MAR does not hold. The performance of the standard imputation methods may not be the same when the data are informative missing data compared to when the MAR assumption is reasonable. The IBCSG dataset was the basis for simulated datasets with a known relationship between quality of life and DFS given different missing data mechanisms. These simulated datasets are used in Chapters 5 and 6 to investigate if the performance of the standard imputation methods given different missing data mechanisms is influenced by the relationship between quality of life and DFS.

5 Applying Simple Imputation Methods to Simulated Datasets with a Positive Relationship Between Quality of Life and Disease-Free Survival and a Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival

5.1 Introduction

There was no evidence of a statistically significant or clinically important relationship between quality of life and DFS in the IBCSG dataset from the time-dependent Cox model analysis following imputation of missing coping scores by standard simple and standard multiple imputation methods described in Chapters 3 and 4 respectively. There was a trend towards a positive relationship between delayed chemotherapy and DFS. The performance of standard imputation methods may be different when there is a positive relationship between quality of life and DFS compared to when there is no relationship, and this is investigated in this chapter.

Complete simulated datasets were generated with a positive relationship between quality of life and DFS and a positive relationship between delayed chemotherapy and DFS. Here, a high quality of life was associated with improved DFS and delayed chemotherapy was associated with improved DFS. These associations were from the parameters β_{sp} and β_{del} of the time-dependent Cox model respectively. The method for generating the complete simulated datasets is described in section 5.2.1 and these datasets are based on the 2231 patients from the IBCSG dataset with an observed baseline coping score (approximately at randomisation). As described in section 5.2.2, four combinations of β_{sp} and β_{del} were considered and for each of the 4 combinations, 150 complete simulated datasets were generated. The time-dependent Cox model analysis of the complete simulated datasets is described in section 5.2.2.

Simulated datasets with missing data were generated by artificially removing coping scores from the complete simulated datasets. An overview of generating the simulated datasets is given in Figure 5.1. The methods for artificially removing coping scores are described in section 5.2.2. The aim of simulating data was to investigate the influence of the missing data mechanism on the performance of standard imputation methods given different combinations of a positive relationship between quality of life and DFS and positive relationship between delayed chemotherapy and DFS. In assessing the performance, whether the positive relationship between quality of life and DFS and between delayed chemotherapy and DFS is found following imputation is of interest as well as the parameter estimates. This chapter describes applying standard simple imputation methods to the simulated datasets. Applying multiple imputation methods is described in Chapter 6.

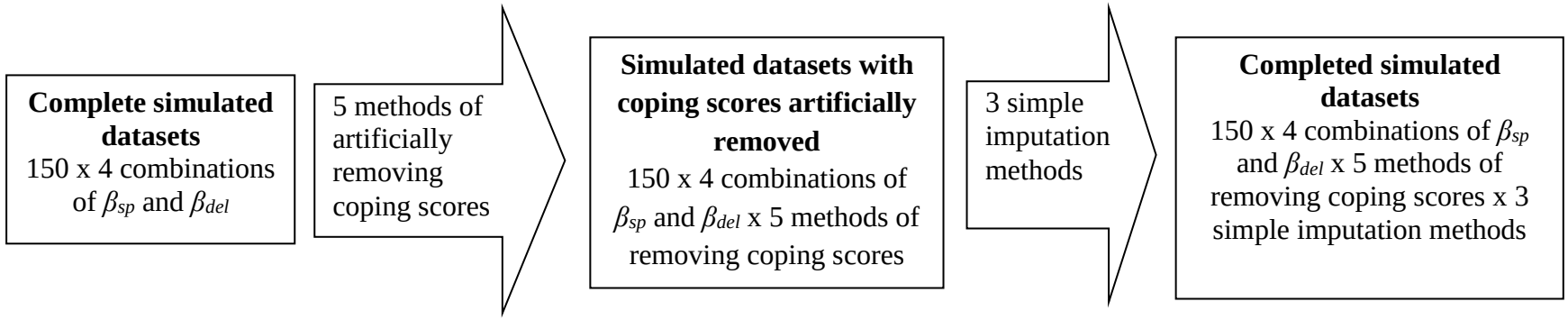
The standard simple imputation methods that are applied to the IBCSG dataset are listed in section 3.1. Three of the standard simple imputation methods applied to the IBCSG dataset were applied to the simulated datasets with a positive relationship between quality of life and DFS and a positive relationship between delayed chemotherapy and DFS:

i) LOCF (see section 3.4.3)
ii) median imputation by patient (see section 3.4.4)
iii) linear regression with previous coping score(s) (see section 3.4.5)

The explanation why the remaining standard simple imputation methods were not applied to the simulated datasets is given in section 3.5.2.

The time-dependent Cox model analysis following standard imputation methods is described in section 3.2.6. The technical details of the patients included in the time-dependent Cox model analysis are described in section 5.3 and the findings are described in section 5.4. The summary of the chapter is presented in section 5.5.

Figure 5.1 Overview of Generating Simulated Datasets



5.2 Simulated Datasets with a Positive Relationship Between Quality of Life and Disease-Free Survival and a Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival

It was noted that in the IBCSG dataset DFS times for patients with a DFS event and patients with no DFS event both approximately follow a Weibull distribution. It was straightforward to simulate DFS times from these Weibull distributions. These DFS times could be matched to the same matrix of coping scores for patients and the same indicator for delayed chemotherapy from the IBCSG dataset. This created complete simulated datasets with a positive relationship between quality of life and DFS and a positive relationship between delayed chemotherapy and DFS. The strength of the relationship between quality of life and DFS and between delayed chemotherapy and DFS was controlled by the parameters β_{sp} and β_{del} respectively (see section 5.2.1). As shown in the overview of generating the simulated datasets in Figure 5.1, there were 600 complete simulated datasets, 150 x 4 combinations of β_{sp} and β_{del}. There were 5 methods of artificially removing coping scores from each of the complete simulated datasets (150 x 4 x 5).

Three different simple imputation methods were applied to these simulated datasets with coping scores artificially removed. Time-dependent Cox model analysis was carried out on the 150 x 4 x 5 x 3 completed simulated datasets. The parameter estimates from the time-dependent Cox model analysis can therefore be thought as arising from an experiment with 150 replications where imputation method is nested within missing data mechanism, both nested within the combination of β_{sp} and $\beta_{del.}$ In this section, the method for simulating time to event data and creating the simulated datasets with coping scores artificially removed is summarised, with the technical details provided in Appendix C.

5.2.1 Simulating Time to Event Data

Method for Simulating Data

In the IBCSG dataset, the median follow-up time is 12.3 years. The DFS survival times in days for patients with a DFS event approximately followed a Weibull distribution with a shape parameter 1.199 and scale parameter 1519. The range was 0 days – 5786 days (190.1 months; 15.8 years). For patients with no DFS event recorded, the follow-up time in the trial approximately followed a Weibull distribution with a shape parameter 5.5014 and scale parameter 4997. The longest censored DFS time was 6212 days (204.1 months; 17.0 years).

MacKenzie and Abrahamowicz (2002) described an algorithm for randomly generating time-to-event data that arises from an interpretation of the expression for the partial likelihood. The method for simulating a positive relationship between quality of life and DFS and a positive relationship between delayed chemotherapy and DFS according to this algorithm is summarised as follows. A matrix of coping scores based on the patients' coping scores in the IBCSG dataset is considered. Simulated DFS times (event or censored) are simulated from Weibull distributions (see equation (C1)). The simulated DFS times are considered in ascending order and matched to a patient. The risk set of patients who have yet to be matched to a DFS time is identified. The probability of selection is calculated for each patient in the risk set of patients, and the patient to whom the DFS time is matched is selected (see equation (C2)). For times to DFS event, this selection probability is based on the covariates i) the centred S_Pacis and ii) indicator for delayed chemotherapy. To create a time- dependent process, the centred S_Pacis at the appropriate time period is used when calculating the selection probability. For censored DFS times, the selection probability is equal for all patients in the risk set. The patient matched is removed from the risk set and the steps repeated until all patients have been matched to a DFS time. The technical details are described in Appendix C, Part 1.

5.2.2 Complete Simulated Datasets and Artificially Removing Data

Complete Simulated Datasets

When generating the complete simulated datasets with a positive relationship between quality of life and DFS and a positive relationship between delayed chemotherapy and DFS, the 2231 patients with an observed baseline coping score (approximately at randomisation) were considered (Appendix C, Part 1).

Four combinations of β_{sp} and β_{del} data were considered. The complete simulated datasets had a positive relationship between quality of life and DFS and a positive relationship between delayed chemotherapy and DFS. Low coping scores correspond to high quality of life. The combinations considered a value for β_{sp} of 0.1 (weak) or 0.4 (strong) and a value for β_{del} of -0.165 (weak) or -0.195 (strong). For each of the 4 combinations, 150 simulated datasets were generated. The considerations for setting these parameter values relate to the time-dependent Cox model analysis of the complete simulated datasets, which is described next.

Time-Dependent Cox Model Analysis of Complete Simulated Datasets

The real date of quality of life assessments were used wherever possible in the complete simulated datasets. However, if using the real date of the quality of life assessment would lead to an interval less than 1 day, then the expected dates of visit calculated from the date of randomisation were used instead.

Example 5.1 Example of Using Expected Dates of Visit in Complete Simulated Datasets

Suppose a patient randomised on 24th October 1990 had the following visit dates:

	Actual	Expected
21 months (Time 8)	27th May 1992	22nd July 1992
24 months (Time 9)	29th July 1994	21st October 1992

If the simulated DFS for this patient was 825 days, ending on 26th January 1993, then the expected date of 21st October 1992 was used for the date of 24 months

(Time 9) quality of life assessment. For 3 patients with a date of assessment taken from the non-compliance form but a missing coping score, the expected date of quality of life assessment was used to prevent an interval of less than 1 day for the time-dependent Cox model analysis.

The plots of Schoenfeld residuals against time for the covariates S_Pacis and delayed chemotherapy from the time-dependent Cox model analysis of a complete simulated dataset for each of the 4 combinations of β_{sp} and β_{del} are shown in Appendix D. As noted, a zero slope indicates that the assumption of proportional hazards is reasonable. Beyond approximately 13.5 years (~ 5000 days), the plots no longer indicated a zero slope for delayed chemotherapy. However, this did not raise concerns about the time-dependent Cox model. The results from the time-dependent Cox model analysis of the complete simulated datasets are shown below in Table 5.1:

Table 5.1 Summary of Time-Dependent Cox Model Analysis Stratified by Trial of Complete Simulated Datasets

Square root of coping score (S_Pacis)					
Combination of *βsp* and *βdel*	Theoretical Value of *βsp*	Mean Parameter Estimate	Mean Standard Error	Number of 95% CIs for Hazard Ratio Containing 1	Number of 95% CIs for Parameter Estimate Containing Theoretical Value
Weak, weak	0.1	0.1007	0.0111	0 (0.0%)	150 (100.0%)
Weak, strong	0.1	0.1024	0.0111	0 (0.0%)	150 (100.0%)
Strong, weak	0.4	0.4019	0.0131	0 (0.0%)	150 (100.0%)
Strong, strong	0.4	0.4025	0.0131	0 (0.0%)	150 (100.0%)
Delayed chemotherapy					
Combination of *βsp* and *βdel*	Theoretical Value of *βdel*	Mean Parameter Estimate	Mean Standard Error	Number of 95% CIs for Hazard Ratio Containing 1	Number of 95% CIs for Parameter Estimate Containing Theoretical Value
Weak, weak	-0.165	-0.1721	0.0549	40 (26.7%)	150 (100.0%)
Weak, strong	-0.195	-0.1866	0.0549	19 (12.7%)	150 (100.0%)
Strong, weak	-0.165	-0.1531	0.0549	41 (27.3%)	150 (100.0%)
Strong, strong	-0.195	-0.1928	0.0549	17 (11.3%)	150 (100.0%)

CI = confidence interval

Of note, the 95% confidence interval for the parameter estimates for all of the complete simulated datasets contained the theoretical value (Table 5.1). The simulation process has been successful and the complete simulated datasets represent the relationship between quality of life and DFS and delayed chemotherapy and DFS intended. There was imprecision in the parameter estimate for delayed chemotherapy from the individual complete simulated datasets, reflected in the standard error of the parameter estimate (~0.055) (Table 5.1).

It is of interest to investigate if a relationship between quality of life and DFS could be masked by the missing data mechanism, or if a relationship is still found following imputation. The values for β_{sp} of 0.1 and 0.4 lead to a relationship between quality of life and DFS being found in all the complete simulated datasets (Table 5.1). This relationship should be found as well as the parameter estimate being unbiased following imputation in order to consider the imputation method performs well. A relationship between delayed chemotherapy and DFS was found in 73% (219/300) and 88% (264/300) of complete simulated datasets when considering the weak and strong relationship respectively (Table 5.1). The values for β_{del} of -0.165 and -0.195 approximately correspond to conventional values for the power of the hypothesis tests (80% and 90%) set in clinical trials (see section 1.1.4).

Artificially Removing Data

For each of the complete simulated datasets 5 different methods of artificially removing coping scores were considered. These 5 methods represent 4 different scenarios and each of the 3 different categories (see section 1.6.1) for the missing data mechanism in the IBCSG dataset:

i) Higher coping scores (lower quality of life) have a higher probability of missingness: **informative missing data**

ii) Lower coping scores (higher quality of life) have a higher probability of missingness: **informative missing data**

iii) Later time periods have a higher probability of missingness: **MAR**

iv) Coping scores missing (completely) at random: **MCAR**

The most likely scenario in the IBCSG dataset was that higher coping scores (poorer quality of life) have a higher probability of missingness. Two methods of artificially removing data under this scenario were considered. Each of the methods were derived in order that in the simulated datasets approximately 30% of the expected coping score were missing, similar to the IBCSG dataset. The technical details of the 5 methods of artificially removing coping scores were as described in Appendix C, Part 2.

5.3 Technical Details of Patients Considered in Time-Dependent Cox Model Analysis of Simulated Datasets

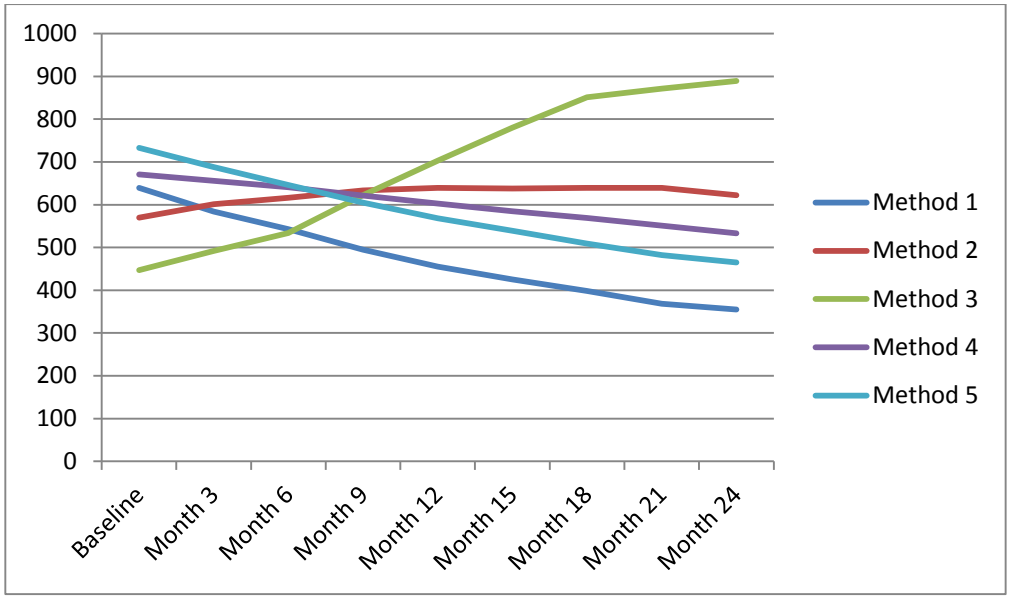
The status of coping scores for time-dependent Cox model analysis from the 600 simulated datasets, 150 simulated datasets in each of the 4 combinations of β_{sp} and β_{del}, with coping scores artificially removed according to a particular method is described in Table 5.2 and Figure 5.2:

Table 5.2 Summary of Status of Coping Scores in Simulated Datasets According to Method of Artificially Removing Coping Scores

Method of Artifically Removing Coping Scores	Time	Mean Number of Observed Coping Scores	Mean Number of Imputed Coping Scores	Mean Total Number of Coping Scores
Method 1	1	1592	639	2231
	3	1587	543	2131
	5	1556	455	2010
	7	1493	398	1891
	9	1422	355	1777
Method 2	1	1661	570	2231
	3	1515	616	2131
	5	1371	639	2010
	7	1252	639	1891
	9	1155	622	1777
Method 3	1	1784	447	2231
	3	1596	534	2131
	5	1307	703	2010
	7	1039	851	1891
	9	888	889	1777
Method 4	1	1560	671	2231
	3	1490	641	2131
	5	1408	603	2010
	7	1322	569	1891
	9	1244	533	1777
Method 5	1	1498	733	2231
	3	1485	646	2131
	5	1442	568	2010
	7	1382	509	1891
	9	1313	465	1777

Method 1: Higher coping scores have a higher chance of being missing.
Method 2: Lower coping scores have a higher chance of being missing.
Method 3: Later time period have a higher chance of being missing
Method 4: 30% of coping scores missing at random
Method 5: Higher coping scores have a higher chance of being missing

A



B

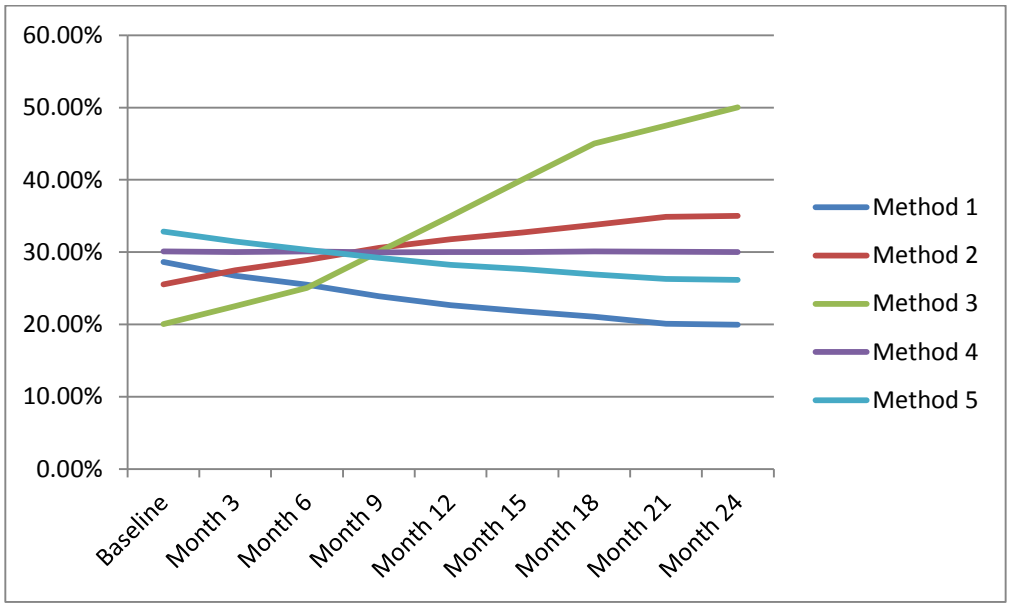


Figure 5.2 Mean number of imputed coping scores (A) and mean percentage of imputed coping scores (B) in simulated datasets according to method of artificially removing coping scores

Method 1: Higher coping scores have a higher chance of being missing.
Method 2: Lower coping scores have a higher chance of being missing.
Method 3: Later time period have a higher chance of being missing
Method 4: 30% of coping scores missing at random
Method 5: Higher coping scores have a higher chance of being missing

When higher coping scores (lower quality of life) were associated with a higher probability of missingness (Methods 1 and 5), the number and percentage of patients with missing coping score decreased across time in an approximately linear fashion. This number remained similar across time when lower coping scores (higher quality of life) were associated with a higher probability of missingness (Method 2), with the percentage of patients with missing coping score increasing in an approximately linear fashion. The method of later time period being associated with a higher probability of missingness (Method 3) was the most different from the other methods in terms of the number of patients and percentage of patients with missing coping score increases across time (Figure 5.2; Table 5.2).

When performing the imputation by LOCF and by linear regression using previous coping scores, the patients with the baseline coping score artificially removed could not be considered in the time-dependent Cox model analysis. This was a noticeable number of patients, with the number varying depending on the method of artificially removing coping scores. The time-dependent Cox model analyses following these imputation methods was based on an average of between 1498 and 1784 patients, compared to 2231 (Table 5.2).

Patients with no observed coping scores could not be considered in the time-dependent Cox model analysis when performing median imputation by patient. The number of such patients is summarised in Table 5.3. As shown in Table 5.3, the number of patients with no observed coping scores increased when higher coping scores (lower quality of life) were associated with a higher probability missingness compared to other methods. When higher coping scores were associated with a higher probability of missingness, this number increased when the strength of the relationship between quality of life and DFS increased from 0.1 to 0.4. This is due to the fact that when there is a strong relationship between quality of life and DFS, the patients with poorer quality of life are more likely to have a DFS event and thus a lower number of expected quality of life assessments

than when there is a weak relationship between quality of life and DFS. The number of patients with no observed coping scores was lowest overall when later time periods were associated with a higher probability of missingness. When considering the weak relationship between quality of life and DFS, the number of patients with no observed coping scores was higher when lower coping scores (higher quality of life) were associated with a higher probability of missingness compared to later time periods. The number of patients with no observed coping scores was similar for these missing data mechanisms when considering the strong relationship between quality of life and DFS.

Table 5.3 Summary of Number of Patients with No Simulated Observed Coping Scores According to Method of Artificially Removing Coping Scores

Combination of β_{sp} and β_{del}	Theoretical Value of β_{sp}	Theoretical Value of β_{del}	Method of Artificially Removing Coping Scores	Mean Number of Patients with No Observed Coping Scores
Weak, weak	0.1	-0.165	Method 1	24
	0.1	-0.165	Method 2	19
	0.1	-0.165	Method 3	13
	0.1	-0.165	Method 4	21
	0.1	-0.165	Method 5	29
Weak, strong	0.1	-0.195	Method 1	25
	0.1	-0.195	Method 2	19
	0.1	-0.195	Method 3	13
	0.1	-0.195	Method 4	21
	0.1	-0.195	Method 5	29
Strong, weak	0.4	-0.165	Method 1	33
	0.4	-0.165	Method 2	12
	0.4	-0.165	Method 3	12
	0.4	-0.165	Method 4	21
	0.4	-0.165	Method 5	38
Strong, strong	0.4	-0.195	Method 1	33
	0.4	-0.195	Method 2	11
	0.4	-0.195	Method 3	12
	0.4	-0.195	Method 4	21
	0.4	-0.195	Method 5	38

Method 1 and Method 5: Higher coping scores have a higher chance of being missing.
Method 2: Lower coping scores have a higher chance of being missing.
Method 3: Later time period have a higher chance of being missing
Method 4: 30% of coping scores missing at random

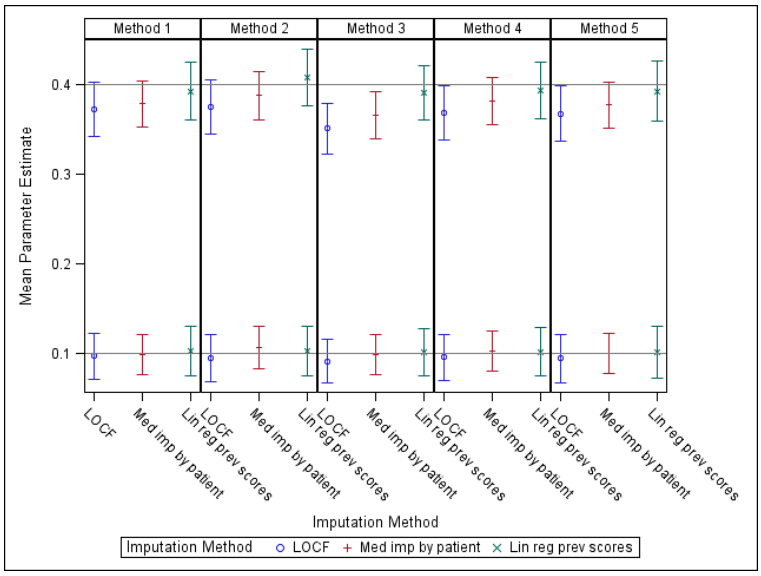
5.4 Findings from Applying Simple Imputation Methods to Simulated Datasets with a Positive Relationship Between Quality of Life and Disease-Free Survival and Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival

This section describes the findings from the time-dependent Cox model analysis of DFS on S_Pacis and delayed chemotherapy stratified by trial for the 4 combinations of a positive relationship between quality of life and DFS and a positive relationship between delayed chemotherapy and DFS following simple imputation of simulated datasets. For each completed simulated dataset, the estimate of β_{sp} and β_{del} from the time-dependent Cox model stratified by trial was recorded. Then the

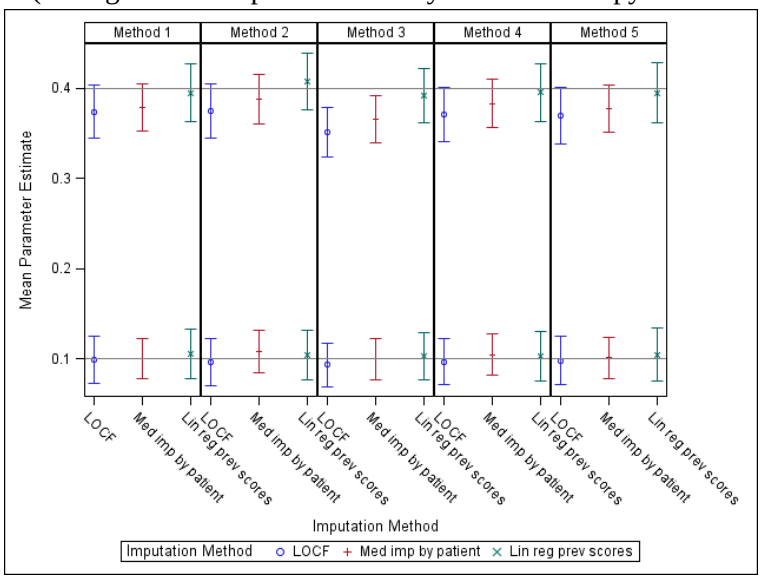
i) mean parameter estimate
ii) mean standard error of the parameter estimate
iii) number of 95% confidence intervals for hazard ratio containing 1
iv) number of 95% confidence intervals for parameter estimate containing the simulated value

were calculated for each combination according to the method of artificially removing coping scores (see Appendix C, Part 2 for technical details of the methods). The results are shown in Appendix E. An overview of the estimate of β_{sp} is shown in Figure 5.3 and the results are summarised in Table 5.4 – Table 5.6.

A (Weak relationship between delayed chemotherapy and DFS)



B (Strong relationship between delayed chemotherapy and disease-free survival)



Footnotes for figure on next page

Figure 5.3 Mean parameter estimate for square root of coping score (S_Pacis) from the time-dependent Cox model analysis and 95% confidence interval based on mean standard error from simple imputation in 150 simulated datasets. The weak and strong relationship between quality of life and disease-free survival is shown in top and lower portion of the figure respectively, with the combination of A) weak relationship between delayed chemotherapy and disease-free survival; (B) strong relationship between delayed chemotherapy and disease-free survival

Footnotes for Figure 5.3:

Method 1: Higher coping scores have a higher chance of being missing
Method 2: Lower coping scores have a higher chance of being missing
Method 3: Later time period have a higher chance of being missing
Method 4: 30% of coping scores missing at random
Method 5: Higher coping scores have a higher chance of being missing

LOCF = last observation carried forward;
Med imp by patient = median imputation by patient;
Lin reg prev scores: linear regression using previous coping score(s)

Table 5.4 Summary of Findings from Applying Last Observation Carried Forward to Simulated Datasets with Positive Relationship Between Quality of Life and Disease-Free Survival and Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival

Combination of β_{sp} and β_{del}	Suggestion of poorest performance [1]	Robustness of Estimate of β_{sp} [2]	Power to Find Significance of β_{del} [3]	Robustness of Estimate of β_{del} [4]
Weak, weak β_{sp}=0.1, β_{del}= -0.165	Little influence of missing data mechanism	Robust to missing data mechanism	Lowest when higher coping scores ⇒ missingness	Robust to missing data mechanism
Weak, strong β_{sp}=0.1, β_{del}= -0.195	Higher coping scores ⇒ missingness	Robust to missing data mechanism	Lowest when higher coping scores ⇒ missingness	Trend to underestimate $\beta_{del,}$ most noticeable when higher coping scores ⇒ missingness
Strong, weak β_{sp}=0.4, β_{del}= -0.165	Higher coping scores ⇒ missingness or later time periods ⇒ missingness	β_{sp} underestimated, bias most extreme when later time periods ⇒ missingness	Lowest when higher coping scores ⇒ missingness	β_{del} underestimated, most extreme when higher coping scores ⇒ missingness
Strong, strong β_{sp}=0.4, β_{del}= -0.195	Higher coping scores ⇒ missingness or later time periods ⇒ missingness	β_{sp} underestimated, bias most extreme ⇒ when later time periods missingness	Lowest when higher coping scores ⇒ missingness	β_{del} underestimated, most extreme when higher coping scores ⇒ missingness

Table 5.5 Summary of Findings from Applying Median Imputation by Patient to Simulated Datasets with Positive Relationship Between Quality of Life and Disease-Free Survival and Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival

Combination of β_{sp} and β_{del}	Suggestion of poorest performance [1]	Robustness of Estimate of β_{sp} [2]	Power to Find Significance of β_{del} [3]	Robustness of Estimate of β_{del} [4]
Weak, weak β_{sp}=0.1, β_{del}= -0.165	Little influence of missing data mechanism	Robust to missing data mechanism	Little or no reduction	Robust to missing data mechanism
Weak, strong β_{sp}=0.1, β_{del}= -0.195	Higher coping scores ⇨ missingness	Robust to missing data mechanism	Trend towards a small reduction	Trend to underestimate β_{del}, most noticeable when higher coping scores ⇨ missingness
Strong, weak β_{sp}=0.4, β_{del}= -0.165	Higher coping scores ⇨ missingness or later time periods ⇨ missingness	β_{sp} underestimated, bias most extreme when later time periods ⇨ missingness	Little or no reduction	β_{del} underestimated, most extreme when higher coping scores ⇨ missingness
Strong, strong β_{sp}=0.4, β_{del}= -0.195	Higher coping scores ⇨ missingness or later time periods ⇨ missingness	Trend to underestimate β_{sp}	Trend towards a small reduction	Trend to underestimate β_{del}, most noticeable when higher coping scores ⇨ missingness

β_{sp} = parameter estimate for square root of coping score (S_Pacis); β_{del} = parameter estimate for delayed chemotherapy

1: Considers the robustness of β_{sp} and β_{del} and the power to find the significance of β_{del}

2: Summarised from columns "Mean Parameter" estimate and "Bias (%): …" for section "Median imputation" in Tables E1.1, E2.1, E3.1 and E4.1

3: Summarised from column "Number of 95% Cis for hazard ratio containing 1" for section "Median imputation" in Tables E1.2, E2.2, E3.2 and E4.2

4: Summarised from columns "Mean Parameter" estimate and "Bias (%): …" for section "Median imputation" in Tables E1.2, E2.2, E3.2 and E4.2

Table 5.6 Summary of Findings from Applying Linear Regression using Previous Coping Scores to Simulated Datasets with Positive Relationship Between Quality of Life and Disease-Free Survival and Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival

Combination of β_{sp} and β_{del}	Suggestion of poorest performance [1]	Robustness of Estimate of β_{sp} [2]	Power to Find Significance of β_{del} [3]	Robustness of Estimate of β_{del} [4]
Weak, weak β_{sp}=0.1, β_{del}= -0.165	Little influence of missing data mechanism	Robust to missing data mechanism	Lowest when higher coping scores ⇒ missingness	Robust to missing data mechanism
Weak, strong β_{sp}=0.1, β_{del}= -0.195	Higher coping scores ⇒ missingness	Robust to missing data mechanism	Lowest when higher coping scores ⇒ missingness	Trend to underestimate β_{del}, most noticeable when higher coping scores ⇒ missingness
Strong, weak β_{sp}=0.4, β_{del}= -0.165	Higher coping scores ⇒ missingness	Robust to missing data mechanism	Lowest when higher coping scores ⇒ missingness, jointly with coping scores missing at random	β_{del} underestimated, most extreme when higher coping scores ⇒ missingness
Strong, strong β_{sp}=0.4, β_{del}= -0.195	Higher coping scores ⇒ missingness	Robust to missing data mechanism	Lowest when higher coping scores ⇒ missingness	Trend to underestimate β_{del}, most noticeable when higher coping scores ⇒ missingness

β_{sp} = parameter estimate for square root of coping score (S_Pacis); β_{del} = parameter estimate for delayed chemotherapy

1: Considers the robustness of β_{sp} and β_{del} and the power to find the significance of β_{del}

2: Summarised from columns "Mean Parameter" estimate and "Bias (%): …" for section "Linear regression" in Tables E1.1, E2.1, E3.1 and E4.1

3: Summarised from column "Number of 95% Cis for hazard ratio containing 1" for section "Linear regression" in Tables E1.2, E2.2, E3.2 and E4.2

4: Summarised from columns "Mean Parameter" estimate and "Bias (%): …" for section "Linear regression" in Tables E1.2, E2.2, E3.2 and E4.2

A relationship between quality of life and DFS was found in all the completed simulated datasets (column Number of 95% CIs for hazard ratio containing 1, Table E1.1; Table E2.1; Table E3.1; Table E4.1). As noted, following imputation by LOCF and by linear regression using previous coping scores, the patients with the baseline coping score artificially removed could not be considered.

The time-dependent Cox model analyses of the completed simulated datasets indicates a lack of precision in the estimates of β_{del} and led to a wide range of the parameter estimates from each of the completed simulated datasets. For example, the range of the parameter estimate following LOCF was (-0.4322, 0.0846) when considering the combination of weak positive relationship between quality of life and DFS and weak positive relationship between delayed chemotherapy and DFS and when higher coping scores (lower quality of life) were associated with a higher probability of missingness (first row of Table E5.1). The imprecision is reflected in the fact that the mean standard error of the parameter estimate was at least 0.055 (Table E1.2; Table E2.2; Table E3.2; Table E4.2). The details on the performance of the individual standard simple imputation methods are given below.

5.4.1 Last Observation Carried Forward

When considering the weak positive relationship between quality of life and DFS, the parameter estimate for S_Pacis was robust to not being able to consider patients with a missing baseline coping score and the missing data mechanism. The mean parameter estimate for S_Pacis was between 0.0914 and 0.0992 compared to the theoretical value of 0.1, with a mean standard error around 0.013 (Figure 5.3; Table E1.1; Table E2.1).

In contrast, the parameter estimate for S_Pacis was generally biased towards 0 following imputation by LOCF when considering the strong positive relationship between quality of life and DFS (Table E3.1; Table E4.1). The mean parameter estimate for S_Pacis was around 0.35-0.38 compared to theoretical value of 0.4,

with a mean standard error around 0.015. This indicates a bias of around 6% to 12%. Only a low number, 62% or less, of the 95% confidence intervals for the parameter estimate for S_Pacis for the completed simulated datasets following imputation by LOCF contained the parameter estimate for the complete simulated dataset (column Number of 95% CIs for parameter estimate containing simulated value, Table E3.1 and Table E4.1). The bias in the parameter estimate was most extreme when later time periods were associated with a higher probability of missingness. In this case, almost none of the completed simulated datasets following imputation by LOCF contained the parameter estimate for the complete simulated dataset (Method 3, Table E3.1; Table E4.1).

When considering the weak positive relationship between delayed chemotherapy and DFS, approximately 27% of the complete simulated datasets failed to find a relationship between delayed chemotherapy and DFS (Table 5.1). Not being able to consider the patients with a missing baseline coping score in the imputation by LOCF led to a noticeable decrease in the probability of finding a relationship between delayed chemotherapy and DFS, where between 31.3% and 44.7% of the completed simulated datasets failed to find this relationship (column Number of 95% CIs for hazard ratio containing 1, Table E1.2 and Table E3.2). The proportion of complete simulated datasets failing to find this relationship was lower when considering the strong positive relationship between delayed chemotherapy and DFS, approximately 12% (Table 5.1). The probability of finding a relationship between delayed chemotherapy and DFS again decreased following LOCF, where between 17.3% and 29.3% of the completed simulated datasets failed to find this relationship (Table E2.2; Table E4.2). The probability of finding this relationship was lowest when higher coping scores (lower quality of life) were associated with a higher probability of missingness (Method 1 and Method 5), when at least 24% of completed simulated datasets failed to find this relationship (Table E2.2; Table E4.2).

The parameter estimate for delayed chemotherapy was robust to not being able to consider patients with a missing baseline coping score and the missing data mechanism when considering the combination of weak positive relationship between quality of life and DFS and weak positive relationship between delayed chemotherapy and DFS. It was close to the theoretical value of -0.165 for each of 5 methods of artificially removing data (Table E1.2). For the remaining combinations of β_{sp} and β_{del}, the trend was the for the parameter estimate to be closer to 0 following imputation by LOCF (Table E2.2; Table E3.2; Table E4.2). The bias indicated ranged from 3% to 18%. This bias was more noticeable when higher coping scores (lower quality of life) were associated with a higher probability of missingness than in other methods (Method 1 and Method 5, Table E2.2; Table E3.2; Table E4.2). The bias was more extreme (> 6%) in the combination of strong positive relationship between quality of life and DFS and weak positive relationship between delayed chemotherapy and DFS (Table E3.2).

With one exception, the performance of LOCF was influenced by the missing data mechanism. For example, consider the combination of strong positive relationship between quality of life and DFS and weak positive relationship between delayed chemotherapy and DFS. Here, i) the bias in the parameter estimates of S_Pacis and delayed chemotherapy was lowest and ii) the probability of finding a relationship between delayed chemotherapy and DFS was highest when lower coping scores (better quality of life) were associated with missingness (Method 2, Table E3.1; Table E3.2). This indicated that the performance of LOCF was better when lower coping scores (better quality of life) were associated with a higher probability of missingness compared to other missing data mechanisms in this setting. The exception was when considering the combination of weak positive relationship between quality of life and DFS and weak positive relationship between delayed chemotherapy and DFS. Here, the performance of LOCF was similar for each of the 5 methods of artificially removing data. As noted, the parameter estimates of S_Pacis and delayed chemotherapy were robust in this

setting. LOCF performed better in this setting than in the other combinations of β_{sp} and β_{del}.

5.4.2 Median Imputation by Patient

Similar to LOCF (section 5.4.1), when considering the weak positive relationship between quality of life and DFS, the parameter estimate for S_Pacis was robust to the missing data mechanism following median imputation by patient. The mean parameter estimate was approximately equal to the theoretical parameter value of 0.1 with a standard error around 0.012 (Figure 5.3; Table E1.1; Table E2.1).

The parameter estimate for S_Pacis following median imputation by patient had a similar pattern to LOCF when considering the strong positive relationship between quality of life and DFS. The bias indicated was around 3% to 9%, with the numerically most extreme bias in the parameter estimate again when later time periods were associated with a higher probability of missingness. The mean parameter estimate was around 0.37-0.39, compared to 0.4, with a mean standard error of around 0.013 (Table E3.1, Table E4.1).

There was little or no reduction in the probability of finding a relationship between delayed chemotherapy and DFS following median imputation by patient. As with the complete simulated datasets (Table 5.1), approximately 27% completed datasets failed to find this relationship when considering the weak positive relationship between delayed chemotherapy and DFS (Table E1.2; Table E3.2). There was a trend towards a small reduction in the probability of finding this relationship when considering the strong positive relationship between delayed chemotherapy and DFS. Here, between 13.3% and 16.7% of the completed simulated datasets failed to find this relationship (Table E2.2; Table E4.2) compared to 12% of the complete simulated datasets (Table 5.1).

The parameter estimate for delayed chemotherapy following median imputation by patient was robust to the missing data mechanism when considering the

combination of weak positive relationship between quality of life and DFS and weak positive relationship between delayed chemotherapy and DFS. In this setting, the parameter estimates, though numerically further away from 0, were similar to the theoretical value and had a bias < 6% (Table E1.2). They can be considered robust, similar to LOCF. The parameter estimate was also approximately equal to the theoretical value when coping scores were missing at random when considering the combination of strong positive relationship between quality of life and DFS and strong positive relationship between delayed chemotherapy and DFS (Method 4, Table E4.2). As with LOCF, for the remaining combinations of β_{sp} and β_{del}, the trend was for the parameter estimate to be closer to 0. This trend was small, corresponding to a bias of between 3.0% and 5.2%, when considering the combination of weak positive relationship between quality of life and DFS and strong positive relationship between delayed chemotherapy and DFS (Table E2.2). Though smaller in magnitude, around 5% to 12%, the bias in the parameter estimate for delayed chemotherapy had a similar pattern to LOCF when considering the strong positive relationship between quality of life and DFS (Table E3.2; Table E4.2).

As with LOCF, the performance of median imputation by patient was influenced by the missing data mechanism, with the exception noted. An example considered the combination of strong positive relationship between quality of life and DFS and weak positive relationship between delayed chemotherapy and DFS (Table E3.1; Table E3.2). In this setting, i) the bias in the parameter estimates of S_Pacis was lowest (3% when the range was around 3% to 8%) and ii) the bias in the parameter estimate for delayed chemotherapy (8.1%) was among the lower values when lower coping scores (higher quality of life) were associated with missingness (Method 2, Table E3.1; Table E3.2). This indicated that the performance of median imputation by patient was again better when lower coping scores (higher quality of life) were associated with a higher probability of missingness compared to other missing data mechanisms in this setting. The performance of median imputation by patient was again better when considering

the combination of weak positive relationship between quality of life and DFS and the weak positive relationship between delayed chemotherapy and DFS than in the other combinations of β_{sp} and β_{del}.

5.4.3 Linear Regression Using Previous Coping Scores

The parameter estimate for S_Pacis was robust to not being able to consider patients with a missing baseline coping score and the missing data mechanism following imputation by linear regression using previous coping scores. For the weak positive relationship between delayed chemotherapy and DFS, the mean parameter estimate was between 0.1009 and 0.1051 compared to 0.1, with a mean standard error around 0.014 (Figure 5.3; Table E1.1; Table E2.1). The corresponding mean parameter estimate was between 0.3910 and 0.4082, with a mean standard error around 0.016, compared to 0.4 for the strong positive relationship between delayed chemotherapy and DFS (Figure 5.3; Table E1.1; Table E2.1).

Not being able to consider the patients with a missing baseline coping score again led to a reduction in the probability of finding a relationship between delayed chemotherapy and DFS. The reduction in this probability following imputation by linear regression using previous coping scores was similar to LOCF (section 5.4.1). When considering the weak positive relationship between delayed chemotherapy and DFS, between 28% and 42% of the completed simulated datasets failed to find a relationship between delayed chemotherapy and DFS (Table E1.2; Table E3.2) compared to approximately 27% (Table 5.1).

The parameter estimate for delayed chemotherapy was again robust to not being able to consider patients with a missing baseline coping score and the missing data mechanism when considering the combination of weak positive relationship between quality of life and DFS and weak positive relationship between delayed chemotherapy and DFS (Table E1.2). Though, as with median imputation by patient, the largest bias of the parameter estimate towards 0 was between 5% and

6% (Method 4 and Method 5, Table E1.2). For the remaining combinations of β_{sp} and β_{del}, the trend was for the parameter estimate to be closer to 0 following imputation by linear regression using previous coping scores. The bias indicated was between 2% and 16%. The parameter estimates followed a similar pattern to LOCF (Table E2.2; Table E3.2; Table E4.2).

The performance of linear regression using previous coping scores was also influenced by the missing data mechanism, with the exception previously noted. The example described considered the combination of strong positive relationship between quality of life and DFS and weak positive relationship between delayed chemotherapy and DFS (Table E3.1; Table E3.2). For similar reasons to LOCF relating to the parameter estimate for delayed chemotherapy, the performance of linear regression using previous coping scores was again better when lower coping scores (higher quality of life) were associated with a higher probability of missingness compared to other missing data mechanisms in this setting. The performance of linear regression using previous coping scores was again better when considering the combination of weak positive relationship between quality of life and DFS and the weak positive relationship between delayed chemotherapy and DFS than in the other combinations of β_{sp} and β_{del}.

5.5 Summary of Applying Simple Imputation Methods to Simulated Datasets

In this chapter, the performance of the standard simple imputation methods when there is a positive relationship between quality of life and DFS and a positive relationship between delayed chemotherapy and DFS was investigated. As noted, the performance of the standard simple imputation methods was better when considering the combination of weak positive relationship between quality of life and DFS and weak positive relationship between delayed chemotherapy and DFS (Table E1.1; Table E1.2) than in the other combinations of β_{sp} and β_{del}.

The standard simple imputation methods involve assumptions described in section 2.2. The investigations considered in this chapter include many scenarios where the assumptions of the standard simple imputation methods were not met. An example is LOCF when higher coping scores were associated with a higher probability of missingness. The standard simple imputation methods performed better when considering the combination of weak positive relationship between quality of life and DFS and weak positive relationship between delayed chemotherapy and DFS than other combinations. This suggests the standard simple imputation methods are less sensitive to the assumptions for the imputation methods in this setting.

A relationship between quality of life and DFS was found in all of the completed simulated datasets (Appendix E). As noted, the parameter estimate for S_Pacis was robust when considering the weak relationship between quality of life and DFS (Table E1.1; Table E2.1). It was also robust following linear regression using previous coping scores when considering the strong relationship between quality of life and DFS (Table E3.1; Table E4.1) The trend was for the parameter estimate to be biased towards 0 following LOCF or median imputation by patient when considering the strong relationship between quality of life and DFS (Table E3.1; Table E4.1). The bias was most extreme when later time periods were associated with a higher probability of missingness (Method 3, Table E3.1; Table E4.1).

Patients who had a missing baseline coping score could not be considered in the time-dependent Cox model analysis following LOCF and linear regression using previous coping scores. This lead to i) a larger mean standard error in the parameter estimates and ii) a lower probability of finding a relationship between delayed chemotherapy and DFS compared to median imputation by patient (Appendix E). Here, the probability of finding a relationship between delayed chemotherapy and DFS was lowest when higher coping scores were associated with a higher probability of missingness (Method 1 and/or Method 5, Table E1.2; Table E2.2; Table E3.2; Table E4.2). This probability was highest when later

time periods were associated with a higher probability of missingness when considering the strong relationship between delayed chemotherapy and DFS (Method 3, Table E2.2; Table E4.2).

As with the parameter estimate for S_Pacis, the parameter estimate for delayed chemotherapy was robust when considering the combination of weak positive relationship between quality of life and DFS and weak positive relationship between delayed chemotherapy and DFS (Table E1.2). For the remaining combinations of β_{sp} and β_{del}, the trend was for the parameter estimate for delayed chemotherapy to be biased towards 0, with the exception noted (section 4.4; Appendix E). The bias in the parameter estimate was most noticeable when higher coping scores were associated with a higher probability of missingness (Method 1 and/or Method 5, Table E2.2; Table E3.2; Table E4.2). The exception to the trend was that when coping scores were missing at random the parameter estimate was approximately equal to the theoretical value following median imputation by patient when considering the combination of strong positive relationship between quality of life and DFS and strong positive relationship between delayed chemotherapy and DFS (Method 4, Table E4.2).

The performance of the standard simple imputation method was influenced by the missing data mechanism except when the combination of weak positive relationship between quality of life and DFS and weak positive relationship between delayed chemotherapy and DFS was considered (Table 5.4 – Table 5.6). However, there was no suggestion that the performance of the standard simple imputation methods was noticeably better when coping scores were missing at random (Method 4) compared to other missing data mechanisms. This is reassuring for the investigation of applying the standard simple imputation methods to the IBCSG dataset in Chapter 3. Among the simulated datasets, the IBCSG dataset most resembles the combination of weak positive relationship between quality of life and DFS and weak positive relationship between delayed chemotherapy and DFS. The results from applying the standard simple imputation

methods to the IBCSG dataset may be more robust, though not necessarily unbiased, than if there was an indication of a strong relationship between quality of life and DFS or a strong relationship between delayed chemotherapy and DFS.

The influence of the missing data mechanism on the performance the standard simple imputation methods in the simulation study illustrates the importance of carefully investigating the missing data mechanism when performing imputation techniques. It also raises the question of the influence of the missing data mechanism on the performance of standard multiple imputation methods. This will be considered in Chapter 6. Unlike simple imputation, the issue of underestimation of the variance of observations is addressed by the standard multiple imputation methods considered in Chapter 6.

Implications of Findings from Applying Simple Imputation Methods

- Simple imputation methods have limitations; the main limitation is the underestimation of the variance of the parameter estimate
- There are only limited circumstances when it is appropriate to draw inferences from the parameter estimate resulting from simple imputation; if the parameter estimates are considered, then justification should be provided
- The simple imputation methods may provide information as part of a sensitivity analysis into the sensitivity of results to the assumptions about the missing data
- The influence of the missing data mechanism on the performance of the standard simple imputation methods in the simulation study illustrates the importance of carefully investigating the missing data mechanism

6 Applying Multiple Imputation Methods to Simulated Datasets with a Positive Relationship Between Quality of Life and Disease-Free Survival and a Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival

6.1 Introduction

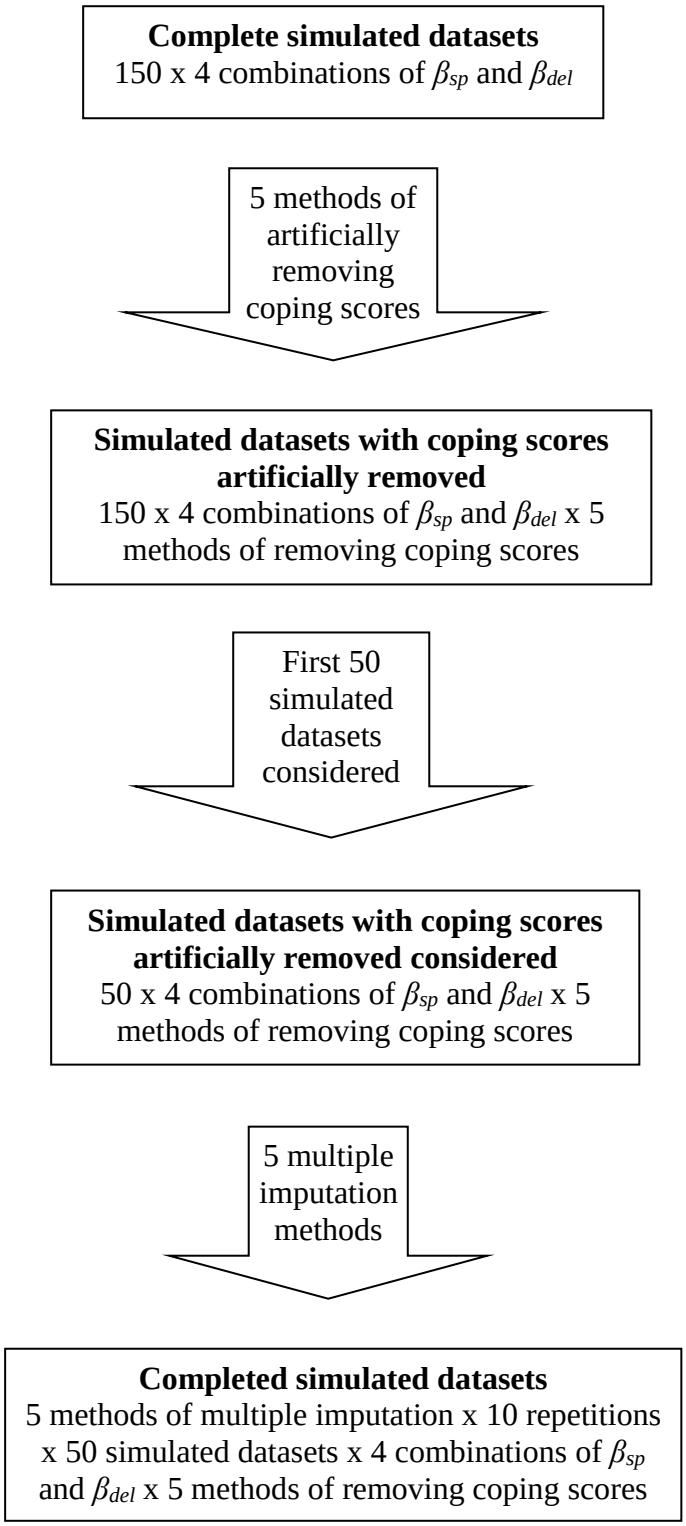
Simulation of datasets took place with the aim of investigating the influence of the missing data mechanism on the performance of standard imputation methods given different combinations of a positive relationship between quality of life and DFS and a positive relationship between delayed chemotherapy and DFS. The influence of the missing data mechanism from the method of artificially removing data on the performance of standard simple imputation methods is described in Chapter 5. This chapter describes applying standard multiple imputation methods to the simulated datasets (see section 5.2). An overview of generating the completed simulated datasets following multiple imputation is given in Figure 6.1.

The standard multiple imputation methods applied were:

i) bootstrapping: subgroups defined by baseline coping score and subgroups defined by previous coping score (see section 4.2.3)
ii) nearest neighbour imputation (see section 4.2.4)
iii) predictive mean matching (see section 4.2.5)
iv) pattern mixture models – Curran's analytical technique (see section 4.2.6)

Similarly to Chapter 5, time-dependent Cox model analysis (see section 3.2.6) was performed following standard multiple imputation methods. The technical details of the patients included in the time-dependent Cox model analysis are described in section 6.2 and the findings from the time-dependent Cox model analysis are described in section 6.3. The summary of the chapter is presented in section 6.4.

Figure 6.1 Overview of Generating Simulated Datasets for Multiple Imputation



6.2 Technical Details of Time-Dependent Cox Model Analysis

Patients Considered in Time-Dependent Cox Model Analysis

As noted in section 4.2.1, when considering the standard multiple imputation methods, the patients with the baseline coping score artificially removed could not be considered in the time-dependent Cox model analysis. A monotone missing data pattern was created by imputing non-monotone missing coping scores by LOCF for the following standard multiple imputation methods:

i) nearest neighbour imputation (see section 4.2.4)

ii) predictive mean matching (see section 4.2.5)

iii) pattern mixture models – Curran's analytical technique (see section 4.2.6)

In contrast, patients with a missing baseline coping score artificially removed can be considered in bootstrapping: subgroups defined by baseline coping score and subgroups defined by previous coping score (see section 4.2.3).

Repetitions of Multiple Imputation

There are 3000 simulated datasets with coping scores artificially removed (150 x 4 combinations of β_{sp} and β_{del} x 5 methods of artificially removing coping scores). This makes performing multiple imputation on each of the simulated datasets or performing a large number of repetitions of multiple imputation impractical. In the context of the multiple imputation in the IBCSG dataset, the efficiency of the estimate was high (94%) after 5 repetitions using bootstrapping, subgroups defined by baseline coping score (Table 4.2). There was more variation between imputed values between simulated datasets with coping scores artificially removed than within repetitions of multiple imputation when estimating the difference between the imputed coping score and the missing coping score in the IBCSG dataset (Table B2.1 and Table B2.2). These were among the considerations in the decision to apply 10 repetitions for 50 simulated datasets with coping scores artificially removed for each scenario in this chapter (Figure 6.1).

6.3 Findings from Applying Multiple Imputation Methods to Simulated Datasets with a Positive Relationship Between Quality of Life and Disease-Free Survival and a Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival

This section describes the findings from the time-dependent Cox model analysis of DFS on S_Pacis and delayed chemotherapy stratified by trial for the 4 combinations of a positive relationship between quality of life and DFS and a positive relationship between delayed chemotherapy and DFS following multiple imputation in simulated datasets. The results are shown in Appendix F and a relationship between quality of life and DFS was found in all the simulated completed datasets (column n (%) of the 50x95% CIs for hazard ratio containing 1, Table F1.1; Table F2.1; Table F3.1; Table F4.1). An overview of the estimate of β_{sp} is shown in Figure 6.2 and a summary of results is provided in Table 6.1-Table 6.4.

Figure 6.2 A (Weak relationship between delayed chemotherapy and DFS)

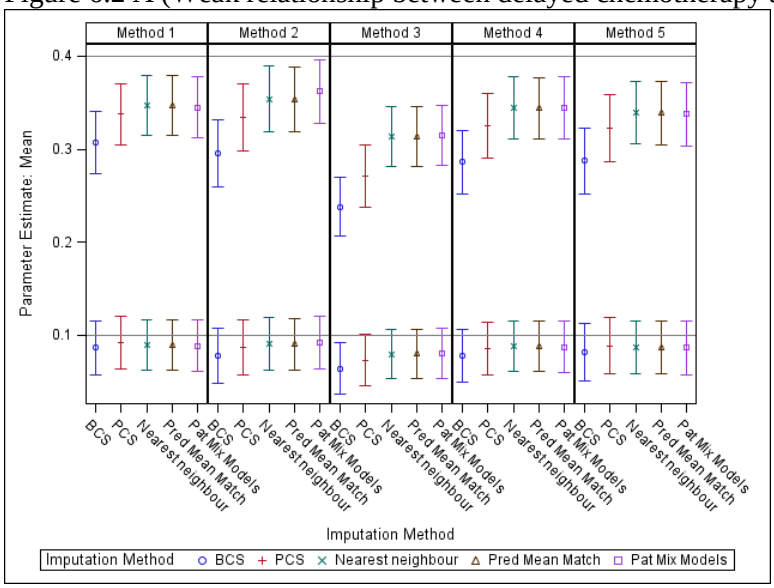
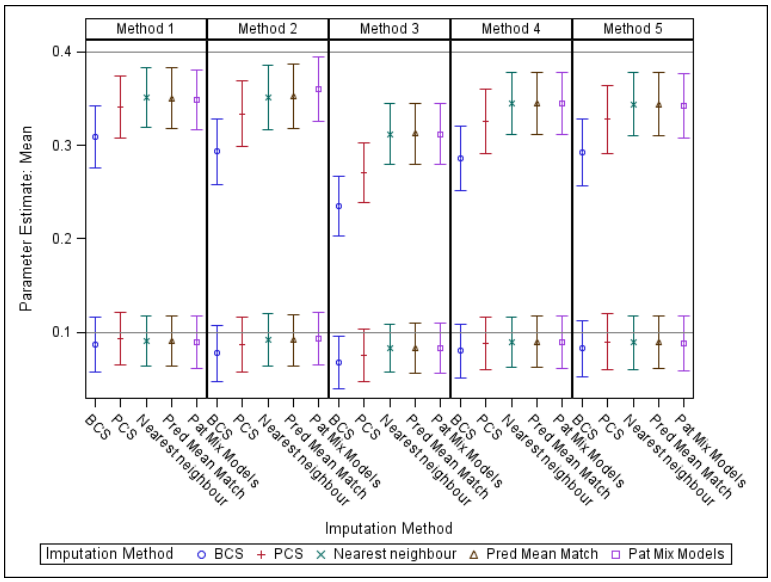


Figure 6.2 B (Strong relationship between delayed chemotherapy and DFS)



Method 1: Higher coping scores have a higher chance of being missing
Method 2: Lower coping scores have a higher chance of being missing
Method 3: Later time period have a higher chance of being missing
Method 4: 30% of coping scores missing at random
Method 5: Higher coping scores have a higher chance of being missing

bcs = bootstrapping, subgroups defined by baseline coping score;
pcs = baseline, subgroups defined by previous coping score;
pat mix models = pattern mixture models – Curran's analytic technique;
pred mean match = predictive mean matching

Figure 6.2 Mean parameter estimate for square root of coping score (S_Pacis) from the time-dependent Cox model analysis and 95% confidence interval based on mean standard error from 10 repetitions of multiple imputation in 50 simulated datasets. The weak and strong relationship between quality of life and disease-free survival is shown in top and lower portion of the figure respectively, with the combination of A) weak relationship between delayed chemotherapy and disease-free survival; (B) strong relationship between delayed chemotherapy and disease-free survival

Table 6.1 Summary of Findings from Applying Bootstrapping, Subgroups Defined by Baseline Coping Score to Simulated Datasets with Positive Relationship Between Quality of Life and Disease-Free Survival and Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival

Combination of β_{sp} and β_{del}	Suggestion of poorest performance	Robustness of Estimate of β_{sp}	Power to Find Significance of β_{del}	Robustness of Estimate of β_{del}
Weak, weak β_{sp}=0.1, β_{del}=-0.165	Later time periods ⇒ missingness	β_{sp} generally underestimated, bias most noticeable when later time periods ⇒ missingness	Lowest when higher coping scores ⇒ missingness according to method 5 Reduction lowest when coping scores missing at random	Trend to overestimate β_{del}, except when higher coping scores ⇒ missingness according to method 1
Weak, strong β_{sp}=0.1, β_{del}=-0.195	Higher coping scores ⇒ missingness according to method 1 or later time periods ⇒ missingness	β_{sp} generally underestimated, bias most noticeable when later time periods ⇒ missingness	Lowest when higher coping scores ⇒ missingness according to method 1	Trend to underestimate β_{del}, most noticeable when higher coping scores ⇒ missingness according to method 1
Strong, weak β_{sp}=0.4, β_{del}=-0.165	Higher coping scores ⇒ according to method 5 or later time periods ⇒ missingness	β_{sp} underestimated, bias most extreme when later time periods ⇒ missingness	Lowest when higher coping scores ⇒ missingness according to method 5 or later time periods ⇒ missingness	β_{del} underestimated, bias most extreme when later time periods ⇒ missingness
Strong, strong β_{sp}=0.4, β_{del}=-0.195	Higher coping scores according to method 5 ⇒ missingness or later time periods ⇒ missingness	β_{sp} underestimated, bias most extreme when later time periods ⇒ missingness	Lowest when higher coping scores ⇒ missingness according to method 5 or later time periods ⇒ missingness	β_{del} underestimated, bias most extreme when later time periods ⇒ missingness

Table 6.2 Summary of Findings from Applying Bootstrapping, Subgroups Defined by Previous Coping Score to Simulated Datasets with Positive Relationship Between Quality of Life and Disease-Free Survival and Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival

Combination of β_{sp} and β_{del}	Suggestion of poorest performance	Robustness of Estimate of β_{sp}	Power to Find Significance of β_{del}	Robustness of Estimate of β_{del}
Weak, weak β_{sp}=0.1, β_{del}=-0.165	Later time periods ⇒ missingness	β_{sp}, generally underestimated, bias most noticeable when later time periods ⇒ missingness	Lowest when higher coping scores ⇒ missingness according to method 5 Reduction lowest when coping scores missing at random	Trend to overestimate β_{del}, except when higher coping scores ⇒ missingness according to method 1
Weak, strong β_{sp}=0.1, β_{del}=-0.195	Higher coping scores ⇒ according to method 1 or later time periods ⇒ missingness	β_{sp}, generally underestimated, bias most noticeable when later time periods ⇒ missingness	Lowest when higher coping scores ⇒ missingness according to method 1	Trend to underestimate β_{del}, most noticeable when higher coping scores ⇒ missingness according to method 1
Strong, weak β_{sp}=0.4, β_{del}=-0.165	Higher coping scores ⇒ according to method 5 or later time periods ⇒ missingness	β_{sp} underestimated, bias most extreme when later time periods ⇒ missingness	Lowest when higher coping scores ⇒ according to method 5 or later periods ⇒ missingness	β_{del} underestimated, bias most extreme when later coping scores ⇒ missingness
Strong, strong β_{sp}=0.4, β_{del}=-0.195	Later time periods ⇒ missingness	β_{sp} underestimated, bias most extreme when later time periods ⇒ missingness	Lowest when later time periods ⇒ missingness	β_{del} generally underestimated, bias most extreme when later coping scores ⇒ missingness

Table 6.3 Summary of Findings from Applying Nearest Neighbour Imputation and Predictive Mean Matching to Simulated Datasets with Positive Relationship Between Quality of Life and Disease-Free Survival and Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival

Combination of β_{sp} and β_{del}	Suggestion of poorest performance	Robustness of Estimate of β_{sp}	Power to Find Significance of β_{del}	Robustness of Estimate of β_{del}
Weak, weak β_{sp}=0.1, β_{del}=-0.165	Later time periods ⇒ missingness	β_{sp}, generally underestimated, bias most noticeable when later time periods ⇒ missingness	Lowest when higher coping scores ⇒ missingness according to method 5 Reduction lowest when coping scores missing at random	Trend to overestimate β_{del}, except when higher coping scores ⇒ missingness according to method 1
Weak, strong β_{sp}=0.1, β_{del}=-0.195	Higher coping scores or later time periods ⇒ missingness	β_{sp}, generally underestimated, bias most noticeable when later time periods ⇒ missingness	Lowest when higher coping scores ⇒ missingness according to method 1	Trend to underestimate β_{del}, most noticeable when higher coping scores ⇒ missingness
Strong, weak β_{sp}=0.4, β_{del}=-0.165	Higher coping scores or later time periods ⇒ missingness	β_{sp} underestimated, bias most extreme when later time periods ⇒ missingness	Lowest when higher coping scores ⇒ missingness	β_{del} underestimated, bias most extreme when higher coping scores ⇒ missingness
Strong, strong β_{sp}=0.4, β_{del}=-0.195	Later time periods ⇒ missingness	β_{sp} underestimated, bias most extreme when later time periods ⇒ missingness	Lowest when higher coping scores or later time periods ⇒ missingness	Trend to underestimate β_{del}, most noticeable when later coping scores ⇒ missingness

Table 6.4 Summary of Findings from Applying Pattern Mixture Models to Simulated Datasets with Positive Relationship Between Quality of Life and Disease-Free Survival and Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival

Combination of β_{sp} and β_{del}	Suggestion of poorest performance	Robustness of Estimate of β_{sp}	Power to Find Significance of β_{del}	Robustness of Estimate of β_{del}
Weak, weak β_{sp}=0.1, β_{del}=-0.165	Later time periods ⇒ missingness	β_{sp}, generally underestimated, bias most noticeable when later time periods ⇒ missingness	Lowest when higher coping scores ⇒ missingness according to method 5 Reduction lowest when coping scores missing at random	Trend to overestimate β_{del}, except when higher coping scores ⇒ missingness according to method 1
Weak, strong β_{sp}=0.1, β_{del}=-0.195	Higher coping scores ⇒ missingnness or later time periods ⇒ missingness	β_{sp}, generally underestimated, bias most noticeable when later time periods ⇒ missingness	Lowest when higher coping scores ⇒ missingness according to method 1	Trend to underestimate β_{del}, most noticeable when higher coping scores ⇒ missingness
Strong, weak β_{sp}=0.4, β_{del}=-0.165	Higher coping scores or later time periods ⇒ missingness	β_{sp} underestimated, bias most extreme when later time periods ⇒ missingness	Lowest when higher coping scores ⇒ missingness	β_{del} underestimated, bias most extreme when higher coping scores ⇒ missingness
Strong, strong β_{sp}=0.4, β_{del}=-0.195	Later time periods ⇒ missingness	β_{sp} underestimated, bias most extreme when later time periods ⇒ missingness	Lowest when higher coping scores or later coping scores ⇒ missingness	Trend to underestimate β_{del}, most noticeable when later coping scores ⇒ missingnes

Similarly to the standard simple imputation methods, the time-dependent Cox model analyses of the completed simulated datasets indicates a lack of precision in the estimates of β_{del} and led to a wide range of the mean parameter estimate based on 10 repetitions of multiple imputation for each of the 50 simulated datasets in each scenario. For example, the range of the mean parameter estimate was (-0.4270 to 0.0865) when considering the combination of weak positive relationship between quality of life and DFS and weak positive relationship between delayed chemotherapy and DFS and when higher coping scores were associated with a higher probability of missingness (first row of Table F1.2). The imprecision is reflected in the fact that the mean standard error of the parameter estimate for delayed chemotherapy was around 0.065 and around 0.068 when considering the weak positive relationship (Table F1.2; Table F2.2) and the strong positive relationship (Table F3.2; Table F4.2) between quality of life and DFS respectively.

In contrast to simple imputation, as previously noted, the standard errors of the parameter estimates reflect the uncertainty in the imputed values. The standard error of the parameter estimate for S_Pacis increased from ~0.011 (Table 5.1) to ~0.015 (Table F1.1; Table F2.1) and from ~0.013 (Table 5.1) to ~0.017 (Table F3.1; Table F4.1) compared to the complete simulated datasets when considering the weak relationship between quality of life and DFS and the strong relationship between quality of life and DFS respectively. The relative increases for the parameter estimate for S_Pacis are similar when considering both the weak and strong relationship between quality of life and DFS. The corresponding values for the parameter estimate for delayed chemotherapy were ~0.065 (Table F1.2; Table F2.2) and ~0.068 (Table F3.2; Table F4.2) compared to ~0.055 (Table 5.1) in the complete simulated datasets respectively.

Overall, the time-dependent Cox model analysis indicated that the 5 standard multiple imputation methods did not perform well in this setting. For example, the bias in the parameter estimate for S_Pacis was at least 6.5% (Table F1.1; Table

F2.1; Table F3.1; Table F4.1). The details on the performance of the individual standard multiple imputation methods are given below.

6.3.1 Bootstrapping, Subgroups Defined by Baseline Coping Score

The parameter estimate for S_Pacis was generally biased towards 0 following bootstrapping, subgroups defined by baseline coping score, with the most extreme bias when later time periods were associated with a higher probability of missingness (Method 3). The bias was large when considering the strong positive relationship between quality of life and DFS, generally of 25%. The mean parameter estimate for S_Pacis was around 0.29-0.31, compared to 0.4, with a mean standard error of ~0.017 (Table F3.1; Table F4.1), with one exception. Almost none of the 95% confidence intervals for the parameter estimate from the 50 completed simulated datasets contained the value of the complete simulated dataset when considering the strong positive relationship between quality of life and DFS (Table F3.1; Table F4.1). The exception was when later time periods were associated with a higher probability of missingness. Here, the bias indicated was around 40% (Method 3, Table F3.1; Table F4.1).

As noted, when considering the strong positive relationship between delayed chemotherapy and DFS, approximately 12% of the complete simulated datasets failed to find a relationship between delayed chemotherapy and DFS (Table 5.1). The probability of finding a relationship between delayed chemotherapy and DFS was noticeably lower following bootstrapping, subgroups defined by baseline coping score, when between 26% and 40% of the completed simulated datasets failed to find this relationship (column n (%) of the 50x95% CIs for hazard ratio containing 1, Table F2.2; Table F4.2). This probability was jointly lowest when there was a strong positive relationship between quality of life and DFS and i) later time periods (Method 3) and ii) higher coping scores were associated with a higher probability of missingness according to method 5 (Table F4.2).

The proportion of complete simulated datasets failing to find a relationship between delayed chemotherapy and DFS was higher when considering the weak positive relationship between delayed chemotherapy and DFS, approximately 27% (Table 5.1). The probability of finding this relationship following bootstrapping, subgroups defined by baseline coping score was reduced when considering the combination of weak positive relationship between quality of life and DFS and weak positive relationship between delayed chemotherapy and DFS. The exception was when coping scores were missing at random (Method 4; Table F1.2). Between 26% and 36% of the completed simulated datasets failed to find this relationship (Table F1.2). The reduction in the probability of finding this relationship was much larger when considering the combination of strong positive relationship between quality of life and DFS and weak positive relationship between delayed chemotherapy and DFS. A high proportion, between 60% and 68%, of the completed simulated datasets failed to find this relationship between delayed chemotherapy and DFS in this case (Table F3.2).

When considering the combination of weak positive relationship between quality of life and DFS and weak positive relationship between delayed chemotherapy and DFS, the general trend was for the parameter estimate for delayed chemotherapy to be further away from 0 following bootstrapping, subgroups defined by baseline coping score. The bias was around 3% - 10% (Table F1.2), with one exception. The exception was when higher coping scores (lower quality of life) were associated with a higher probability of missingness according to method 1, when the parameter estimate was robust (Table F1.2). In contrast, the parameter estimate for delayed chemotherapy was biased towards 0 when considering the combination of weak positive relationship between quality of life and DFS and strong positive relationship between delayed chemotherapy and DFS. Here, the bias was around 6%-14% (Table F2.2). In both combinations, almost all of the 95% confidence intervals for the parameter estimate from the 50 completed simulated datasets contained the value of the complete simulated dataset (Table F1.2; Table F2.2). The bias towards 0 in the parameter estimate for

delayed chemotherapy was higher when considering the strong relationship between quality of life and DFS, at least 35% and around 15%-30% when considering the combination with a weak (Table F3.2) and a strong (Table F4.2) relationship between delayed chemotherapy and DFS respectively. The bias was most extreme when later time periods were associated with a higher probability of missingness (Method 3; Table F3.2 and Table F4.2).

The bias in the parameter estimate for S_Pacis was largest when later time periods were associated with a higher probability of missingness for the 4 combinations of a positive relationship between quality of life and DFS and a positive relationship between delayed chemotherapy and DFS (Method 3, Table F1.1; Table F2.1; Table F3.1; Table F4.1). In addition, the number of 95% confidence intervals for the parameter estimate from the 50 completed simulated datasets that contained the value of the complete simulated dataset was lowest (less than 25%) when considering the weak positive relationship between quality of life and DFS (Method 3, Table F1.1; Table F2.1).

Bootstrapping, subgroups defined by baseline coping score did not perform as well when later time periods were associated with a higher probability of missingness compared to other missing data mechanisms. The probability of finding a relationship between delayed chemotherapy and DFS was lowest and the bias in the parameter estimate for delayed chemotherapy was most extreme for a particular combination when higher coping scores (lower quality of life) were associated with a higher probability of missingness. The particular combination was the combination of weak relationship between quality of life and DFS and strong relationship between delayed chemotherapy and DFS (Method 1, Table F2.2). Therefore, bootstrapping, subgroups defined by baseline coping score also did not perform as well when higher coping scores (lower quality of life) were associated with a higher probability of missingness compared to other missing data mechanisms when considering this combination.

As noted, the bias in the parameter estimate for S_Pacis was large when considering the strong positive relationship between quality of life and DFS. In addition, a high proportion of the completed simulated datasets failed to find a relationship between delayed chemotherapy and DFS. Bootstrapping, subgroups defined by baseline coping score did not perform well in this setting.

6.3.2 Bootstrapping, Subgroups Defined by Previous Coping Score

Though closer to the theoretical parameter value, the parameter estimates of S_Pacis had a similar pattern to bootstrapping, subgroups defined by baseline coping score (section 6.3.1). The most extreme bias in the parameter estimates of S_Pacis was again when later time periods were associated with a higher probability of missingness (Method 3, Table F1.1; Table F2.1; Table F3.1; Table F4.1). The bias remained large when considering the strong positive relationship between quality of life and DFS, generally of 18%. Here, the mean parameter estimate for S_Pacis was around 0.33-0.35, compared to 0.4, with a mean standard error of ~0.017 (Table F3.1; Table F4.1), with the exception noted.

The probability of finding a relationship between delayed chemotherapy and DFS was also always reduced following bootstrapping, subgroups defined by previous coping score, with the exception noted. This probability had a similar pattern to bootstrapping, subgroups defined by baseline coping score, but was generally higher (column n (%) of the 50x95% CIs for hazard ratio containing 1, Table F1.2; Table F2.2; Table F3.2; Table F4.2). The reduction in this probability was again particularly noticeable when considering the combination of strong positive relationship between delayed chemotherapy and DFS and weak positive relationship between delayed chemotherapy and DFS (Table F3.2).

When considering the weak positive relationship between quality of life and DFS, the parameter estimate for delayed chemotherapy also had a similar pattern to bootstrapping, subgroups defined by baseline coping score. The bias was around 5% to 12% further away from 0 in combination with the weak positive

relationship between delayed chemotherapy and DFS (Table F1.2), with the exception noted. Though the bias was of a similar magnitude when considering the combination with the strong positive relationship between delayed chemotherapy and DFS, around 4% to 13%, it was in the opposite direction, towards 0 (Table F2.2). The parameter estimate was closest to the theoretical value when coping scores were missing at random (Method 4, Table F2.2).

6.3.3 Nearest Neighbour and Predictive Mean Matching

The parameter estimates and mean standard errors were similar following both nearest neighbour imputation and predictive mean matching. This is as expected given that i) the imputation methods follow the same initial steps and ii) a monotone missing data pattern was created by imputing non-monotone missing coping scores by LOCF. The parameter estimates of S_Pacis was generally biased towards 0, with the most extreme bias when later time periods were associated with a higher probability of missingness (Method 3, Table F1.1; Table F2.1; Table F3.1; Table F4.1). When considering the strong positive relationship between quality of life and DFS, the parameter estimates were closer to the theoretical parameter value than following bootstrapping (section 6.3.1 and section 6.3.2). However, the bias remained large, generally 14%. The mean parameter estimate for S_Pacis was around 0.34 – 0.35, compared to 0.4, with a mean standard error of ~0.017 (Table F3.1; Table F4.1), with the same exception noted as for bootstrapping.

The probability of finding a relationship between delayed chemotherapy and DFS was also almost always reduced following the nearest neighbour imputation and predictive mean matching. This probability was generally higher than following bootstrapping but had a similar pattern (column n (%) of the 50x95% CIs for hazard ratio containing 1, Table F1.2; Table F2.2; Table F3.2; Table F4.2). The reduction in this probability was again particularly noticeable when considering the combination of strong positive relationship between quality of life and weak positive relationship between delayed chemotherapy and DFS. A high proportion,

between 44% and 64%, of the completed datasets failed to find a relationship between delayed chemotherapy and DFS in this case (Table F3.2).

When considering the weak positive relationship between quality of life and DFS, the parameter estimate for delayed chemotherapy had a similar pattern to bootstrapping. The bias was around 8% to 13% further away from 0 in combination with the weak positive relationship between delayed chemotherapy and DFS (Table F1.2), with the exception noted as for bootstrapping. The parameter estimates were further away from the theoretical parameter value than following bootstrapping. In contrast, when considering the combination with the strong positive relationship between delayed chemotherapy and DFS, the parameter estimates were generally closer to the theoretical parameter value than following bootstrapping. Here, the bias was around 5% to 13% towards 0 (Table F2.2).

6.3.4 Pattern Mixture Models - Curran's Analytical Technique

The parameter estimates and mean standard errors following imputation by pattern mixture models – Curran's analytical technique (Appendix F) were similar to those following both nearest neighbour imputation and predictive mean matching (section 6.3.3). This may be influenced by the fact that in applying these 3 standard imputation methods, a monotone missing data pattern was created by imputing non-monotone missing coping scores by LOCF.

6.4 Summary of Applying Multiple Imputation Methods to Simulated Datasets

The time-dependent Cox model analysis of completed simulated datasets in Chapter 4 suggested that the performance of the standard simple imputation methods was influenced by the missing data mechanism except when considering the combination of weak positive relationship between quality of life and DFS and

weak positive relationship between delayed chemotherapy and DFS. This chapter investigated the performance of the standard multiple imputation methods.

As noted, a relationship between quality of life and DFS was found in all the completed simulated datasets. However, the fact that the 5 standard multiple imputation methods led to i) a biased parameter estimate for S_Pacis closer to 0 than the theoretical parameter value and ii) a reduction in the probability of finding a relationship between delayed chemotherapy and DFS in almost all cases indicates that they did not perform well (section 6.3). The bias of the parameter estimate for S_Pacis was most extreme when later time periods were associated with a higher probability of missingness (Method 3). This most extreme bias was higher following bootstrapping than following the other multiple imputation methods, around 30% to 40% compared to around 17% to 27% (Table F1.1; Table F2.1; Table F3.1; Table F4.1). It was also noticeable that, as with simple imputation, there was imprecision in the parameter estimate for delayed chemotherapy (section 6.3).

The standard multiple imputation methods did not perform well in the context of the simulation study. This is influenced by the fact i) for two of the scenarios the simulated datasets have informative missing data and ii) all the simulated datasets have a general missing data pattern. While none of the standard multiple imputation methods could be recommended for the scenarios considered in the simulation study, there were differences in the bias seen in the parameter estimate of S_Pacis. This bias was smaller following pattern mixture models - Curran's analytical technique than following the other multiple imputation methods when lower coping scores (higher quality of life) was associated with a higher probability of missingness (Method 2).The bias in the parameter estimate for S_Pacis was largest following bootstrapping, subgroups defined by baseline coping score (Figure 6.2). The probability of finding a relationship between delayed chemotherapy and DFS was generally lowest following bootstrapping, subgroups defined by baseline coping score (section 6.3).

The largest influence of the performance of the standard multiple imputation methods was the combination of a positive relationship between quality of life and DFS and a positive relationship between delayed chemotherapy and DFS. The relative bias in the parameter estimates of S_Pacis was lower when there was a weak relationship between quality of life and DFS than when there was a strong relationship between quality of life and DFS (Table F1.1 and Table F2.1 vs Table F3.1 and Table F4.1). The probability of finding a relationship between delayed chemotherapy and DFS was lowest when considering the combination of strong relationship between quality of life and DFS and weak relationship between delayed chemotherapy and DFS (Table F3.2). There was a trend for parameter estimates of delayed chemotherapy to be further away from 0 than the theoretical parameter value when considering the combination of weak relationship between quality of life and DFS and weak relationship between delayed chemotherapy and DFS (Table F1.2), with the exception noted (section 6.3). In contrast, the trend was for the parameter estimates of delayed chemotherapy to be closer to 0 when considering the other combinations (Table F2.2; Table F3.2; Table F4.2).

The performance of the standard multiple imputation methods was influenced by the missing data mechanism. The bias in the parameter estimate for S_Pacis was largest when later time periods were associated with a higher probability of missingness. The parameter estimate for delayed chemotherapy could only be considered robust, in one combination, when higher coping scores (lower quality of life) were associated with a higher probability of missingness according to method 1. The applicable combination was the combination of weak relationship between quality of life and DFS and weak relationship between delayed chemotherapy and DFS (Table F1.2). In the remaining combinations, the bias in the parameter estimate for delayed chemotherapy was lowest when coping scores were missing at random (Method 4, Table F2.2; Table F3.2; Table F4.2). The fact the that the standard multiple imputation methods did not perform well and that the performance of the standard multiple imputation methods was influenced by the missing data mechanism further illustrates the importance of carefully

investigating the missing data mechanism when performing imputation techniques.

The standard errors of the parameter estimates of S_Pacis and delayed chemotherapy following standard multiple imputation methods were larger than from the complete simulated datasets, reflecting the uncertainty in the imputed values (section 6.3). This is the advantage of the standard multiple imputation methods compared to the standard simple imputation methods, where the standard error did not increase. However, bias in the parameter estimate for S_Pacis was more apparent following the standard multiple imputation methods than the standard simple imputation methods. For example, the bias was around 6% to 12% (Table E3.1; Table E4.1) and generally of 25% (Table F3.1; Table F4.1) following imputation by LOCF and by bootstrapping, subgroups defined by baseline coping score respectively when considering the strong positive relationship between quality of life and DFS. The parameter estimate for S_Pacis was robust following the standard simple imputation methods (Table E1.1; Table E2.1) but were generally biased towards 0 (Table F1.1; Table F2.1) following the standard multiple imputation methods when considering the weak positive relationship between quality of life and DFS. Considering the bias in the parameter estimates, there was no indication that the standard multiple imputation methods were more useful than the standard simple imputation methods in the context of the simulated datasets.

As noted, the square root of the coping score in the time-dependent Cox model analysis of the IBCSG dataset and simulated completed datasets in Chapters 3 – 6 was an example of a missing explanatory variable. Missing coping scores were imputed by standard imputation methods before the time-dependent Cox model analysis. The most appropriate way of dealing with missing data may not be the same for a missing explanatory variable as a missing outcome. The influence of a missing outcome variable will be investigated in Chapter 7.

Implications of Findings from Applying Multiple Imputation Methods

- Multiple imputation methods generally assume the data are MAR; pattern mixture models were developed to analyse informative missing data
- A monotone missing data pattern was required to implement pattern mixture models – Curran's analytical technique and the remaining standard multiple imputation techniques apart from bootstrapping
- In the context of the simulation study, the standard multiple imputation methods did not perform well; this is influenced by the fact
    i) for two of the scenarios the simulated datasets have informative missing data
    ii) all the simulated datasets have a general missing data pattern
- The bias of the parameter estimate for S_Pacis was most extreme when later time periods were associated with a higher probability of missingness
- The influence of the missing data mechanism on the performance of the standard multiple imputation methods in the simulation study again illustrates the importance of carefully investigating the missing data mechanism

7 Cardiac Safety in the HERA trial

Trastuzumab treatment had been shown to benefit patients with metastatic breast cancer; however, it is associated with congestive heart failure (CHF) (Slamon et al. 2001; Vogel et al. 2002; Baselga et al. 2005; Marty et al. 2005). Therefore, in the HERA trial, which investigated trastuzumab treatment in early stage breast cancer, only patients with healthy heart function were eligible and cardiac function was monitored in all patients. Symptomatic or asymptomatic CHF was a cardiac endpoint. As part of the cardiac monitoring, the patients' left ventricular ejection fraction (LVEF) was assessed throughout the study and the patients' LVEF assessments. The missing LVEF values give an example of a missing outcome variable in analysis of the change from baseline LVEF over time and repeated measures analysis of the LVEF values over time.

The initial aim of further cardiac analysis described in this chapter was to investigate the influence of missing LVEF assessments on change in LVEF from baseline. Then, to investigate the relationship between a noticeable LVEF drop from baseline and a cardiac endpoint and to model the patients' LVEF values over time. First, the background to the HERA trial is described in section 7.1. The status of LVEF assessments and whether the LVEF assessments were missing at random is described in section 7.2. Multiple imputation was applied to the missing LVEF values in section 7.3 in order to assess the impact of missing LVEF values on the change in LVEF from baseline. The observed and imputed LVEF values were considered in regression models investigating the change in LVEF from baseline at Week 13 as a risk factor for later developing a cardiac endpoint. The occurrence of a noticeable drop in LVEF from baseline was used as a time-dependent covariate in a time-dependent Cox model for time to a cardiac endpoint or competing event of a DFS event in section 7.4. This time-dependent covariate was based on i) observed LVEF values only and ii) observed and imputed LVEF values. Next, repeated measures analysis of LVEF values over time was

performed using a mixed model with baseline information as covariates in section 7.5. The summary of the chapter is presented in section 7.6.

7.1 Description of the HERA Trial

The description of the HERA trial in this section begins with the background (section 7.1.1) and study design (section 7.1.2). The definitions of cardiac endpoints (section 7.1.3) and administration of trastuzumab (section 7.1.4) are described. The published analysis on efficacy and cardiac safety, with statistical analyses performed by the author, is summarised in section 7.1.5. Lastly, the further cardiac analysis described in this chapter is outlined in section 7.1.6.

7.1.1 Background to the HERA Trial

In breast cancer patients, overexpression of the HER2 protein, amplification of the HER2 gene, or both, are associated with aggressive tumours (Slamon et al. 1987; Slamon et al. 1989). The HER2 gene is involved in controlling the growth and survival of cells (Yarden et al. 2001; Gschwind et al. 2004). Trastuzumab is an antibody against HER2 which has been shown to benefit patients with HER2-positive metastatic breast cancer when administered alone (Vogel et al. 2002; Baselga et al. 2005) or in combination of chemotherapy (Slamon et al. 2001; Marty et al. 2005).

Trastuzumab is not associated with the adverse events, such as nausea and vomiting, which often occur during chemotherapy treatment (Bell 2002). However, occasionally hypersensitivity to trastuzmab is seen, generally during or immediately after the first infusion. Trastuzumab treatment is associated with congestive heart failure (CHF) and cardiac dysfunction. Therefore, in the HERA trial cardiac function was prospectively monitored in all patients. The HERA trial investigated whether trastuzumab treatment was effective as adjuvant treatment for HER2-positive breast cancer if used after the completion of the primary treatment.

7.1.2 Study Design

The HERA trial is an international, intergroup, open-label, randomised trial considering women with HER2-positive early-stage breast cancer who completed primary therapy of surgery, radiotherapy if indicated and a minimum of four courses of chemotherapy given preoperatively (neo-adjuvant), postoperatively (adjuvant) or both. The HER2-positive status of the breast cancer tumour was centrally confirmed before randomisation. Patients were randomised between three treatment groups on a 1:1:1 basis; observation only, 1 year of trastuzumab treatment intravenously with one loading dose at 8 milligram/kilogram (mg/kg) then at a dose of 6 mg/kg every three weeks; and 2 years of trastuzumab treatment intravenously also with one loading dose at 8 mg/kg then at a dose of 6 mg/kg every three weeks. A minimisation procedure according to Pocock and Simon (1975) was used with the stratification factors of nodal status, type of chemotherapy, hormone receptor status and intention to use endocrine therapy, age at randomisation and region of the world.

The primary endpoint was DFS, defined as the time from randomisation to the first occurrence of a DFS event: any local, regional or distant recurrence of breast cancer, the development of contralateral breast cancer, including ductal carcinoma in situ but not lobular carcinoma in situ; second non-breast malignant disease other than basal-cell or squamous-cell carcinoma of the skin or carcinoma in situ of the cervix; or death from any cause without documentation of a cancer-related event. Secondary endpoints included overall survival, time to distant recurrence, time to recurrence and cardiac safety. Overall survival was defined as time from randomisation to death due to any cause.

The trial involved the collaboration of 17 Breast International Group (BIG) groups, 9 other cooperative groups, 91 independent sites and the pharmaceutical sponsor, Roche. The institutional review board at each of the 478 sites in 39 countries approved the trial protocol. All patients gave informed consent.

Eligibility Criteria

Eligible patients had histologically confirmed, completely excised invasive breast cancer with HER2 positive status as assessed locally and confirmed centrally. The oestrogen receptor status of the tumour must be known. Eligible patients had node-positive disease or node-negative disease if the pathological tumour size was larger than 1cm. Adjuvant chemotherapy, neo-adjuvant chemotherapy or both from an approved list consisting of at least four cycles was completed before randomisation. Patients with chemotherapy treatment with anthracyclines above a maximum cumulative dose were excluded. Patients given stem-cell support for chemotherapy treatment were excluded. Patients must have acceptable baseline hepatic, renal and bone marrow function.

Adjuvant endocrine therapy, most commonly tamoxifen, was given after chemotherapy to women with hormone receptor positive disease unless contraindicated. During the trial, a protocol amendment allowed aromatase inhibitors to be given as endocrine therapy instead of, or in sequence with, tamoxifen. Patients were required to use contraception, excluding hormone-based methods, if indicated.

Patients were excluded if they had distant metastases, a previous invasive breast carcinoma, or a neoplasm not involving the breast, except for curatively treated basal-cell or squamous-cell carcinoma of the cervix. Patients with clinical stage T4 tumours, including inflammatory breast cancers or involvement of supraclavicular nodes, were excluded. Patients with suspicious internal mammary nodes were excluded unless radiotherapy was given to the internal mammary nodes. Patients with prior mediastinal irradiation (except internal mammary node irradiation for primary breast cancer) were excluded. Patients with chemotherapy treatment with anthracyclines above a maximum cumulative dose were excluded. Patients given stem-cell support for chemotherapy treatment were excluded.

Only patients with a normal baseline LVEF of 55% or more measured by echocardiography or multiple-gated acquisition (MUGA) scan after completion of chemotherapy and radiotherapy were eligible. Patients with a history of documented CHF, coronary heart disease (angina pectoris requiring antianginal medication or transmural infarction on ECG), uncontrolled hypertension (blood pressure systolic > 180 mmHg or diastolic > 100 mmHg), high-risk arrhythmias or clinically significant valvular disease were excluded.

Follow-up Procedures

All patients followed the same schedule of follow-up visits, which required the recording of symptoms, side effects (graded according to the National Cancer Institute Common Toxicity Criteria [NCI-CTC] version 2.0), and findings on clinical examination every three months for the first two years, with haematologic and chemistry studies performed every six months. From year 3 to year 10 after randomisation, these assessments were scheduled to occur annually. Annual chest radiography was required to year 5 and annual mammography to year 10.

Cardiac Monitoring

A cardiac questionnaire, physical examination, ECG and an assessment of LVEF by echocardiography or MUGA scan were performed in all patients at baseline, 3, 6, 9, 12, 18, 24, 30, 36 and 60 months after randomisation. Among the first 900 patients randomised, echocardiograms up to six months (the first three LVEF assessments) were reviewed by a core laboratory blinded to treatment group as a quality control measure and feedback to the site was given where necessary. The results of the review of echocardiograms by the core laboratory were presented to the Independent Data Monitoring Committee (IDMC). Three pre-specified interim cardiac safety analyses were performed after 300, 600 and 900 patients were treated or observed for at least six months. As part of cardiac monitoring, the patients' LVEF assessments were analysed by summarising the change from baseline over time.

7.1.3 Cardiac Definitions

Cardiac safety and tolerability of trastuzumab were assessed on the basis of pre-specified cardiac endpoints, which must take place between randomisation and the start date of new therapy for recurrent disease. Cardiac death was defined as death definitely due to cardiac failure, myocardial infarction or documented arrhythmia, or probable cardiac death within 24 hours of a cardiac event. A significant LVEF drop was defined as an absolute decline of at least 10 percentage points from baseline LVEF and to below a value of 50% identified by MUGA scan or echocardiogram. Severe CHF was defined as New York Heart Association (NYHA) class III or IV, confirmed by a cardiologist and a significant LVEF drop. Symptomatic CHF was defined as symptomatic CHF confirmed by a cardiologist and a significant LVEF drop. Confirmed significant LVEF drop was defined as an asymptomatic (NYHA class I) or mildly symptomatic (NYHA class II) significant LVEF drop, unless the next subsequent LVEF assessment indicated a return to levels that did not meet the definition of a significant LVEF drop; or as identified by the treatment unblinded Cardiac Advisory Board. A repeat LVEF assessment was to be performed approximately 3 weeks after the first documented LVEF drop. The primary cardiac endpoint of the trial was cardiac death or severe CHF. The secondary cardiac endpoint of the trial was confirmed significant LVEF drop.

7.1.4 Administration of Trastuzumab

Trastuzumab was administered intravenously over a 90-minute period at all doses. Patients were closely observed for at least six hours after the start of the first infusion of trastuzumab, a loading dose of 8 mg/kg. Subsequent maintenance doses were 6 mg/kg every three weeks. Trastuzumab was permanently discontinued in patients who experienced severe CHF (a primary cardiac endpoint) and heart failure treatment was recommended. If the patient reached a confirmed significant LVEF drop (a secondary cardiac endpoint) trastuzumab was permanently discontinued. In patients who reached a significant LVEF drop, trastuzumab was temporarily suspended and a repeat LVEF assessment performed

three weeks later. If the repeat LVEF assessment indicated a return to levels that did not meet the criteria for significant LVEF drop, then trastuzumab was resumed.

7.1.5 Published Statistical Analysis

Randomisation of 4482 patients was planned to detect a 23 percent relative reduction in the risk of DFS event with 80% power, with a two-sided significance level of 0.025 for each pairwise comparison of 1 year trastuzumab vs observation and 2 years trastuzumab vs observation. A total of 951 DFS events were required for the final analysis. One interim efficacy analysis was planned after 475 DFS events, with a sequential plan according to the O'Brien-Fleming boundary as implemented by Lan and DeMets (1983). Each of these pair wise comparisons was made according to the method of Holm (1979) so that the overall trial-wide alpha level was 0.05. The significance level for the interim efficacy analysis was 0.002.

The null hypothesis was tested with unstratified log-rank tests (two-sided) following a step-down adjustment procedure of the Bonferroni method as proposed by Holm. In this procedure the testing is conducted in decreasing order of significance. The smallest of the p-values is tested at significance level alpha/2. If the corresponding hypothesis is rejected then the second p-value is be tested at the level of alpha. Efficacy analyses were conducted according to the intent-to-treat principle. Log-rank tests for time-to-event endpoints provide two-sided p-values. Kaplan-Meier curves were presented. Cox proportional hazards model analysis was used to estimate hazard ratios and 95% confidence intervals. A stratified log-rank test was not part of the primary analysis of DFS described in the protocol and not part of the publications.

The interim efficacy analysis after 475 DFS events showed a highly significant improvement of DFS for patients who were randomised to both 1 year trastuzumab and 2 years trastuzumab compared with observation. The IDMC

recommended the release of the 1 year trastuzumab vs observation results. The detailed efficacy results for 1 year trastuzumab vs observation with a median follow-up of 1 year were published (Piccart-Gebhart et al. 2005). Follow-up for the 1 year trastuzumab vs 2 years trastuzumab group continued and details of the 2 years trastuzumab group were not published. Subsequently, manuscripts detailing the cardiac safety (Suter et al. 2007; Procter et al. 2010; de Azambuja et al. 2014) and efficacy results with longer median follow-up and details of the 2 year trastuzumab group (Goldhirsch et al. 2013) were published. At the completion of follow-up, 10 years after the randomisation of the last patient, established benefit of 1 year trastuzumab vs observation was shown and cardiotoxicity remained low (Jackisch et al. 2015).

7.1.6 Further Cardiac Analysis

The clinical cut-off date for the further cardiac analysis described in this chapter was 29th March 2005. Data were available for 3386 patients randomised between December 2001 and March 2005; 1693 were randomised to observation and 1693 were randomised to 1 year trastuzumab. The safety analysis population groups were defined by whether or not a patient received trastuzumab before disease recurrence. There were 19 patients randomised to 1 year trastuzumab who did not receive any trastuzumab before recurrence and 4 patients randomised to observation who received at least one dose of trastuzumab before recurrence. Therefore, there were 1678 patients in the trastuzumab safety analysis population group and 1708 patients in the observation safety analysis population group.

The further cardiac analysis had three aims. These were to i) investigate the influence of missing LVEF assessments on the change in LVEF from baseline (section 6.3) ii) investigate the relationship between a noticeable LVEF drop from baseline and a cardiac endpoint (section 6.4) and iii) to model the five LVEF values from Week 13 up to Week 103/Month 24 (section 6.5). The change from baseline was analysed by i) summarising the change in LVEF from baseline over time and ii) regression models to investigate if, for patients in the trastuzumab

group, the change in LVEF from baseline at Week 13 was related to later development of a cardiac endpoint. The further cardiac analysis was performed in SAS.

When investigating the relationship between the occurrence of an LVEF drop greater than 5 LVEF points from baseline and later development of a cardiac endpoint, the LVEF values throughout the study were considered. This means that a time-dependent Cox model was appropriate. As a cardiac endpoint must occur before the start of new therapy for recurrent disease, the occurrence of a DFS event was a competing risk (see section 1.5). The cause-specific hazards of a cardiac endpoint and the cause-specific hazards of a DFS event from the time-dependent Cox model analysis were calculated. The five LVEF values from Week 13 up to Week 103/Month 24 were modelled by safety analysis population group using a mixed model with the stratification factors and the baseline LVEF and Eastern Cooperative Oncology Group (ECOG) performance score as covariates.

The further cardiac analysis began by investigating if the assumption that the LVEF values are MAR was reasonable (section 7.2). The impact of missing LVEF values on the change in LVEF from baseline was considered by multiple imputation of the missing LVEF values by bootstrapping, subgroups defined by baseline LVEF and safety analysis population group (section 7.3). It was also considered by defining the time-dependent covariate of the occurrence of an LVEF drop greater 5 LVEF points in the time-dependent Cox model analysis based on observed LVEF values and imputed LVEF values (section 7.4). There was no benefit found from imputing LVEF values.

7.2 LVEF Assessments in the HERA Trial

7.2.1 Percentage of Missing LVEF Assessments

The baseline LVEF value was measured at the screening visit. Out of the 3386 patients all except 3 had a baseline LVEF value. There were 31 patients with baseline LVEF of less than 55%, a protocol violation. The status of LVEF assessments by safety analysis population group (see section 7.1.6) is shown below in Table 7.1:

Table 7.1 Status of LVEF Assessments by Visit and Safety Analysis Population Group

	Screening	Week 13	Week 25	Week 52	Week 79/ Month 18	Week 103/ Month 24	Month 30	Month 36
Observed								
Trastuzumab	1677	1586	1454	1033	595	266	65	4
Observation	1706	1503	1383	931	535	251	65	2
All patients	3383	3089	2837	1964	1130	517	130	6
Missing								
Trastuzumab	1	45	64	69	72	66	25	4
Observation	2	109	91	89	68	49	19	6
All patients	3	154	155	158	140	115	44	10
Not reached								
Trastuzumab	0	26	114	490	886	1201	1433	1513
Observation	0	27	114	468	824	1101	1307	1381
All patients	0	53	228	958	1710	2302	2740	2894
Post-recurrence								
Trastuzumab	0	10	25	50	79	90	93	95
Observation	0	22	51	114	159	171	179	179
All patients	0	32	76	164	238	261	272	274
Lost to follow-up								
Trastuzumab	0	11	18	25	29	29	31	31
Observation	0	47	61	85	94	100	100	100
All patients	0	58	79	110	123	129	131	131
Dead								
Trastuzumab	0	0	3	11	17	26	31	31
Observation	0	0	8	21	28	36	38	40
All patients	0	0	11	32	45	62	69	71
Total								
Trastuzumab	1678	1678	1678	1678	1678	1678	1678	1678
Observation	1708	1708	1708	1708	1708	1708	1708	1708
All patients	3386	3386	3386	3386	3386	3386	3386	3386

LVEF = left ventricular ejection fraction

Considering the LVEF assessments that were expected, the percentage of missing LVEF values was low (2.8% [45/1631] in the trastuzumab group and 6.8% [109/1612] in the observation group at Week 13 and 6.3% [69/1102] in the trastuzumab group and 8.8% [89/1020] in the observation group at Week 52; Table 7.1). The percentage of missing LVEF values increased across time, and the missing LVEF values have a general missing data pattern. Among the 779 missing LVEF assessments (Table 7.1), approximately half had an expected date between January and March 2005. For approximately 75% of the missing LVEF assessments with such an expected date, the value of the LVEF assessment was observed in later databases.

As expected, the status of LVEF assessments was not the same for both safety analysis population groups. The number of patients lost to follow-up is higher in the observation group compared to the trastuzumab group. The number of patients with recurrence of disease was higher in the observation group compared to the trastuzumab group (Table 7.1). Therefore the number of observation patients expected to have an LVEF assessment at each visit was lower than the number of trastuzumab patients.

7.2.2 Missing at Random vs Informative Missing Data

It might be expected that the status of LVEF assessments was related to DFS or whether the patient had a severe or life-threatening adverse event. DFS for patients where the LVEF value was expected was compared in each safety analysis population group by the status of LVEF assessment at each visit in Table 7.2. Visits up to Week 79/Month 18 were considered due to the low number of patients at risk of a DFS event beyond the subsequent visit at Week 103/Month 24. The 2-year DFS rate and 95% confidence interval was calculated for these cohorts. The null hypothesis that DFS was the same among patients in the same safety analysis population group with an observed and with a missing LVEF value at a particular visit was tested using a log-rank test.

Table 7.2 Summary of Disease-Free Survival by Safety Analyses Population Group and Status of LVEF Assessment

Trastuzumab						
Visit	Status of LVEF Assessment	Number of Patients	Number of Patients with a DFS Event	Percentage of Patients with a DFS Event	2-year DFS rate (%) (95% CI)	p-Value
Week 13	Observed	1586	116	7.3	86.2 (83.4, 89.1)	0.734
Week 13	Missing	45	1	2.2	95.2 (86.1, 100)	
Week 25	Observed	1454	98	6.7	87.3 (84.4,90.1)	0.737
Week 25	Missing	64	1	1.6	94.1 (82.9, 87.1)	
Week 52	Observed	1033	64	6.2	89.9 (87.1, 92.6)	0.683
Week 52	Missing	69	2	2.9	85.2 (64.3, 100)	
Week 79/ Month 18	Observed	595	28	4.7	94.0 (91.4, 96.5)	0.216
Week 79/ Month 18	Missing	72	3	4.2	85.9 (71.0, 100)	
Observation						
Visit	Status of LVEF Assessment	Number of Patients	Number of Patients with a DFS Event	Percentage of Patients with a DFS Event	2-year DFS rate (%) (95% CI)	p-Value
Week 13	Observed	1503	186	12.4	79.4 (76.3, 82.5)	0.299
Week 13	Missing	109	11	10.1	79.8 (68.0, 91.7)	
Week 25	Observed	1383	157	11.4	81.1 (78.0, 84.3)	0.314
Week 25	Missing	91	3	3.3	89.6 (77.0, 100)	
Week 52	Observed	931	82	8.8	86.6 (83.5, 89.8)	0.247
Week 52	Missing	89	2	2.3	93.9 (85.7, 100)	
Week 79/ Month 18	Observed	535	29	5.4	93.2 (90.5, 96.0)	0.516
Week 79 / Month 18	Missing	68	3	4.4	96.0 (88.3, 100)	

LVEF = left ventricular ejection fraction

Of note, there was a small percentage of patients with a DFS event at each visit. The number of patients with a DFS event was particularly small among patients with a missing LVEF assessment (Table 7.2). The uncertainty in the Kaplan-Meier estimates for the cohorts with a missing LVEF assessment is reflected in the wide 95% confidence interval for the 2-year DFS rate. These 95% confidence intervals included the estimated 2-year DFS rate among the corresponding cohort with an observed LVEF value (Table 7.2). The p-values from the log-rank test

were not significant and were not adjusted for multiple testing (Table 7.2). While the percentage of patients with a DFS event was numerically smaller among patients with a missing LVEF assessment, there was no suggestion that the status of LVEF assessments was related to DFS.

A logistic regression model was used to investigate if the status of LVEF value was related to whether the patient had a severe or life-threatening adverse event for each safety population group. Here, the number of LVEF assessments at Week 103/Month 24 was sufficient for this visit also to be included. Any severe or life-threatening adverse event reported was considered. The results from the logistic regression model analysis are shown in Table 7.3:

Table 7.3 Occurrence of at Least One Severe or Life-Threatening Adverse Event as Risk Factor for Missing LVEF Assessment by Safety Analysis Population Group

Visit	Safety Analysis Population Arm	Status of Risk Factor	Number of Patients	Incidence	Odds Ratio	95% Confidence Interval for Odds Ratio
Week 13	Trastuzumab	At least one severe or life-threatening adverse event	153	6 (3.9%)	1.51	(0.63, 3.62)
Week 13	Trastuzumab	No severe or life-threatening adverse events	1478	39 (2.6%)	Reference	
Week 13	Observation	At least one severe or life-threatening adverse event	88	1 (1.1%)	0.15	(0.02, 1.09)
Week 13	Observation	No severe or life-threatening adverse events	1524	108 (7.1%)	Reference	
Week 25	Trastuzumab	At least one severe or life-threatening adverse event	147	6 (4.1%)	0.96	(0.41, 2.27)
Week 25	Trastuzumab	No severe or life-threatening adverse events	1371	58 (4.2%)	Reference	
Week 25	Observation	At least one severe or life-threatening adverse event	85	1 (1.2%)	0.21	(0.03, 1.56)
Week 25	Observation	No severe or life-threatening adverse events	1389	73 (5.3%)	Reference	

LVEF = left ventricular ejection fraction

Table 7.3 Occurrence of at Least One Severe or Life-Threatening Adverse Event as Risk Factor for Missing LVEF Assessment by Safety Analysis Population Group (continued)

Visit	Safety Analysis Population Arm	Status of Risk Factor	Number of Patients	Incidence	Odds Ratio	95% Confidence Interval for Odds Ratio
Week 52	Trastuzumab	At least one severe or life-threatening adverse event	114	10 (8.8%)	1.51	(0.75, 3.05)
Week 52	Trastuzumab	No severe or life-threatening adverse events	988	59 (6.0%)	Reference	
Week 52	Observation	At least one severe or life-threatening adverse event	64	6 (9.4%)	1.09	(0.46, 2.6)
Week 52	Observation	No severe or life-threatening adverse events	956	83 (8.7%)	Reference	
Week 79/ Month 18	Trastuzumab	At least one severe or life-threatening adverse event	71	7 (9.9%)	0.89	(0.39, 2.03)
Week 79/ Month 18	Trastuzumab	No severe or life-threatening adverse events	596	65 (10.9%)	Reference	
Week 79/ Month 18	Observation	At least one severe or life-threatening adverse event	48	3 (6.3%)	0.5	(0.15, 1.66)
Week 79/ Month 18	Observation	No severe or life-threatening adverse events	555	65 (11.7%)	Reference	
Week 103/ Month 24	Trastuzumab	At least one severe or life-threatening adverse event	32	6 (18.8%)	0.92	(0.36, 2.34)
Week 103/ Month 24	Trastuzumab	No severe or life-threatening adverse events	300	60 (20.0%)	Reference	
Week 103/ Month 24	Observation	At least one severe or life-threatening adverse event	30	6 (20.0%)	1.32	(0.51, 3.42)
Week 103/ Month 24	Observation	No severe or life-threatening adverse events	270	43 (15.9%)	Reference	

LVEF = left ventricular ejection fraction

The number of patients with at least one severe or life-threatening adverse event was small compared to the number of patients with no severe or life-threatening adverse event (Table 7.3). The uncertainty in the odds ratio estimate is reflected in the width of the confidence interval. The 95% confidence interval for odds ratio at each visit contained 1 (Table 7.3).

There was no evidence that the probability of an LVEF value being missing was related to DFS (Table 7.2) or whether the patient had a severe or life-threatening adverse event (Table 7.3). Though not a formal test for informative missing data, this suggests that it is reasonable to assume the LVEF values are MAR. As noted, the number of patients expected to have LVEF assessments at each visit was not the same in each safety analysis population group. Among the patients who were expected to have an LVEF assessment at each visit, the proportion of missing LVEF assessments was not the same for each safety analysis population group (Table 7.1). The probability of a missing LVEF assessment was not independent of the safety analysis population group and it is not reasonable to assume the LVEF values are MCAR.

7.3 Change in LVEF from Baseline

The further cardiac analysis in this section investigated the influence of missing LVEF assessments on the change in LVEF from baseline. The change in LVEF from baseline over time was summarised (section 7.3.2). For patients in the trastuzumab safety analysis population group, regression models were used to investigate if the change in LVEF from baseline at Week 13 was related to later development of a cardiac endpoint (section 7.3.3). Missing LVEF values required to calculate the change from baseline were imputed using multiple imputation. The results of these analyses based on i) observed LVEF values only ii) observed and imputed LVEF values were compared.

7.3.1 Description of Analysis of Change in LVEF from Baseline

Calculating Change in LVEF from Baseline

Change in LVEF from baseline could only be only be calculated when both LVEF assessments were by the same method (i.e. echocardiography or MUGA scan). If the change in LVEF from baseline could not be calculated, the LVEF assessment was set to missing. The 3 patients with a missing baseline LVEF value and the patient with baseline LVEF only reported as a range > 55% could not be considered in the analysis of change in LVEF from baseline. Among the 3 patients with a missing baseline LVEF value, 2 patients were lost to follow-up by Week 13. The remaining patient with a missing baseline LVEF value reached Week 13 but could not be considered in the analysis of change in LVEF from baseline. The Week 13 LVEF value for one patient was incorrectly reported as 2% and was not considered in the analysis of change in LVEF from baseline.

LVEF values reported as unscheduled LVEF assessments were considered when summarising the change in LVEF from baseline over time. For the unscheduled LVEF assessments, a visit window was assigned according to the length of time the patient had been on study at the time of the LVEF assessment.

Decrease from Baseline Greater than 5% and Developing Cardiac Endpoint

Otterstad et al. (1997) investigated sources of variability in echocardiograms. Among the 12 healthy volunteers, the standard deviation of the baseline (first echocardiogram recording) LVEF assessment was 5.2 LVEF points (%) and the mean was 54.3%. In the HERA trial, LVEF $\geq$ 50% was considered healthy heart function in defining the cardiac endpoints. Amongst the considerations in setting the eligibility criteria baseline LVEF $\geq$ 55% was the possibility that patients with an LVEF between 50% and 55% may have been assessed as LVEF less than 50% by a different cardiologist or at a slightly different time. Based on this, it was reasonable to consider that a decrease greater than 5 LVEF points from baseline indicates a genuine decrease in LVEF from baseline LVEF. Decrease from baseline LVEF greater than 5% is considered when investigating change in LVEF

from baseline at Week 13 as a risk factor for later development of a cardiac endpoint (section 7.3.3) and as a time-dependent covariate in a time-dependent Cox model analysis for time to cardiac endpoint (section 7.4).

Technical Details of Log-Binomial Regression and Logistic Regression Models Considering Change in LVEF from Baseline at Week 13 as a Risk Factor for Later Development of a Cardiac Endpoint

Change in LVEF from baseline at Week 13 was considered as a continuous variable in a log-binomial regression model to investigate if the change in LVEF from baseline at Week 13 was related to later development of a cardiac endpoint. This change was then considered as a categorical variable in a logistic regression model. The categories are i) decrease $\geq$ 5% LVEF points and ii) decrease < 5% LVEF points, no change or increase (see paragraph above). This change was considered as a categorical variable as well as a continuous variable given the fact change from baseline is considered categorically in the definition of a significant LVEF drop. Patients with a baseline LVEF of less than 55%, a protocol violation, were not considered. The change from baseline at Week 13 calculated considered LVEF values reported as a Week 13 LVEF assessment. The exception was for 2 patients where no Week 13 LVEF assessment was reported and an unscheduled assessment reported around the time of Week 13 was considered.

Percentage of Patients with Missing LVEF Change from Baseline

The status of the change in LVEF from baseline by safety analysis population group is shown in Table 7.4. Considering the patients expected to have an LVEF assessment, the percentage of patients with a missing change in LVEF from baseline was higher in the observation group than the trastuzumab group for the visits that a large number of patients had reached.

Table 7.4 Status of Change in LVEF from Baseline by Visit and Safety Analysis Population Group

	Observation			Trastuzumab			All patients		
	Observed	Missing	Total	Observed	Missing	Total	Observed	Missing	Total
Week 13	1496 (92.8%)	116 (7.2%)	1612	1584 (97.2%)	45 (2.8%)	1629	3081	160	3241
Week 25	1376 (93.3%)	99 (6.7%)	1475	1453 (95.7%)	65 (4.3%)	1518	2829	164	2993
Week 52	921 (90.1%)	101 (9.9%)	1022	1024 (93.0%)	77 (7.0%)	1101	1945	178	2123
Week 79/Month 18	523 (86.7%)	80 (13.3%)	603	587 (88.0%)	80 (12.0%)	667	1110	160	1270
Week 103/Month 24	244 (81.3%)	56 (18.7%)	300	256 (77.3%)	75 (22.7%)	331	500	131	631
Month 30	65 (76.5%)	20 (23.5%)	85	62 (68.1%)	29 (31.9%)	91	127	49	176
Month 36	2 (25.0%)	6 (75.0%)	8	4 (50.0%)	4 (50.0%)	8	6	10	16

LVEF = left ventricular ejection fraction

Imputing Missing LVEF Assessments

The impact of missing LVEF assessments on the change in LVEF from baseline was investigated by comparing the results following imputation of the missing LVEF values required to calculate the change in LVEF from baseline with the results considering the observed LVEF values. As it is expected that the patients' LVEF values throughout the study are influenced by the baseline LVEF value and the safety analysis population group, the missing LVEF values required to calculate the change in LVEF from baseline were imputed by bootstrapping, subgroups defined by baseline LVEF and safety analysis population group. This means that the baseline LVEF values are considered when imputing missing LVEF assessments in a standard multiple imputation method.

The patients were divided into 10 subgroups defined according to baseline LVEF value and safety analysis population group. The number of patients in each subgroup is shown in Table 7.5. The 31 patients with a baseline LVEF of less than 55% were not considered in the analysis of change in LVEF from baseline following imputation. An LVEF value by the same method as the baseline LVEF is imputed where the change in LVEF from baseline could not be calculated due to different methods of assessment.

Table 7.5 Summary of Subgroups Defined by Baseline LVEF and Safety Analysis Population Group

	Observation	Trastuzumab
Baseline LVEF ≥ 55% and < 60%	379	377
Baseline LVEF ≥ 60% and < 63%	337	333
Baseline LVEF ≥ 63% and < 65%	190	207
Baseline LVEF ≥ 65% and < 70%	474	450
Baseline LVEF ≥ 70%	308	296

LVEF = left ventricular ejection fraction
Note: 31 patients (13 in the trastuzumab group and 18 in the observation arm) excluded due to a protocol violation of baseline LVEF < 55%
1 patient in the trastuzumab group with baseline LVEF reported only as a range > 55% excluded

Parameter Estimate from Log-Binomial Model and Reason for 50 Repetitions of Multiple Imputation

The log-Binomial model for later development of a cardiac endpoint was based on the change in LVEF from baseline at Week 13 as a continuous variable and considered a log link:

$$\log(\pi) = \beta_0 + \beta_1 x \quad (7.1)$$

$$Y \sim \text{Binomial}(1, x) \quad (7.2)$$

where

π is the predicted probability that Y =1 (a later cardiac endpoint occurs)

x is the change from baseline LVEF at Week 13

The log-Binomial model was considered in order to estimate the relative risk and, when considering observed values, the 95% confidence interval directly. As with the parameter estimate for S_Pacis (Table 4.2), the efficiency of the parameter estimate for change in LVEF from baseline at Week 13 was very high after 5 imputations (> 99%). Similarly to section 4.2.6, the efficiency of the parameter estimate and the work of Graham et al. (2007) indicated that there would be no benefit from performing more than 50 repetitions of imputation.

7.3.2 Summary of LVEF Over Time

The summary of change in LVEF from baseline over time for observed LVEF assessments is shown in Table 7.6. The most important category was LVEF of less than 50% and at least 10 EF points from baseline. This category is the intersection of the category decrease $\geq$ 10 and the category LVEF < 50%. As shown in Table 7.6, the percentage of patients with an LVEF of less than 50% and at least 10 EF points from baseline was low for both the trastuzumab and observation group (7.4% in the trastuzumab group and 2.2% in the observation group considering the worst LVEF value).

The mean summary of LVEF over time following imputation by bootstrapping, subgroups defined by baseline LVEF and safety analysis population group is also

shown in Table 7.6. Though the number of patients summarised at each visit has increased, there appeared to be little difference between the percentage of patients in each category from the summary of LVEF over time considering observed LVEF values. There was no suggestion that the percentage of patients who experienced an LVEF of less than 50% and at least 10 EF points from baseline increased. Imputing the missing LVEF values to calculate the change in LVEF from baseline had little influence on the summary of LVEF over time (Table 7.6).

Table 7.6 Summary of LVEF Over Time Based on Observed LVEF Values and Based on Observed and Imputed LVEF Values

	Observation		Trastuzumab	
	Observed Values	Including Imputed Values*	Observed Values	Including Imputed Values*
Week 13				
n	1496	1594	1584	1616
Increase or no change	811 (54.2%)	858 (53.8%)	614 (38.8%)	624 (38.6%)
Decrease < 10	593 (39.6%)	637 (40.0%)	786 (49.6%)	802 (49.6%)
Decrease ≥ 10	92 (6.1%)	99 (6.2%)	184 (11.6%)	190 (11.8%)
LVEF < 50%	18 (1.2%)	19 (1.2%)	65 (4.1%)	64 (4.0%)
LVEF < 50% and decrease ≥ 10	11 (0.7%)	12 (0.8%)	50 (3.2%)	50 (3.1%)
Week 25				
n	1376	1459	1453	1506
Increase or no change	734 (53.3%)	772 (52.9%)	571 (39.3%)	593 (39.4%)
Decrease < 10	535 (38.9%)	572 (39.2%)	686 (47.2%)	708 (47.0%)
Decrease ≥ 10	107 (7.8%)	115 (7.9%)	196 (13.5%)	205 (13.6%)
LVEF < 50%	18 (1.3%)	19 (1.3%)	62 (4.3%)	64 (4.2%)
LVEF < 50% and decrease ≥ 10	14 (1.0%)	15 (1.0%)	48 (3.3%)	50 (3.3%)
Week 52				
n	921	1010	1024	1096
Increase or no change	515 (55.9%)	563 (55.7%)	380 (37.1%)	406 (37.0%)
Decrease < 10	345 (37.5%)	380 (37.6%)	488 (47.7%)	520 (47.4%)
Decrease ≥ 10	61 (6.6%)	68 (6.7%)	156 (15.2%)	170 (15.5%)
LVEF < 50%	15 (1.6%)	16 (1.6%)	40 (3.9%)	42 (3.8%)
LVEF < 50% and decrease ≥ 10	13 (1.4%)	14 (1.4%)	34 (3.3%)	37 (3.4%)

*Mean of 50 repetitions of imputation; LVEF = left ventricular ejection fraction

Table 7.6 Summary of LVEF Over Time Based on Observed LVEF Values and Based on Observed and Imputed LVEF Values (continued)

	Observation		Trastuzumab	
	Observed Values	Including Imputed Values*	Observed Values	Including Imputed Values*
Week 79/Month 18				
n	523	595	587	664
Increase or no change	297 (56.8%)	336 (56.5%)	284 (48.4%)	320 (48.2%)
Decrease < 10	185 (35.4%)	212 (35.6%)	249 (42.4%)	282 (42.5%)
Decrease ≥ 10	41 (7.8%)	47 (7.9%)	54 (9.2%)	63 (9.5%)
LVEF < 50%	4 (0.8%)	5 (0.8%)	14 (2.4%)	16 (2.4%)
LVEF < 50% and decrease ≥ 10	4 (0.8%)	5 (0.8%)	13 (2.2%)	15 (2.3%)
Week 103/Month 24				
n	244	297	256	329
Increase or no change	135 (55.3%)	164 (55.2%)	139 (54.3%)	180 (54.7%)
Decrease < 10	92 (37.7%)	112 (37.7%)	93 (36.3%)	119 (36.2%)
Decrease > 10	17 (7.0%)	21 (7.1%)	24 (9.4%)	29 (8.8%)
LVEF < 50%	4 (1.6%)	5 (1.7%)	3 (1.2%)	4 (1.2%)
LVEF < 50% and decrease > 10	1 (0.4%)	1 (0.3%)	3 (1.2%)	4 (1.2%)
Month 30				
n	65	84	62	90
Increase or no change	34 (52.3%)	46 (54.8%)	37 (59.7%)	54 (60.0%)
Decrease < 10	27 (41.5%)	34 (40.5%)	22 (35.5%)	32 (35.6%)
Decrease ≥ 10	4 (6.2%)	4 (4.8%)	3 (4.8%)	5 (5.6%)
LVEF < 50%	2 (3.1%)	2 (2.4%)	1 (1.6%)	1 (1.1%)
LVEF < 50% and decrease ≥ 10	0 (0.0%)	0 (0.0%)	1 (1.6%)	1 (1.1%)
Overall (worst value)				
n	1544	1594	1600	1616
Increase or no change	516 (33.4%)	482 (30.2%)	323 (20.2%)	292 (18.1%)
Decrease < 10	816 (52.8%)	850 (53.3%)	883 (55.2%)	881 (54.5%)
Decrease ≥ 10	212 (13.7%)	232 (14.6%)	394 (24.6%)	412 (25.5%)
LVEF < 50%	45 (2.9%)	49 (3.1%)	144 (9.0%)	146 (9.0%)
LVEF < 50% and decrease ≥ 10	34 (2.2%)	38 (2.4%)	118 (7.4%)	123 (7.6%)

*Mean of 50 repetitions of imputation; LVEF = left ventricular ejection fraction

7.3.3 Change in LVEF from Baseline at Week 13 as a Risk Factor for Developing a Later Cardiac Endpoint (Trastuzumab Group)

Change in LVEF from baseline at Week 13 as a risk factor for later development of a cardiac endpoint is shown in Table 7.7 based on observed LVEF values and in Table 7.8 based on observed and imputed LVEF values.

Table 7.7 Change in LVEF from Baseline at Week 13 as Risk Factor for Later Development of a Cardiac Endpoint (Observed LVEF Values)

	Number of patients	Incidence	Risk Ratio	95% Confidence Interval for Risk Ratio
Change from baseline at Week 13 (continuous)	1547	34 (2.2%)	0.943	(0.897, 0.992)
Decrease > 5 LVEF points	407	15 (3.7%)	2.211	(1.134, 4.310)
Decrease <= 5 LVEF points, no change or increase	1140	19 (1.7%)	Reference	

LVEF = left ventricular ejection fraction
Note: For change from baseline at Week 13 as a continuous variable, change in risk ratio is per 1 LVEF point increase in the change from baseline at Week 13

Table 7.8 Change in LVEF from baseline at Week 13 as Risk Factor for Later Development of a Cardiac Endpoint
(Missing LVEF Values Required to Calculate Change from Baseline Imputed)

	Mean Number of Patients	Mean Incidence	Mean Risk Ratio	Range of Risk Ratio
Change from baseline at Week 13 (continuous)	1589	34 (2.1%)	0.944	0.943 to 0.948
Decrease > 5 LVEF points	421	15 (3.6%)	2.193	2.148 to 2.226
Decrease <= 5 LVEF points, no change or increase	1168	19 (1.6%)	Reference	

LVEF = left ventricular ejection fraction
Note: For change from baseline at Week 13 as a continuous variable, change in mean risk ratio is per 1 LVEF point increase in the change from baseline at Week 13

The incidence of a later cardiac endpoint was low (2.2%). Each 1 unit increase in change in LVEF from baseline at Week 13 was associated with a reduction in risk of later development of a cardiac endpoint ("Change from baseline at Week 13 (continuous)" row, Table 7.7). A decrease of greater than 5 LVEF points from

baseline at Week 13 was associated with later development of a cardiac endpoint (Table 7.7). This indicates that LVEF is an important measure of cardiac function in patients receiving trastuzumab. Though the number of patients considered in the analysis has increased following multiple imputation, imputing the missing LVEF values required to calculate the change in LVEF from baseline at Week 13 had little influence on the risk ratios (Table 7.8).

7.4 Time-Dependent Cox Model Analysis of Time to Cardiac Endpoint or Disease-Free Survival Event

Time-dependent Cox Model

Patients with a noticeable LVEF drop (> 5%) may not suffer a cardiac endpoint. They may instead have LVEF values which return to baseline values or which, though lower than baseline, do not meet the definition of a significant LVEF drop. Further, patients who suffer a cardiac endpoint may not have a noticeable LVEF drop beforehand. An aim of the further cardiac analysis was to investigate the relationship between the occurrence of a noticeable LVEF drop and a cardiac endpoint. As LVEF was measured throughout the study, a time-dependent Cox model was appropriate. An indicator for the occurrence of an LVEF drop greater than 5 LVEF points from baseline was included in a time-dependent Cox model for time to cardiac endpoint or the competing event of a DFS event. Previous analysis (Suter et al. 2006) found that low baseline LVEF was a risk factor for a cardiac endpoint in the trastuzumab safety analysis population. Therefore, indicators for two categories of baseline LVEF, 55% $\leq$ baseline LVEF < 60% and 60% $\leq$ baseline LVEF < 65%, were included in the model as well as the randomised group. An interaction term for category of baseline LVEF and randomised group was not included in the model described as it did not add information to the main effects. The influence of the category of baseline LVEF was assumed to be the same in both randomised groups. The cause-specific hazards of a cardiac endpoint and the cause-specific hazards of a DFS event from the specified time-dependent Cox model analysis were considered. The time-

dependent Cox model analysis considered i) observed LVEF values and ii) observed and imputed LVEF values.

Defining Time-Dependent Indicator for Occurrence of an LVEF Drop Greater than 5 LVEF Points from Baseline

LVEF assessments were considered where change in LVEF from baseline could be calculated as described in section 7.3.1. As noted, 36 patients (19 randomised to observation; 17 randomised to 1 year trastuzumab) were excluded due to the baseline LVEF value or because of an incorrect LVEF value reported at Week 13. There were also 99 patients (56 randomised to observation; 43 randomised to 1 year trastuzumab) excluded as no post-screening information was available (DFS censored at date of randomisation). In addition, 2 patients randomised to observation were excluded due to an incorrect date of LVEF assessment at Week 13 or Week 25 which was before the date of randomisation. Therefore, 3249 patients were considered in the time-dependent Cox model analysis; 1616 randomised to observation and 1633 randomised to 1 year trastuzumab. In defining an indicator for the occurrence of an LVEF drop greater than 5 LVEF points from baseline, missing LVEF assessments were ignored.

The first occurrence of an LVEF drop greater than 5 LVEF points from baseline for each patient was identified. For patients with no occurrence of an LVEF drop, the time-dependent indicator for occurrence of an LVEF drop was set to 0 from the date of randomisation until the date of cardiac endpoint or DFS event or censoring. For patients with an occurrence of an LVEF drop, the time-dependent indictor for occurrence of an LVEF drop was set to 0 from the date of randomisation to the date of first LVEF drop and set to 1 from the date of first LVEF drop to the date of cardiac endpoint or DFS event or censoring.

7.4.1 Cause-Specific Hazards from Time-Dependent Cox Model Analysis with Competing Risks (Observed LVEF Values)

The status of the competing risks of cardiac endpoint and DFS event by randomised group and occurrence of an LVEF drop greater than 5 LVEF points from baseline is summarised in Table 7.9:

Table 7.9 Summary of Cardiac Endpoint and DFS Event by Randomised Group and Occurrence of an LVEF Drop Greater Than 5 LVEF Points from Baseline (Observed LVEF Values)

	Observation (N=1616)			Trastuzumab (N=1633)			Total
	Cardiac endpoint	DFS event	No event (censored)	Cardiac endpoint	DFS event	No event (censored)	
LVEF drop > 5 LVEF points from baseline	2	47	357	17	53	549	1025
No LVEF drop > 5 LVEF points from baseline	8	166	1036	44	71	899	2224

DFS = disease-free survival; LVEF = left ventricular ejection fraction

The cause-specific hazards for a cardiac endpoint and the competing risk of a DFS event from the time-dependent Cox model analysis are shown in Table 7.10:

Table 7.10 Cause-Specific Hazards for Cardiac Endpoint and DFS Event from Time-Dependent Cox Model Analysis (Observed LVEF Values)

Cardiac Endpoint Parameter	Parameter estimate	Standard error	Hazard ratio	95% CI for hazard ratio
LVEF drop > 5 LVEF points from baseline	0.4807	0.2941	1.6172	(0.909, 2.878)
Randomized group (trastuzumab vs obs.)	1.7247	0.3421	5.6110	(2.869, 10.972)
55% ≤ Baseline LVEF < 60%	1.3192	0.3106	3.7403	(2.035, 6.876)
60% ≤ Baseline LVEF < 65%	0.6918	0.3187	1.9974	(1.070, 3.730)
DFS Event **Parameter**	**Parameter estimate**	**Standard error**	**Hazard ratio**	**95% CI for hazard ratio**
LVEF drop > 5 LVEF points from baseline	0.0705	0.1336	1.0730	(0.826, 1.394)
Randomized group (trastuzumab vs obs.)	-0.6001	0.1139	0.5488	(0.439, 0.686)
55% ≤ Baseline LVEF < 60%	-0.0341	0.1493	0.9665	(0.721, 1.295)
60% ≤ Baseline LVEF < 65%	0.0661	0.1290	1.0683	(0.830, 1.376)

DFS = disease-free survival; LVEF = left ventricular ejection fraction; obs = observation

There was a trend for the occurrence of an LVEF drop greater than 5 LVEF points from baseline to be associated with a cardiac endpoint, though this was not statistically significant. Low baseline LVEF, 55% ≤ baseline LVEF < 60% and 60% ≤ baseline LVEF < 65%, were significantly associated with a cardiac endpoint. This is consistent with the previous finding that low baseline LVEF was a risk factor for a cardiac endpoint. As expected from the efficacy analysis (Piccart-Gebhart et al. 2005), observation rather than trastuzumab treatment was associated with recurrence of disease (a DFS event). There was no association between low baseline LVEF and a DFS event or occurrence of an LVEF drop greater than 5 LVEF points from baseline and a DFS event (Table 7.10).

7.4.2 Cause-Specific Hazards from Time-Dependent Cox Model Analysis with Competing Risks (Observed and Imputed LVEF Values)

The mean status of the competing risks of cardiac endpoint and DFS event by randomised group and occurrence of an LVEF drop greater than 5 LVEF points from baseline following multiple imputation is summarised in Table 7.11.

Table 7.11 Summary of Cardiac Endpoint and DFS Event by Randomised Group and Occurrence of an LVEF Drop Greater Than 5 LVEF Points from Baseline (Missing LVEF Values Required to Calculate Change from Baseline Imputed)

	Observation (N=1616)			Trastuzumab (N=1633)			Total
	Cardiac endpoint	DFS event	No event (censored)	Cardiac endpoint	DFS event	No event (censored)	
LVEF drop > 5 LVEF points from baseline	2	50	381	19	55	570	1077
No LVEF drop > 5 LVEF points from baseline	8	163	1012	42	69	878	2172

DFS = disease-free survival; LVEF = left ventricular ejection fraction

The cause-specific hazards for a cardiac endpoint and the competing risk of a DFS event from the time-dependent Cox model analysis following multiple imputation are shown below in Table 7.12. The derivations in the summary table were as described in section 3.5.1.

Table 7.12 Cause-Specific Hazards for Cardiac Endpoint and DFS Event from Time-Dependent Cox Model Analysis
(Missing LVEF Values Required to Calculate Change from Baseline Imputed)

Cardiac Endpoint Parameter	Mean Parameter estimate	Range	Mean Standard Error	Mean *t* statistic
LVEF drop > 5 LVEF points from baseline	0.6458	(0.607, 0.791)	0.2865	2.26
Randomized group (trastuzumab vs obs.)	1.7122	(1.699, 1.717)	0.3421	5.00
55% ≤ Baseline LVEF < 60%	1.3662	(1.353, 1.408)	0.3109	4.40
60% ≤ Baseline LVEF < 65%	0.7310	(0.721, 0.764)	0.3188	2.29
DFS Event **Parameter**	**Mean Parameter estimate**	**Range**	**Mean Standard Error**	**Mean *t* statistic**
LVEF drop > 5 LVEF points from baseline	0.0968	(0.047, 0.159)	0.1324	0.73
Randomized group (trastuzumab vs obs.)	-0.6026	(-0.609, -0.597)	0.1139	-5.29
55% ≤ Baseline LVEF < 60%	-0.0253	(-0.041, -0.006)	0.1494	-0.17
60% ≤ Baseline LVEF < 65%	0.0735	(0.060, 0.090)	0.1291	0.57

DFS = disease-free survival; LVEF = left ventricular ejection fraction; obs = observation

Following imputation, the number of patients with an occurrence of an LVEF drop greater than 5 LVEF points from baseline increased from 1025 (Table 7.9) to a mean of 1077 (Table 7.11). The parameter for an LVEF drop greater than 5 LVEF points from baseline was statistically significant in the time-dependent Cox model analysis of each of the completed datasets. Following imputation, the association between an LVEF drop greater than 5 LVEF points from baseline and a cardiac endpoint (~0.65) (Table 7.12) was stronger than when only observed LVEF values are considered (0.48) (Table 7.10). This may be influenced by the fact that in the multiple imputation by bootstrapping, subgroups were defined by baseline LVEF, and low baseline LVEF was a risk factor for a cardiac endpoint. The remaining parameter estimates for the cause-specific hazard of cardiac

endpoint were little influenced following multiple imputation. Similarly, the parameter estimates for the cause-specific hazard of DFS event were little influenced following multiple imputation.

The low percentage of missing LVEF assessments and the fact that it is reasonable to assume the missing LVEF assessments are MAR makes the missing data less of a concern than the missing quality of life assessments in the IBCSG dataset. Simple imputation methods will lead to an underestimation of the variance of the LVEF assessments (see section 2.2). The possible influence of the fact baseline LVEF is a risk factor for a cardiac endpoint when applying bootstrapping (subgroups defined by baseline LVEF) in the time-dependent Cox model analysis is noted in the paragraph above. The remaining standard multiple imputation methods require a monotone missing data pattern, which is not the case for LVEF assessments in the HERA trial. Here, there was no benefit in imputation of missing LVEF values. No imputation will be applied in the repeated measures analysis in the next section.

7.5 Repeated Measures Analysis of LVEF Over Time

The five LVEF values from Week 13 up to Week 103/Month 24 were modeled in this section. The repeated measures analysis incorporates correlations for the LVEF values from the same patient. In the mixed model considered, the overall linear trend in LVEF values is defined, among cohorts of patients according to the stratification factors (see section 7.1.2 and Table 7.13) and the baseline ECOG performance status, by the parameter corresponding to the visit. A different intercept is defined for patients according to the safety analysis population group. The interaction term for safety analysis population group and visit makes the slopes different over time for each group. The parameters from this mixed model are estimated and hypothesis tests on the parameter estimates carried out in order to describe the pattern of the LVEF values. The repeated measures analysis can be applied by *proc mixed* in SAS.

7.5.1 Description of the Mixed Model

Henderson (1990) and Searle et al. (1992) describe the historical development of the mixed model. The mixed model is written as:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z\omega} + \boldsymbol{\varepsilon} \tag{6.3}$$

where

$\boldsymbol{X}$ is a matrix of independent coefficients (fixed effects)

$\boldsymbol{\beta}$ is a unknown design matrix

$\boldsymbol{X\beta}$ is the fixed component

$\boldsymbol{Z}$ is a known design matrix

$\boldsymbol{\omega}$ is a vector of unknown random-effects parameters and

$$\boldsymbol{\omega} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{G}) \tag{6.4}$$

$$\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{R}) \tag{6.5}$$

and $\boldsymbol{G}$ and $\boldsymbol{R}$ are unknown variances of a multivariate Normal distribution.

A special case of the mixed model is where there are no random-effects and the term $\boldsymbol{Z\omega}$ disappears. As noted, it is expected that the patients' LVEF values throughout the study are influenced by the baseline LVEF value and the safety analysis population group. In modeling the 5 LVEF values from Week 13 up to Week 103/Month 24, visit is the within-subject factor. The between-subject factor of safety analysis population group (treatment) and the potential interaction between safety analysis population and visit are of most interest.

By considering the time from randomisation of the LVEF assessments as continuous, the regression model for each patient could be considered as a random deviation from some population regression model, with a random intercept and slope from each patient. This allows the effect of time from randomisation on LVEF to differ between patients. However, it does not appear reasonable to use the 5 LVEF assessments at predefined visits up to 2 years from randomisation to model LVEF at time from randomisation as a continuous variable.

It was more appropriate to consider visit as a categorical within-subject factor in a mixed model and consider the between-subject factors as systematic (fixed effects). The 5 LVEF values from Week 13 up to Week 103/Month 24 were modelled using a mixed model that included an interaction term for safety analysis population group and visit. The stratification factors and the baseline LVEF and ECOG performance score were included as covariates. Due to eligibility criteria LVEF $\geq$ 55%, baseline LVEF values have a truncated distribution and baseline LVEF was considered as a categorical variable. As baseline LVEF was included as a covariate, the type of assessment was not considered. The particular mixed model considered for the LVEF values over times is a special case of the mixed model where there are no random-effects.

The repeated measures analysis considered patients where at least one LVEF assessment between Week 13 and Week 103/Month was expected. LVEF values were included in the repeated measure analysis over time regardless of the method of assessment, echocardiography or MUGA scan. Patients with a baseline LVEF of less than 55%, a protocol violation, were not considered. As shown in Table 7.13, the stratification factors and baseline ECOG performance status were balanced by randomised group.

Table 7.13 Summary of Stratification Factors and Baseline ECOG Performance Status by Randomised Group

	Observation N=1693	Trastuzumab N=1693
Nodal Status		
Any Nodal Status, neo-adj chemotherapy	176 (10.4%)	190 (11.2%)
No Positive Nodes, no neo-adj chemotherapy	555 (32.8%)	543 (32.1%)
1-3 Nodes Positive, no neo-adj chemotherapy	490 (28.9%)	483 (28.5%)
>=4 Nodes Positive, no neo-adj chemotherapy	471 (27.8%)	477 (28.2%)
Missing	1 (0.1%)	0 (0.0%)
Adjuvant Chemotherapy Regimen[1]		
No Anthracyclines or Taxanes	99 (5.8%)	97 (5.7%)
Anthracyclines but no Taxanes	1154 (68.2%)	1150 (67.9%)
Anthracyclines + Taxanes	438 (25.9%)	443 (26.2%)
Other (not predefined)	2 (0.1%)	3 (0.2%)
Receptor Status and Endocrine Therapy		
Negative	841 (49.7%)	838 (49.5%)
Positive and no Endocrine Therapy	34 (2.0%)	53 (3.1%)
Positive and Endocrine Therapy	818 (48.3%)	802 (47.4%)
Age group		
< 35 years	126 (7.4%)	126 (7.4%)
35 - 49 years	749 (44.2%)	751 (44.4%)
50 - 59 years	546 (32.3%)	546 (32.3%)
>= 60 years	272 (16.1%)	270 (15.9%)
Region		
Western and Northern Europe, Canada,	1222 (72.2%)	1208 (71.4%)
South Africa, Australia, New Zealand		
Asia Pacific and Japan	202 (11.9%)	202 (11.9%)
Eastern Europe	175 (10.3%)	189 (11.2%)
Central and South America	94 (5.6%)	94 (5.6%)
Baseline ECOG performance status		
0	1542 (91.1%)	1550 (91.6%)
1	149 (8.8%)	143 (8.4%)
Missing	2 (0.1%)	0 (0.0%)

ECOG = Eastern Cooperative Oncology Group

7.5.2 Estimating Parameters and Missing and Unexpected Values in the Mixed Model

Estimating Parameters

Estimation of the parameters in the mixed model is complicated by having unknown parameters $\boldsymbol{\omega}$, $\boldsymbol{G}$ and $\boldsymbol{R}$ as well as unknown design matrixs $\boldsymbol{\beta}$.

The variance of $\boldsymbol{\omega}$ is

$$\boldsymbol{V}=\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}+\boldsymbol{R} \tag{6.6}$$

in the special case where $\boldsymbol{Z}=\boldsymbol{0}$ as there are no random-effects $\boldsymbol{V}=\boldsymbol{R}$.

Restricted or residual maximum likelihood (REML) estimation (Patterson and Thompson 1971; Harville 1977) is a particular form of maximum likelihood estimation which does not use all observations available but uses a likelihood function calculated from a transformed set of observations in order to consider only the parameters of interest.

The approach used in SAS® *proc mixed* is to use maximum likelihood and REML to find a reasonable estimate of $\boldsymbol{G}$ and $\boldsymbol{R}$. Maximum likelihood and REML are valid when missing data is MAR (Rubin 1976; Little 1995). Then an estimate of $\boldsymbol{V}$ can be used in estimated generalised least squares, minimising the expression:

$$(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^{T}\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \tag{7.7}$$

The standard method of estimating $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ is to solve Henderson's mixed model equations (Henderson 1984) and the solutions can be written as:

$$\widehat{\boldsymbol{\beta}}=(\boldsymbol{X}^{\mathrm{T}}\widehat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\widehat{\boldsymbol{V}}^{-1}\boldsymbol{y} \tag{7.8}$$

$$\widehat{\omega}=\widehat{\boldsymbol{G}}^{\mathrm{T}}\boldsymbol{Z}^{T}\widehat{\boldsymbol{V}}^{-1}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}) \tag{7.9}$$

Missing and Unexpected Values

As noted, patients with a baseline LVEF of less than 55%, a protocol violation, were not considered in the repeated measures analysis over time. The 3 patients with a missing baseline LVEF value and the patient with baseline LVEF only reported as a range > 55% could also not be considered. There were 8 patients excluded due to a missing stratification factor or baseline ECOG performance status. Therefore, there were 3202 patients identified in the repeated measures analysis, of which 3111 patients had at least one observed LVEF value between Week 13 and Week 103/Month 24.

The Week 13 LVEF value for one patient was incorrectly reported as 2% and was not considered in the repeated measures analysis, but instead set to missing. The Week 79/Month 18 and Week 103/Month 24 LVEF assessment for one patient were only reported as a range 55% – 65% and were also set to missing.

It is reasonable to assume that the LVEF values are MAR, and therefore the parameter estimates from repeated measures analysis of available LVEF values over time are unbiased. As the proportion of missing LVEF value is low, it is reasonable to assume that the missing LVEF values do not lead to a noticeable loss of power in the repeated measures analysis over time. This indicates that it is appropriate to perform the repeated measures analysis of LVEF values over time using observed LVEF values only.

7.5.3 Structure of Covariance Matrix

The mean LVEF value by visit and safety analysis population group is shown in Table 7.14 and the covariance matrix for LVEF values is shown in Table 7.15.

Table 7.14 Mean LVEF Value by Visit and Safety Analysis Population Group

Safety analysis population group	Visit	n	Mean LVEF	Standard Deviation	Minimum	Maximum
Observation	Week 13	1502	64	6.5	32	97
Observation	Week 25	1383	64	6.7	32	85
Observation	Week 52	931	64	6.7	41	87
Observation	Week 79/ Month 18	535	64	6.5	39	88
Observation	Week 103/ Month 24	251	64	6.4	46	88
Trastuzumab	Week 13	1586	62	7.1	26	91
Trastuzumab	Week 25	1454	62	7.2	20	83
Trastuzumab	Week 52	1033	62	7.1	33	85
Trastuzumab	Week 79/ Month 18	594	63	6.9	34.7	83
Trastuzumab	Week 103/ Month 24	265	64	6.7	45	81.5

LVEF = left ventricular ejection fraction

Table 7.15 Variance/Covariance Matrix of LVEF Values

	Week 13	Week 25	Week 52	Week 79/ Month 18	Week 103/ Month 24
Week 13	47.6	28	24.7	19.4	18.8
Week 25	28	49.3	28.6	23.3	19.8
Week 52	24.7	28.6	49.3	27	21.1
Week 79/Month 18	19.4	23.3	27	44.9	21.9
Week 103/Month 24	18.8	19.8	21.1	21.9	42.5

LVEF = left ventricular ejection fraction

While the covariance between LVEF values decreased as the time between the LVEF values increased, the covariance matrix does not appear to follow the first order autoregressive pattern of decrease by a constant factor (Table 7.15). The value of the fit statistic -2(residual log likelihood) from the mixed model considering an unstructured covariance matrix and a first order autoregressive covariance matrix is shown in Table 7.16.

Table 7.16 Restricted Maximum Likelihood Residuals from the Mixed Model for LVEF Values Up to Week 103/Month 24 by Safety Analysis Group with Stratification Factors, Baseline LVEF, LVEF Method and ECOG Performance Status as Covariates

Type of Covariance Matrix	-2 Residual Log Likelihood	Number of Covariance Parameters	Difference in -2 Residual Log Likelihood	Difference in Number of Covariance Parameters
Unstructured	59435.7	15	204	13
Autoregessive	59639.7	2		

LVEF = left ventricular ejection fraction

The covariance matrix of LVEF values (Table 7.15) and the value of the fit statistic -2(residual log likelihood) from the mixed model (Table 7.16) indicate that it is appropriate to consider an unstructured covariance matrix in the repeated measures analysis. In the mixed model considered, the ***R*** matrix was a block diagonal 3202 blocks, each block consisting of identical 5x5 unstructured matrices.

7.5.4 Solution of Repeated Measures Analysis of LVEF Values

An unstructured covariance matrix was used in the repeated measures analysis of LVEF values over time. The solution to the repeated measures analysis of LVEF values over time is shown in Table 7.17. The test of fixed-effects from the repeated measures analysis of LVEF values over time is shown in Table 7.18.

Table 7.17 Repeated Measures of LVEF Values up to Week 103/Month 24 by Safety Analysis Population Group with Stratifications Factors, Baseline LVEF and ECOG Performance Status as Covariates

	Level of Variable	Safety Analysis Population Group	Visit	Estimate	Standard Error	p-Value
Intercept				60.343	0.572	<0.0001
Safety group		Observation		0		
Safety group		Trastuzumab		-1.909	0.221	<0.0001
Visit			Week 13	0		
Visit			Week 25	-0.049	0.171	0.7735
Visit			Week 52	0.402	0.210	0.0556
Visit			Week 79/Month 18	0.575	0.265	0.0300
Visit			Week 103/Month 24	0.400	0.367	0.2767
Baseline LVEF category	55 - <60%			0		
Baseline LVEF category	60 - <65%			3.006	0.250	<0.0001
Baseline LVEF category	≥65%			6.691	0.236	<0.0001
Nodal status	Any Nodal Status, neo-adj chemotherapy			-0.305	0.346	0.3780
Nodal status	No Positive Nodes, no neo-adj chemotherapy			0		
Nodal status	1-3 Nodes Positive, no neo-adj chemotherapy			0.106	0.236	0.6531
Nodal status	>= 4 Nodes Positive, no neo-adj chemotherapy			-0.322	0.249	0.1955

LVEF = left ventricular ejection fraction; ECOG = Eastern Cooperative Oncology Group; neo-adj = neo-adjuvant

Table 7.17 Repeated Measures of LVEF Values up to Week 103/Month 24 by Safety Analysis Population Group with Stratifications Factors, Baseline LVEF and ECOG Performance Status as Covariates (continued)

	Level of Variable	Safety Analysis Population Group	Visit	Estimate	Standard Error	p-Value
Chemotherapy regimen	No Anthracyclines or Taxaes			0		
Chemotherapy regimen	Anthracyclines but no Taxanes			-0.729	0.389	0.0614
Chemotherapy regimen	Anthracyclines + Taxanes			0.031	0.442	0.9432
Receptor status and endocrine therapy	Negative			0		
Receptor status and endocrine therapy	Positive and no endocrine therapy			1.067	0.597	0.0737
Receptor status and endocrine therapy	Positive and endocrine therapy			-0.342	0.187	0.0668
Age category	< 35 years			0		
Age category	35-49 years			-0.332	0.365	0.3626
Age category	50-59 years			0.241	0.375	0.5210
Age category	>= 60 years			0.296	0.413	0.4734
Region	Western and Northern Europe, Canada, South Africa, Australia, New Zealand			0		
Region	Asia Pacific and Japan			1.107	0.285	0.0001
Region	Eastern Europe			1.362	0.299	<0.0001
Region	Central and South America			1.069	0.398	0.0073

LVEF = left ventricular ejection fraction; ECOG = Eastern Cooperative Oncology Group

Table 7.17 Repeated Measures of LVEF Values up to Week 103/Month 24 by Safety Analysis Population Group with Stratifications Factors, Baseline LVEF and ECOG Performance Status as Covariates (continued)

	Level of Variable	Safety Analysis Population Group	Visit	Estimate	Standard Error	p-Value
ECOG performance status	0			0		
ECOG performance status	1			-0.467	0.320	0.1446
Safety group*Visit		Observation	Week 13	0		
Safety group*Visit		Observation	Week 25	0		
Safety group*Visit		Observation	Week 52	0		
Safety group*Visit		Observation	Week 79/Month 18	0		
Safety group*Visit		Observation	Week 103/Month 24	0		
Safety group*Visit		Trastuzumab	Week 13	0		
Safety group*Visit		Trastuzumab	Week 25	-0.067	0.239	0.7793
Safety group*Visit		Trastuzumab	Week 52	-0.678	0.290	0.0193
Safety group*Visit		Trastuzumab	Week 79/Month 18	1.116	0.365	0.0022
Safety group*Visit		Trastuzumab	Week 103/Month 24	1.894	0.512	0.0002

LVEF = left ventricular ejection fraction; ECOG = Eastern Cooperative Oncology Group

Table 7.17 Repeated Measures of LVEF Values up to Week 103/Month 24 by Safety Analysis Population Group with Stratifications Factors, Baseline LVEF and ECOG Performance Status as Covariates (continued)

ECOG performance status	0			0		
ECOG performance status	1			-0.467	0.320	0.1446
Safety group*Visit		Observation	Week 13	0		
Safety group*Visit		Observation	Week 25	0		
Safety group*Visit		Observation	Week 52	0		
Safety group*Visit		Observation	Week 79/Month 18	0		
Safety group*Visit		Observation	Week 103/Month 24	0		
Safety group*Visit		Trastuzumab	Week 13	0		
Safety group*Visit		Trastuzumab	Week 25	-0.067	0.239	0.7793
Safety group*Visit		Trastuzumab	Week 52	-0.678	0.290	0.0193
Safety group*Visit		Trastuzumab	Week 79/Month 18	1.116	0.365	0.0022
Safety group*Visit		Trastuzumab	Week 103/Month 24	1.894	0.512	0.0002

LVEF = left ventricular ejection fraction; ECOG = Eastern Cooperative Oncology Group

Table 7.18 Test of Fixed-Effects for Repeated Measures of LVEF Assessments up to Week 103/Month 24 by Safety Analysis Population Group with Stratification Factors, Baseline LVEF and ECOG Performance Status and Covariates

	Numerator Degrees of Freedom	Denominator Degrees of Freedom	p-Value (Wald Chi-Square Test)	p-Value (F Test)
Safety group	1	3093	<0.0001	<0.0001
Visit	4	3093	<0.0001	<0.0001
Baseline LVEF category	2	3093	<0.0001	<0.0001
Nodal status	3	3093	0.3025	0.3027
Chemotherapy regimen	2	3093	0.0012	0.0013
Receptor status and endocrine therapy	2	3093	0.0206	0.0207
Age category	3	3093	0.0233	0.0234
Region	3	3093	<0.0001	<0.0001
ECOG performance status	1	3093	0.1445	0.1446
Safety group*Visit	4	3093	<0.0001	<0.0001

LVEF = left ventricular ejection fraction; ECOG = Eastern Cooperative Oncology Group

The interaction term between safety analysis population group and visit was statistically significant because the LVEF values over time for patients in the trastuzumab group did not follow the same pattern as the LVEF values for patients in the observation group (Table 7.18).

The treatment or observation period lasts for 1 year. For trastuzumab patients, the trend was for LVEF values during the treatment period to decrease from the reference visit of Week 13 and increase from the reference visit of Week 13 after the treatment period was completed. LVEF values after the treatment period was completed were statistically significantly higher than the reference visit of Week 13 (Table 7.17). However, the increase in LVEF value at Week 103/Month 24 compared to reference visit of Week 13 was less than 3 LVEF points (Table 7.17) and is not clinically important. This suggests that during the treatment period, trastuzumab treatment was associated with a statistically significant but clinically unimportant decrease in LVEF values which then returned to baseline LVEF values after trastuzumab treatment was completed.

For observation patients, there was no suggestion of a trend in direction of LVEF values during the period from Week 13 to Week 103/Month 24. The LVEF value at Week 103/Month 24 was not statistically significantly different from the reference visit of Week 13. There was no suggestion of a clinically important change in LVEF values during the period from Week 13 to Week 103/Month 24 (Table 7.17).

Patients with a baseline LVEF 60% or more and less than 65% and patients with a baseline LVEF 65% or more had a statistically significant higher LVEF value than patients with a baseline LVEF 55% or more and less than 60%. The mean LVEF value for patients with baseline LVEF 55% or more and less than 60% and patients with a baseline LVEF 60% or more and less than 65% for visits between Week 13 and Week 103/Month 24 remained above the eligibility criteria of 55% and is considered a normal value (Table 7.17). The incidence of cardiac endpoints among patients with a baseline LVEF 55% or more and less than 60% and patients with a baseline LVEF 60% or more and less than 65% was statistically significantly higher than patients with a higher baseline LVEF (Suter et al. 2007). This suggests patients with a LVEF 55% or more and less than 60% and patients with a baseline LVEF 60% or more and less than 65% had clinically poorer cardiac function than patients with baseline LVEF 65% or more.

Amongst the 5 stratification factors, nodal status was not a statistically significant fixed effect (Table 7.18). There was no suggestion of a clinically important difference in LVEF values among any of the categories of these stratification factors (Table 7.17). However, there were some trends or statistically significant differences that were not clinically important. Firstly, there was a trend towards patients treated with anthracyclines but no taxanes having a lower LVEF value than patients treated with no anthracyclines or taxanes and patients treated with anthracycline and taxanes. Secondly, there was a trend towards the LVEF in the receptor status and endocrine therapy categories being in ascending order i) positive and endocrine therapy, ii) negative and iii) positive and endocrine

therapy. Thirdly, patients from Western and Northern Europe, Canada, South Africa, Australia and New Zealand had a statistically significant lower LVEF value than patients from each of the other 3 regions (Table 7.17). Similarly to the trends described for the stratification factors, there was a trend towards patients with ECOG performance status 0 having a higher LVEF than patients with ECOG performance status 1 (Table 7.17).

While randomisation ensured the number of patients in each category of the stratification factors or baseline characteristic was balanced among the treatment groups, the number of patients in each of these categories may not be balanced (Table 7.13). In particular, there was:

i) a small number of patients who did not receive anthracyclines
ii) a small number of patients with hormone receptor positive tumours who did not receive endocrine therapy
iii) a large number of patients aged 35-49 years compared to patients aged less than 35 years and patients aged 60 years or more
iv) a large number of patients from Western and Northern Europe, Canada, South Africa, Australia and New Zealand compared to each of the other 3 regions
v) a large number of patients with ECOG performance status 0 compared to the number of patients with ECOG performance status 1

This may influence the estimates from the repeated measures analysis for the corresponding stratification factor or baseline characteristic.

7.6 Summary

Trastuzumab treatment has established efficacy in early breast cancer but is not appropriate for patients with a history of cardiac disease (Goldhirsch et al. 2013; de Azambuja et al. 2014). In the HERA trial, LVEF assessments were performed throughout the study to assess cardiac safety. Further cardiac analysis was performed with the aim of investigating the influence of missing LVEF assessments on the change in LVEF from baseline and to model the LVEF values over time. It was reasonable to assume the LVEF assessments were MAR and the percentage of missing LVEF values was low (section 7.2).

The percentage of patients with an LVEF of less than 50% and at least 10 LVEF points from baseline throughout the study was low for both the observation and trastuzumab group (Table 7.6). Among trastuzumab patients, a decrease of greater than 5 LVEF points from baseline at Week 13 was associated with later development of a cardiac endpoint. There was little influence on the findings following multiple imputation (section 7.3.3). The occurrence of an LVEF drop greater than 5 LVEF points was considered as a time-dependent covariate in a time-dependent Cox model analysis based on competing risks (section 7.4). The finding that low baseline LVEF was significantly associated with a cardiac endpoint (Table 7.10) was consistent with the previous finding that low baseline LVEF was a risk factor for a cardiac endpoint. Considering imputed LVEF values led to a stronger association between an LVEF drop greater than 5 LVEF points from baseline and a cardiac endpoint (~0.65) (Table 7.12) than when only observed LVEF values are considered (0.48) (Table 7.10). As noted, this may be influenced by the fact that in the multiple imputation by bootstrapping subgroups were defined by baseline LVEF and low baseline LVEF was a risk factor for a cardiac endpoint. In context of the HERA trial, it was not beneficial to impute LVEF values.

It was appropriate to perform the repeated measures analysis of LVEF over time using observed LVEF values only (section 7.5). The five LVEF values from Week 13 up to Week 103/Month 24 were modeled by safety analysis population group using a mixed model with the stratification factors and the baseline LVEF and ECOG performance score as covariates. The LVEF values over time for patients in the trastuzumab group did not follow the same pattern as the LVEF values for patients in the observation group (Table 7.18). For trastuzumab patients, the repeated measures analysis suggests that during the treatment period, trastuzumab treatment was associated with a statistically significant but clinically unimportant decrease in LVEF values which then returned to baseline LVEF value after trastuzumab treatment was completed. For observation patients, there was no suggestion of trend in direction of LVEF values or a clinically important change in the LVEF values during the period between Week 13 and Week 103/Month 24 (Table 7.17).

The repeated measures analysis of LVEF values over time and the fact that the incidence of cardiac endpoints among patients with a baseline LVEF $\underline{\geq}$ 55% and < 60% and patients with a baseline LVEF $\underline{\geq}$ 60% and < 65% was statistically significantly higher than patients with a higher baseline LVEF suggested that patients with a LVEF $\underline{\geq}$ 55% and < 65% had clinically poorer cardiac function than patients with baseline LVEF $\underline{\geq}$ 65%.

The missing LVEF assessments in the further cardiac analysis are an example of a missing outcome variable and an example where the assumption that missing values are MAR was reasonable. Here, there was no benefit in multiple imputation. In contrast, the missing quality of life assessments in the time-dependent Cox model analysis of DFS in the IBCSG dataset are an example of a missing explanatory variable and are informative missing data. Imputation of the missing quality of life scores was appropriate. The differences in the appropriate approaches in addressing the missing values again highlights the importance of carefully considering the missing data mechanism

8 Conclusions

Therapeutic trials in breast cancer are conducted to compare the effectiveness of treatment regimens. The main treatment comparisons in breast cancer clinical trials are generally based on DFS and OS and the time to event endpoints are generally analysed in a Cox proportional hazards model. More recently, clinical trials have been designed with endpoints which include the patient's perception of his or her well-being. Often quality of life assessments are repeated throughout the study and so are suitable as a time-dependent covariate in a time-dependent Cox model analysis of DFS. It is common for medical assessments which monitor patient safety to be repeated throughout the study.

Generally when assessments are repeated throughout the study, some missing observations are expected. The potential problems associated with missing observations such as missing quality of life assessments include bias of the parameter estimates and loss of power to detect clinically important differences among treatment groups over time. Imputation-based procedures, where the missing values are filled-in and the completed data are analysed by standard methods, have been developed in the statistical literature for the analysis of data with missing values with the aim of reducing bias from missing data. A review of articles published in the Journal of Clinical Oncology indicated that between 2006 and 2011 it had become more common to describe missing data and the associated assumptions but that imputation was not widely applied. In particular, imputation in the context of time-dependent covariates that may be informative missing data has not been studied in detail in the statistical literature and is the focus of this thesis.

8.1 Imputation of Missing Observations

Simple imputation involves replacing the missing observation with a single plausible value and there are several standard methods of simple imputation. When appropriately performed, simple imputation allows valid inferences from

standard procedures with modifications to account for the different status of observed and imputed values and allows all patients where the observation was expected to contribute to the analysis. However, inappropriate use of simple imputation methods may increase bias. An important disadvantage of simple imputation methods is that, as they do not reflect the uncertainty in imputed value, they lead to the underestimation of the variance of the observations. There are only limited circumstances when it is appropriate to draw inferences from the parameter estimate resulting from simple imputation. Of note, LOCF cannot always be assumed to be a conservative analysis. However, simple imputation methods can give useful information about the sensitivity of the results to assumptions about missing data. In considering the performance of imputation methods, simple imputation methods such as LOCF have been considered in the statistical literature.

The basic strategy of multiple imputation is to generate K (K > 1) completed datasets in order to analyse each of the completed datasets by standard methods and then combine the results to produce estimates and confidence intervals that take account of the missing-data uncertainty. The number of repetitions of multiple imputation is commonly set to 5. Multiple imputation can avoid the disadvantage of underestimation of the variance of the observations and allow sensitivity analysis of the imputation methods on the results from the analyses. As with simple imputation, it has the advantage of including information from patients with incomplete observations.

Careful consideration should be given to the final choice of imputation method. It is important to remember that if the imputation model does not capture the missing data mechanism, then any analysis based on the imputation is flawed. Most of the methods for multiple imputation assume that the missing observations are missing at random and many methods assume a multivariate normal distribution for the data. Among the considerations when deciding on the final imputation method will be previous experience with similar missing data issues,

the ease of communicating the methods used and the resources required to perform the imputation.

Imputation in quality of life data is expected to be beneficial when additional information, such as intensity of adverse events, that is related to quality of life is available when the quality of life assessment is both observed and missing. This additional patient information can be considered in the imputation process, for example by using MCMC methods. In quality of life data, the few items or scales which have been identified as the most important should be the focus of choosing the imputation method.

The focus of this thesis is the influence of missing data on explanatory variables that may be related to DFS, such as quality of life, and also missing longitudinal assessments which are performed to assess major safety endpoints. The application is to breast cancer clinical trials. Standard imputation methods were reviewed and outlined. Missing values of quality of life were imputed using standard imputation methods before analysis investigating the possible relationship between quality of life and DFS. The performance of standard imputation methods was considered. This was done by generating simulated datasets with a known relationship between quality of life and DFS. Simulated datasets with each of the three types of missing data mechanism (MCAR, MAR and informative missing data) were considered. In the last part of the thesis, the influence of missing values of an outcome variable assessing safety was considered. Repeated measures analysis of a safety assessment was performed.

8.2 Main Findings from Applying Imputation Methods in Breast Cancer Trials

Quality of Life in the IBCSG Dataset

Quality of life was an important endpoint in the adjuvant breast cancer trials IBCSG Trials VI and VII. The main publication on quality of life compared quality of life among treatment groups based on observed values (Hürny et al.

1996). As noted, Coates et al. (2000) indicated that DFS was not significantly predicted by quality of life scores or by changes in quality of life scores from baseline. However, Herring et al. (2004) indicated poor baseline coping score was associated with improved prognosis in Swiss postmenopausal patients.

Previous work investigating the potential relationship between quality of life and DFS was extended in this thesis by considering quality of life as a time-dependent covariate. The quality of life assessments focus on the impact of adjuvant treatment in the early stage of the trial. The high proportion of missing coping scores and the indication from previous statistical analysis that the coping scores are informative missing data (Herring et al. [2004]) together lead to the conclusion that imputation is appropriate for the IBCSG dataset. The question of interest is whether quality of life is related to prognosis in breast cancer patients. The illustration of the standard imputation methods in a small example dataset indicated that using several standard imputation methods may provide useful information in this investigation. Standard imputation methods were used before analysis of DFS, in order to use the coping score as a covariate in a time-dependent Cox model for DFS considering the treatment effect of delayed chemotherapy. Here, the coping score was an example of a missing explanatory variable.

In the IBCSG dataset there was no evidence of a statistically significant or clinically important relationship between quality of life and DFS from the standard imputation methods. The multiple imputation methods showed hazard ratios which were similar for each repetition. There was a small increase in the standard error of the parameter estimates following the standard multiple imputation methods compared to the analysis of all available coping scores. The parameter estimate for the square root of the coping score (S_Pacis) was close to 0 following imputation by all of the standard imputation methods. Similarly, the parameter estimates for S_Pacis from the time-dependent Cox model analyses without imputation carried out for reference and illustrative purposes were also

close to 0. The trend towards a positive relationship between delayed chemotherapy and DFS, with the parameter estimate around -0.1, is consistent with the finding from the main efficacy analysis that there may be a therapeutic benefit from delayed chemotherapy.

Considering coping scores throughout the study in a time-dependent Cox model led to parameter estimates in the opposite direction and of a smaller magnitude than when considering baseline quality of life in Herring et al. (2004). There are differences of note in the analyses in Chapters 3 and 4 compared to Herring et al. (2004). Firstly, in Chapters 3 and 4 premenopausal patients and postmenopausal patients outside Switzerland were considered, giving a broader and larger population of patients. Secondly, the outcome of relapse-free survival considered by Herring et al. (2004) did not include second primary cancer or death without prior event as events. Lastly, further covariates such as age and interaction terms were considered in the analysis by Herring et al. (2004), whereas the time-dependent Cox model in the main investigation in Chapters 3 and 4 was parsimonious.

There was large within and between patient variability in coping scores in the IBCSG dataset. Simulated datasets based on the patients with a complete set of coping scores were used to estimate the difference between the imputed coping score and the missing coping score. For all the standard imputation methods considered, the estimated variance of this difference was high, indicating a lack of accuracy when imputing the missing coping score. In the setting of the IBCSG dataset, the more complex standard multiple imputation methods did not perform better than the standard simple imputation methods. There were similarities between the performance of the imputation methods in the IBCSG dataset in imputing quality of life scores and the performance of imputation methods in the literature. Ramstam et al. (2012) also noted that the alternative imputation methods produced similar results. Peyre et al. (2011) noted a small bias following imputation by personal mean score. This is similar to the finding in the IBCSG

dataset that the coping score may be systematically underestimated following median imputation by patient. However, unlike the imputation in the IBCSG dataset, Peyre et al. (2011) found that multiple imputation improved accuracy and precision compared to personal mean score.

The imputed values form the standard simple and multiple imputation methods in the IBCSG dataset reflected the imputation method applied. There was no suggestion that there is a completed dataset which would be achieved regardless of the imputation method applied. Thus, it is possible that the performance of the imputation methods in this setting was influenced by the relationship between quality of life and DFS. Further, the performance of the standard imputation methods may not be the same when the data are informative missing data compared to when the MAR assumptions is reasonable or when there is a strong relationship between quality of life and DFS. Simulated datasets were generated in order to investigate if the performance of the standard imputation methods given different missing data mechanisms is influenced by the relationship between quality of life and DFS. The IBCSG dataset was the basis of these simulated datasets.

Simulation of Data with a Positive Relationship Between Quality of Life and DFS and a Positive Relationship Between Delayed Chemotherapy and DFS

Complete simulated datasets were generated with four different combinations of a positive relationship between quality of life and DFS and a positive relationship between delayed chemotherapy and DFS. From these complete simulated datasets, simulated datasets containing missing data were generated by artificially removing coping scores. Standard imputation methods were applied to the simulated datasets in order to investigate the influence of the missing data mechanism on the performance of standard imputation methods given different combinations of a positive relationship between quality of life and DFS and a positive relationship between delayed chemotherapy and DFS.

There was a suggestion that the performance of the standard simple imputation methods was influenced by the missing data mechanism except when the combination of weak positive relationship between quality of life and DFS and weak positive relationship between delayed chemotherapy and DFS was considered. The performance of the standard simple imputation methods was poorest when lower quality of life was associated with a higher probability of missingness, a likely scenario for missing quality of life assessments.

The standard simple imputation methods may be less sensitive to the assumptions for the method in the combination of weak positive relationship between quality of life and DFS and weak positive relationship between delayed chemotherapy and DFS than in other settings. This suggests the results from applying the standard simple imputation methods to the IBCSG dataset, may be more robust, though not necessarily unbiased, than if there was an indication of a strong relationship between quality of life or delayed chemotherapy and DFS.

With the one exception for a particular setting noted, the standard multiple imputation methods led to a trend for the parameter estimates of S_Pacis and delayed chemotherapy to be biased towards 0. The bias towards 0 of the parameter estimate for S_Pacis was most extreme when later time periods were associated with a higher probability of missingness. Unlike in simple imputation, the uncertainty in the imputed values was reflected in the fact that the resulting standard errors of the parameter estimates of S_Pacis and delayed chemotherapy were larger than those from the complete simulated datasets. Countering this was the fact that, in this setting, the bias in the parameter estimate for S_Pacis was more apparent following standard multiple imputation methods than standard simple imputation methods. Here, the standard multiple imputation methods were not more useful in achieving unbiased parameter estimates than the standard simple imputation methods. The investigation of applying the standard imputation methods to the simulated datasets illustrates the importance of carefully

investigating the missing data mechanism when performing imputation techniques.

8.3 Findings on Cardiac Safety in the HERA Trial

While trastuzumab treatment has established efficacy in early breast cancer, it is not appropriate for patients with a history of cardiac disease. LVEF assessments were performed throughout the study to assess cardiac safety. Publications on cardiac safety considered observed LVEF values and indicated that the incidence of cardiac endpoints was low. In the trastuzumab group, low baseline LVEF was a risk factor for a cardiac endpoint, which mostly occurred during the trastuzumab period and is generally reversible (Suter al. 2007; Procter et al. 2014).

Further cardiac analysis was performed with the aim of investigating the influence of missing LVEF assessments on the change in LVEF from baseline and to model the LVEF values over time. LVEF was an important safety variable in the HERA trial and considering the LVEF assessments that were expected, the percentage of missing LVEF values was low. This is in contrast to the high percentage of missing quality of life assessments in the IBCSG dataset, where the quality of life assessments were likely considered of less importance during the study. The missing LVEF values give an example of a missing outcome variable and it was reasonable to assume the LVEF values are MAR. In the context of the HERA trial, it was not beneficial to impute LVEF values.

The percentage of patients with an LVEF of less than 50% and at least 10 EF points from baseline throughout the study was low for both the observation and trastuzumab group. Among trastuzumab patients, a decrease of more than 5 LVEF points from baseline at Week 13 was associated with later development of a cardiac endpoint, indicating that LVEF is an important measure of cardiac function in patients receiving trastuzumab. The further cardiac analysis suggested that patients with a LVEF 55% or more and less than 65% had clinically poorer cardiac function than patients with baseline LVEF 65% or more.

The occurrence of an LVEF drop greater than 5 LVEF points from baseline was considered as a time-dependent covariate, similarly to coping score. In line with the previous finding, low baseline LVEF was significantly associated with a cardiac endpoint following multiple imputation. Considering imputed LVEF values led to a stronger association between an LVEF drop greater than 5 LVEF points from baseline and a cardiac endpoint than when only observed LVEF values are considered. The stronger association may be influenced by the fact that in the multiple imputation by bootstrapping, subgroups were defined by baseline LVEF and low baseline LVEF was a risk factor for a cardiac endpoint.

In the HERA trial, it was appropriate to perform the repeated measures analysis of the LVEF values over time in a mixed model using observed LVEF values only. The 5 LVEF values from Week 13 up to Week 103/Month 24 were modelled using a mixed model that included an interaction term for safety analysis population group and visit. Repeated measures analysis suggests that during the treatment period, trastuzumab treatment was associated with a statistically significant but clinically unimportant decrease in LVEF values which then returned to baseline LVEF values after trastuzumab treatment was completed. In contrast, for observation patients, there was no suggestion of a trend in direction of LVEF values or clinically important change in LVEF values up to Week 103/Month 24. The pattern in LVEF values for trastuzumab patients suggested by repeated measures analysis was consistent with the findings from the previous publications that the majority of cardiac endpoints occur during the scheduled treatment period and may be reversible.

LVEF was an important safety variable with a low percentage of missing values. Loss of statistical power from missing safety variables such as LVEF values is less likely to be a concern than when considering quality of life assessments, which may have a high proportion of missing values. The missing LVEF assessments in the further cardiac analysis are an example of where the assumption that missing values are MAR was reasonable and in this setting there

was no benefit in multiple imputation. In contrast, the missing quality of life assessments in the time-dependent Cox model analysis of DFS in the IBCSG trial are an example of informative missing data and imputation of the missing quality of life scores was appropriate. The differences in the appropriate approaches in addressing the missing values again highlights the importance of carefully the missing data mechanism.

8.4 Further Work and Limitations

Observed and imputed coping scores could be used to compare quality of life among treatment groups, repeating the analysis in Hürny et al. (1996). This was not done as the focus of this thesis is investigating the potential association between quality of life and DFS. The work in this thesis focused on standard imputation methods. It could be extended by applying further multiple imputation methods to the missing coping scores in the IBCSG dataset and simulated datasets. In particular, chained equations (section 2.3.7), implemented in the statistical software MICE, has become of more common since work began on this thesis. Of note, the statistical literature on correctly specifying the imputation model continues to be developed (e.g. White et al. 2011). Though chained equations were not considered in this thesis, they are related to multiple imputation methods that were applied: multiple imputation using explicit univariate (section 2.3.4), nearest neighbour imputation and predictive mean matching (section 2.3.5).

Qi et al. (2010) compared multiple imputation by chained equations and fully augmented weighted (FAW) equations (section 1.6.2). The comparison indicated that FAW equations show potential to be a useful tool in survival analysis with missing covariates. It would be of interest to apply FAW equations to the missing coping scores in the IBCSG dataset and simulated datasets.

Jolani et al. (2014) proposed a dual imputation model (DIM) for incomplete longitudinal data. This method integrates multiple imputation and doubly robust

weighing-based methods to protect against misspecifications of the imputation model under a MAR assumption. A key feature of the DIM strategy is to iteratively estimate propensities for each incomplete variable conditional on the other variables, and to impute missing values on that variable by including a function of propensities into the imputation model. The proposed method works well with an intermittent pattern of missingness, as in the IBCSG dataset.

Morris et al. (2014) proposed imputation by local residual draws (LRD). LRD was developed from predictive mean matching, but in LRD the donar's residual is borrowed during imputation rather than an observed value. They compared the performance of predictive mean matching and LRD to fully parametric imputation in simulation studies. Morris et al. (2014) note that that predictive mean matching and LRD may have a role for imputing covariates which i) which are not strongly associated with outcome; and ii) where the imputation model is thought to be slightly but not grossly misspecified. It would be of interest to investigate specification of the imputation model in the IBCSG dataset and whether LRD could be of use.

He et al. (2011) proposed a functional multiple imputation approach. Here, longitudinal response profiles are modeled as smooth curves of time under a functional mixed effects model. They developed a Gibbs sampling algorithm to draw model parameters and imputations for missing values. A simulation study demonstrated that this approach performs well under varying modelling assumptions on the time trajectory and missingness patterns.

The time-dependent Cox model analysis in this thesis assumes proportional hazards. Song and Wang (2013) investigated this proportional hazards assumption and proposed inverse selection probability weighted estimators based on the local partial likelihood approach. These estimators could be used to investigate the sensitivity of the findings of the time-dependent Cox model analysis of quality of life and DFS and cardiac safety to the assumption of proportional hazards.

A further method of obtaining the parameter estimates from the time-dependent Cox model that is of interest is by joint modeling of event time and informative missing values of a time-dependent Covariate (Dupuy and Mesbah 2002). Here, a semi-parametric maximum likelihood estimator of the parameter from the time-dependent Cox model was obtained. An estimator for the variance of these parameter estimates was proposed by Dupuy and Mesbah (2004) to allow hypothesis testing. It would also be of interest to consider the parameter estimate from the selection model proposed by Bradshaw et al. (2010). The selection model is defined by the joint distribution of the event times, missing covariates and the missing data mechanism. It extends work by Herring et al. (2004) to allow time-dependent covariates, such as quality of life in the IBCSG dataset.

Though there was no evidence in the IBCSG dataset of a statistically significant or clinically important relationship between quality of life and DFS, the topic remains of scientific interest in breast cancer clinical trials and clinical trials more widely. It would be of interest to investigate if the findings on the performance of the standard imputation methods in Chapters 4 and 5 are replicated in a simulation study based on quality of life in another clinical trial. Such a simulation study could take place based on a clinical trial from a different application from breast cancer. The two findings that would be of particular interest to investigate the replicability of are:

i) the standard multiple imputation methods were not more useful than the standard simple imputation methods in terms of achieving an unbiased parameter estimate
ii) the bias of the parameter estimate relating to quality of life was largest when later time periods were associated with a higher probability of missingness following the standard multiple imputation methods

A limitation of the further cardiac analysis in the HERA trial is the short median follow-up time of patients (1 year). It would be of interest to consider the time-dependent Cox model analysis of time to cardiac endpoint and the repeated

measures analysis of LVEF over time from a database with longer follow-up. This would incorporate a larger number of cardiac endpoints and a larger number of LVEF assessments beyond the scheduled treatment/observation period. The LVEF values from assessments further from randomisation could then also be included in the repeated measures analysis.

8.5 Summary

The most important concern arising from missing data in clinical trials is the potential to introduce bias to the findings from the clinical trial. The simplest approach to the analysis of data with missing observations is to disregard incomplete patients and only analyse patients with complete data. Though this may be satisfactory when there are only small amounts of missing data, it may lead to misleading results and lacks efficiency.

The aim of the imputation-based procedures for the analysis of data with missing observations which have been developed in the statistical literature was reducing bias from missing data. The review of standard imputation methods and applying standard imputation methods to the IBCSG dataset and the simulation study found:

- There are only limited circumstances when it is appropriate to draw inferences from the parameter estimate resulting from simple imputation. Justification should be provided if the parameter estimates are considered.
- The simple imputation methods may provide information as part of a sensitivity analysis into the sensitivity of results to the assumptions about the missing data
- Multiple imputation methods generally assume the data are MAR; several multiple imputation methods assume a monotone missing data pattern
- The standard multiple imputation methods did not perform well in the simulation study; this is influenced by the fact

i) for two of the scenarios the simulated datasets have informative missing data
ii) all the simulated datasets have a general missing data pattern

- The influence of the missing data mechanism on the performance of the standard imputation methods in the simulation study illustrates the importance of carefully investigating the missing data mechanism

Imputation may not be the most appropriate method of analysing missing data, as illustrated by the further cardiac analysis in the HERA trial. The LVEF assessments were MAR and there was a low percentage of missing LVEF values. It was appropriate to perform the repeated measures of LVEF values based on the observed LVEF values. Complete case analysis may be acceptable when the proportion of missing assessments is small (< 5%). In some scenarios, 10%-20% of missing data will have little or no effect on the results of the study. The reason for the missing data needs to be considered as well as the amount of missing data (Fairclough 2010, p.126-127; Little and Rubin 2002, p.41-42). Special care should be taken if imputation is being applied when more than 30-50% of the data are missing (White et al. 2011), and in this scenario the conclusions that can be drawn are restricted (Fairclough 2010, p.127).

When appropriately performed, imputation allows valid inferences from standard procedures. Development of imputation-based procedures continues, while recognising that imputation-based procedures are not always the best approach to analysing missing data. When appropriately performed, imputation allows valid inferences from standard procedures. However, it is important to investigate why observations are missing and to give careful consideration to the final choice of imputation method used as imputation methods involve untestable assumptions. While statistical methods for dealing with missing data exist, it is always preferable to have the actual data and it is important to minimise the amount of missing data in a clinical trial.

References

1 Aaronson, N.K., Ahmedzai, S., Bergman B. et al. (1993). The European Organization for Research and Treatment of Cancer QLQ-30: a quality-of-life instrument for use in international clinical trials in oncology, *Journal of the National Cancer Institute*, 85, 365-376.

2 Altman, D.G. and De Stavola, B.L. (1994). Practical problems in fitting a proportional hazards model to data with updated measurements of the covariates. *Statistics in Medicine*, 13, 301–341, 1994.

3 Andersen, P.K. and Gill, R.D. (1982). Cox's regression model for counting processes: a large sample study. *Annals of Statistics*, 10, 1100-1120.

4 Bang, H and Robins, J.M. (2005). Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics*, 61, 962-972.

5 Baselga, J., Carbonnell X., Castañeda-Soto, N.J. et al. (2005). Phase II study of efficacy, safety and pharmacokinetics of trastuzumab monotherapy administered on a 3-weekly schedule. *Journal of Clinical Oncology*, 23, 2162-2171.

6 Bell, R. (2002). What can we learn from Herceptin trials in metastatic breast cancer? *Oncology,* 63(Supplement 1), 39-46.

7 Bordeleau, L., Szalai, J.P., Ennis, M. et al. (2003). Quality of life in a randomised trial of group psychosocial support in metastatic breast cancer: overall effects of the intervention and an exploration of missing data. *Journal of Clinical Oncology*, 21: 1944-1951.

8 Bradshaw, P.T., Ibrahim, J.G and Gammon, M.D. (2010). A Bayesian proportional hazards regression model with non-ignorably missing time-varying covariates. *Statistics in Medicine*, 29, 3017-3029.

9 Buck, S.F. (1960). A Method of Estimation of Missing Values in Multivariate Data Suitable For Use With an Electronic Computer, *Journal of the Royal Statistical Society, Series B*, 22, 302-306.

10 Carlin, J.B. (2014). Multiple Imputation: Perspective and Historical Overview in Molenberghs, G., Fitzmaurice, G., Kenward, M.G. et al. (eds.) (2014). *Handbook of Missing Data Methodology* (pp. 239-266).

11 Cella, D.F. and Bonomi, A.E. (1995). Measuring quality of life: 1995 update. *Oncology*, 9(11 Supplement), 47-60.

12 Cella, D.F., Tulsky, D.S., Gray, G., et al. (1993). The Functional Assessment of Cancer Therapy Scale: development and validation of the general measure. *Journal of Clinical Oncology*, 11, 570-579.

13 Chen, H. Y. and Little, R. J. (1999). Proportional hazards regression with missing covariates. *Journal of the American Statistical Association*, 94, 896–908.

14 Chlebowski, R. T., Hendrix, S. L., Langer, R. D. et al. (2003). Influence of estrogen plus progestin on breast cancer and mammography in healthy postmenopausal women: the Women's Health Initiative Randomized Trial. *Journal of the American Medical Association*, 289, 3243-3253.

15 Coates A.S., Fischer-Dillenbeck, C., McNeil D.R. et al. (1983). On the receiving end, II: linear analogue self-assessment in the evaluation of quality of life of cancer patients. *European Journal of Cancer and Clinical Oncology*, 19, 1633-1638.

16 Coates A.S., Glasziou, P.P. and McNeil D. (1990). On the receiving end, III: measurement of quality of life during cancer chemotherapy. *Annals of Oncology*, 1, 213-217.

17 Coates, A.S., Hürny, C., Peterson, H.F. et al. (2000). Quality-of-life scores predict outcome in metastatic but not early breast cancer. *Journal of Clinical Oncology*, 18, 3768-3774.

18 Cochran, W.G. (1977). *Sampling Techniques*, John Wiley & Sons, New York

19 Collett, D. (1994). *Modelling Survival Data in Medical Research*, Chapman and Hall, London.

20 Cook-Bruns, N. (2001). Retrospective analysis of the safety of Herceptin immunotherapy in metastatic breast cancer. *Oncology*, 61(Supplement 2), 58-66.

21 Cooperberg, M.R., Cowan, J.E., Hilton, J.F. et al. (2011). Outcomes of Active Surveillance for Men With Intermediate-Risk Prostate Cancer. *Journal of Clinical Oncology*, 29, 228-234.

22 Cox, D.R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34, 187-220.

23 Cox, D. R and Oakes. D. (1984). *Analysis of Surviva1 Data.* Chapman and Hall, London.

24 Crawford, S.L., Tennstedt, S.L. and McKinlay, J.B. (1995). A Comparison of Analytic Methods for Non-Random Missingness of Outcome Data, *Journal of Clinical Epidemiology*, 48, 209-219.

25 Curran, D. (2000), Analysis of Incomplete Longitudinal Data. Doctoral dissertation, Linsburg Universitair Centrum, as cited in Fairclough, D.L. (2010). *Design and Analysis of Quality of Life Studies in Clinical Trials 2nd edition*. Chapman and Hall, Boca Raton.

26 de Azambuja, E., Procter, M.J., van Veldhuisen D.J. et al. (2014). Trastuzumab-Associated Cardiac Events at 8 Years of Median Follow-Up in the Herceptin Adjuvant Trial (BIG 1-01), *Journal of Clinical Oncology*, 32, 2159-2165.

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

28 Diggle, P.J. and Kenward, M.G. (1994). Informative dropout in longitudinal data analysis (with discussion), *Applied Statistics*, 43, 49-93.

29 Diggle, P.J., Heagerty, P.J., Liang, K.Y. and Zeger, S.L. (2002). *Analysis of Longitudinal Data 2nd edition*, Oxford University Press, Oxford.

30 Dupuy, J.-F. and Mesbah, M. (2002). Joint modeling of event time and nonignorable missing longitudinal data, *Lifetime Data Analysis.*, 8, 99-115.

31 Dupuy, J.-F. and Mesbah, M. (2004). Estimation of the Asymptotic Variance of Semi-parametric Maximum Likelihood Estimators in the Cox Model with a

Missing Time-Dependent Covariate, *Communications in Statistics*, 33, 1385-1401.

32 Efron, B. (1986). How biased in the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, 81, 461-470.

33 Epplein, M., Zheng, Y., Zheng, W. et al. (2011). Quality of life after breast cancer diagnosis and survival. *Journal of Clinical Oncology*, 29, 406-412.

34 Everitt, B. (1993). *Cluster Analysis 3rd edition*, John Wiley & Sons, Chichester.

35 Fairclough, D.L. et al. for the Eastern Cooperative Oncology Group (1999). Quality of Life and Quality Adjusted Survival for Breast Cancer Patients Receiving Adjuvant Therapy, *Quality of Life Research*, 8; 723- 731.

36 Fairclough, D.L. (2002). *Design and Analysis of Quality of Life Studies in Clinical Trials*. Chapman and Hall, Boca Raton.

37 Fairclough, D.L. (2010). *Design and Analysis of Quality of Life Studies in Clinical Trials 2nd edition*. Chapman and Hall, Boca Raton.

38 Fayers, P.M. and Machin, D. (2000). *Quality of Life: Assessments, Analysis and Interpretation*. John Wiley & Sons, Chichester.

39 Fayers, P.M. and Machin, D. (2007). *Quality of Life: Assessments, Analysis and Interpretation 2nd edition*. John Wiley & Sons, Chichester.

40 Fielding, S., Fayers, P.M. and Ramsay, C.R. (2009). Investigating the missing data mechanism in quality of life outcomes: a comparison of approaches. *Health and Quality of Life Outcomes*, 7, 57-66.

41 Ganz, P.A. Land, S.R., Geyer, C.E. et al. (2011). Menstrual History and Quality-of-Life Outcomes in Women With Node-Positive Breast Cancer Treated With Adjuvant Therapy on the NSAPB B-30 Trial. *Journal of Clinical Oncology*, 29, 1110-1116.

42 Gelman, A., Carlin, J.B., Stern, H.S. and Rubin D.B. (2004). *Bayesian Data Analysis 2nd edition*, Chapman and Hall, Boca Raton.

43 Geman, S. and Geman, D. (1984). Stochastical Relaxation, Gibbs Distributions and the Bayesian Restoration of Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.
44 Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (editors) (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall, Boca Raton.
45 Gilson, B.S., Gilson, J.S., Bergner, M. et al. (1975). The sickness impact profile. Development of an outcome measure of health care. *American Journal of Public Health*, 65, 1304-1310.
46 Gleason, T.C. and Staelin, R. (1975). A proposal for handling missing data. *Psychometrika*, 40, 229-252.
47 Goldhirsch, A., Gelber, R.D., Piccart-Gebhart, M.J. et al. (2013). 2 years versus 1 year of adjuvant trastuzumab for HER2-positive breast cancer (HERA): An open-label, randomised controlled trial. *Lancet*, 382, 1021-1028.
48 Graham, J.W., Olchowski, A.E, and Gilreath, T.D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, 8, 206-213.
49 Greenwood, M. (1926). The Natural Duration of Cancer. *Reports on public health and medical subjects,* vol. 33, 1–26. H.M. Stationary Office, London, UK.
50 Grossman, S.A., Sheilder, V.A., Sweedeen K. et al. (1991). Correlation of a patient and caregiver ratings of cancer pain. *Journal of Pain and Symptom Management*, 6, 53-57.
51 Grothey, A. Nikcevich, D.A., Sloan, J.A. et al. (2011). Intravenous Calcium and Magnesium for Oxaliplatin-Induced Sensory Neurotoxicity in Adjuvant Colon Cancer: NCCTG N04C7. *Journal of Clinical Oncology*, 29, 421-427.
52 Gschwind, A., Fischer, O.M. and Ullrich, A. (2004). The discovery of receptor tyrosine kinases: targets for cancer therapy. *Nature Reviews Cancer*, 4, 361-370.
53 Harville, D.A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association*, 72, 320–338.

54 He, Y., Yucel, R. and Raghunathan, T.E. (2011). A functional multiple imputation approach to incomplete longitudinal data. *Statistics in Medicine*, 30, 1137-1156.

55 Heitjan, D.A. and Landis, J.R. (1994). Assessing Secular Trends in Blood Pressure: A Multiple Imputation Approach, *Journal of the American Statistical Association*, 89, 750-759.

56 Henderson, C. R. (1984). *Applications of Linear Models in Animal Breeding*, University of Guelph, Guelph

57 Henderson, C. R. (1990). Statistical Method in Animal Improvement: Historical Overview, in Hammond, K. and Gianola, D. (eds.) (1990). *Advances in Statistical Methods for Genetic Improvement of Livestock (pp. 1-14)***,** Springer-Verlag, Berlin.

58 Herring, A.H, Ibrahim, J.G. and Lipsitz, S.R. (2004). Non-ignorable missing covariate data in survival analysis: a case-study of an International Breast Cancer Study Group trial, *Journal of the Royal Statistical Society, Series C*, 53, 293-310.

59 Holm, S. (1979). A simple sequential rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, 65-70.

60 Holt, J.D. (1978). Competing risk analyses with special reference to matched pair experiments. *Biometrika***,** 65, 159–165.

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

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

63 Hürny, C., Bernhard, J., Bacchi, M., et al. (1993). The Perceived Adjustment to Chronic Illness Scale (PACIS): a global indicator of coping for operable breast cancer patients in clinical trials. *Supportive Care in Cancer*, 1, 200-208.

64 Hürny, C., Bernhard, J., Coates, A.S. et al. (1996). Impact of adjuvant therapy on quality of life in women with node-positive operable breast cancer. *Lancet*, 347, 1279-1284.

65 Hürny, C., Bernhard, J., Coates, A.S., et al. (1996a). Responsiveness of a single-item indicator versus a multi-item scale: assessment of emotional wellbeing in an international adjuvant breast cancer trials. *Medical Care*, 34, 234-248.

66 Hurwitz, H.I., Saltz, L.B. and Cutsem, E.V. (2011). Venous Thromboembolic Events With Chemotherapy Plus Bevacizxumab: A Pooled Analysis of Patients in Randomised Phase II and III Studies. *Journal of Clinical Oncology*, 29, 1757-1764.

67 International Conference on Harmonisation (1996). Guideline for Good Clinical Practice. ICH Secretariat, Geneva (http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E6/E6_R1_Guideline.pdf; last accessed 17May2016)

68 International Conference on Harmonisation (1998). *Statistical Principles for Clinical Trials*. ICH Secretariat, Geneva (c/o European Medicines Agency http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002928.pdf; last accessed 17May2016).

69 Jackisch C., Piccart M., Gelber R.D. et al. (2015). HERA Trial: 10-Year Follow Up of Trastuzumab after Adjuvant Chemotherapy in HER2 Positive Early Breast Cancer – Final Analysis (http://sabcs15.posterview.com/nosl/i/PD5-01; last accessed 17May2016)

Poster presented at 38th San Antonio Breast Cancer Symposium, San Antonio, Texas, December 2015

70 Jansen, I., Hens, N., Molenberghs, G. et al. (2006). The nature and sensitivity in monotone missing not at random models. *Computational Statistics & Data Analysis*, 50, 830-858.

71 Jett, A.M., Davies, A.R., Cleary, P.D. et al. (1986). The functional status questionnaire: its reliability and validity when used in primary care. *Journal of General Internal Medicine*, 1, 143-149.

72 Jolani, S., Frank, L. E. and van Buuren, S. (2014). Dual imputation model for incomplete longitudinal data. *British Journal of Mathematical and Statistical Psychology*, 67: 197–212.
73 Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis* of *Failure Time Data*. John Wiley & Sons Inc., New York.
74 Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis* of *Failure Time Data 2nd edition*, John Wiley & Sons Inc., Hoboken
75 Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal* of *the American Statistical Association*, 53, 457-481.
76 Kaplan, R.M. and Bush, J.W. (1982). Health-related quality of life measurement for evaluation and research and policy analysis, *Health Psychology*, 1, 61-80.
77 Kenne Sarenmalm, E., Odén, B., Öhlén, J. et al. (2009). Changes in health-related quality of life may predict recurrent breast cancer. *European Journal of Oncology Nursing*, 13, 323-329.
78 Kenward, M. G. (1998). Selection models for repeated measurements for nonrandom dropout: an illustration of sensitivity. *Statistics in Medicine*, 17, 2723–2732.
79 Kenward, M.G and Carpenter, J. (2007). Multiple imputation: current perspectives. *Statistical Methods in Medical Research*, 16, 199-218.
80 Kitahara, C.M., Berrington de González, A., Freedman, N.D. et al. (2011). Total Cholesterol and Cancer Risk in a Large Prospective Study in Korea. *Journal of Clinical Oncology*, 29, 1592-1598.
81 Kornblith, A.B., Lan, L. and Archer, L. (2011). Quality of Life of Older Patients with Early-Stage Breast Cancer Receiving Adjuvant Chemotherapy: A Companion Study to Cancer and Leukemia Group B 49907. *Journal of Clinical Oncology*, 29, 1022-1028.
82 Kudoh, S., Takeda K., Nakagawa, K. et al. (2006). Phase III study of docetaxel compared with vinorelbine in elderly patients with advanced non–small-cell lung

cancer: results of the West Japan Thoracic Oncology Group Trial (WJTOG 9904). *Journal of Clinical Oncology*, 24, 3657-3663.

83 Lan, K.K.G. and DeMets, D.L. (1983). Discrete sequential boundaries for clinical trials. *Biomatrika*, 70, 659-663.

84 Lang, T.A. and Secic, M. (2006). *How to report statistics in medicine 2nd edition*, American College of Physicians, Philadelphia

85 Lessler, J.T. and Kalsbeek, W.D. (1992). *Nonsampling Errors in Surveys*, John Wiley & Sons Inc., New York.

86 Leong, T., Lipsitz, S. R. and Ibrahim, J. G. (2001). Incomplete covariates in the Cox model with applications to biological marker data. *Journal of the Royal Statistical Society, Series C*, 50, 467–484.

87 Likert, R.A. (1932). A technique for the measurement of attitude. *Archives of Psychology*, 22, 1-55.

88 Lin, D. Y. and Ying, Z. (1993). Cox regression with incomplete covariate measurements. *Journal* of *the American Statistical Association*, 88, 1341–1349.

89 Lipsitz, S. R. and Ibrahim, J. G. (1998) Estimating equations with incomplete categorical covariates in the Cox model. *Biometrics*, 54, 1002–1013.

90 Little, R.J.A. (1988). A test of missing completely at random for multivariate data with missing values, *Journal* of *the American Statistical Association*, 83, 1198-1202.

91 Little, R.J.A. (1993). Pattern-Mixture Models for Multivariate Incomplete Data, *Journal of the American Statistical Association*, 88, 125-134.

92 Little, R.J.A. (1994). A Class of Pattern Mixture Models for Multivariate Incomplete Data, *Biometrica*, 81, 471-483.

93 Little, R.J.A. (1995). Modelling the Dropout Mechanism in Repeated-Measures Studies, *Journal of the American Statistical Association*, 90, 1112-1121.

94 Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis With Missing Data*, John Wiley & Sons Inc., New York.

95 Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis With Missing Data, Man2nd edition,* John Wiley & Sons Inc., Hoboken

96 Little, R. and Wang, L. (1996). Pattern mixture models for multivariate incomplete data with covariates, *Biometrics*, 52, 98-111.
97 Little, R.A. and Yau L. (1996). Intent-to-Treat Analysis for Longitudinal Studies With Dropout, *Biometrics*, 52, 1324-1333.
98 Lessler, V.M. and Kalsbeek, W.D. (1992). *Nonsampling Errors in Surveys*, John Wiley & Sons Inc., New York.
99 Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS -- a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10, 325-337.
100 MacKenzie, T. and Abrahamowicz, M. (2002). Marginal and hazard ratio specific random data generation: Applications to semi-parametric bootstrapping, *Statistics and Computing*, 12, 245-252.
101 Mantel, N. (1966). Evaluation of Survival Data and Two New Rank Order Statistics Arising in its Consideration. *Cancer Chemotherapy Reports*, 50, 163-170.
102 Marker, D.A., Judkins, D.R. and Winglee, M. (2002). Large-Scale Imputation for Complex Surveys with Imputed Data in Groves, R.M. et al. (editors), *Survey Nonresponse* chapter 22, John Wiley & Sons Inc., New York.
103 Marshall, A., Altman, G., Royston, P. and Holder, R.L. (2010). Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. *BMC Medical Research Methodology*, 10, 7
104 Martinussen, T. (1999). Cox regression with incomplete covariate measurements using the EM-algorithm. *Scandinavian Journal of Statistics*, 26, 479–491.
105 Marty, M., Cognetti, F., Maraninchi, D. et al. (2005) Randomised phase II trial of the efficacy and safety of trastuzumab combined with docetaxel in patients with human epidermal growth factor receptor 2-positive metastatic breast cancer administered as first-line treatment: the M77001 Study group. *Journal of Clinical Oncology*, 23:4256-4274.

106 Meng, X.-L. (2002). A Congenial Overview and Investigation of Multiple Imputation Inferences under Uncongeniality in Groves, R.M. et al. (editors), *Survey Nonresponse* chapter 23, John Wiley & Sons Inc., New York.
107 Miller, G.A. (1956). The magic number of seven plus or minus two: some limits on our capacity for information processing, *Psychological Review*, 63, 81-97.
108 Molenberghs, G. and Kenward, M.G. (2007). *Missing Data in Clinical Studies*, John Wiley & Sons, Chichester.
109 Morris, T.P., White, I. R. and Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology*; 14:75. doi:10.1186/1471-2288-14-75
110 Osborne, R.J., Fillaci, V., Schink, J.C. et al. (2011). Phase III Trial of Weekly Methotrexate or Pulse Dactinomycin for Low-Risk Gestational Trophoblastic Neoplasia: A Gynecologic Oncology Group Study. *Journal of Clinical Oncology*, 29, 825-831.
111 Otterstad, J. E., Froeland, G., St John Sutton, M., and Holme, I. (1997). Accuracy and reproducibility of biplane two-dimensional echocardiographic measurements of left ventricular dimensions and function. *European Heart Journal*, 18, 507-513.
112 Paik, M. C. (1997). Multiple imputation for the Cox proportional hazards model with missing covariates. *Lifetime Data Analysis*, 3, 289–298.
113 Paik, M. C. and Tsai, W.-Y. (1997). On using the Cox proportional hazards model with missing covariates. *Biometrika*, 84, 579–593.
114 Patterson, H.D and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58 (3): 545–554.
115 Patrick, D. and Erickson, P. (1993). *Health Studies and Health Policy: Allocating Resources to Health Care*, Oxford University Press, New York.
116 Peyre, S., Coste, J. and Leplège, A. (2010). Identifying type and determinants of missing items in quality of life questionnaire: Application to the SF-36 French

version of the 2003 decennial health survey. *Health and Quality of Life Outcome*s, 8: 16-21.

117 Peyre, S., Leplège, A. and Coste, J. (2011). Missing data methods for dealing with missing items in quality of life questionnaires. A comparison by simulation of personal mean score, full information maximum likelihood, multiple imputation, and hot deck techniques applied to the SF-36 in the French 2003 decennial health survey. *Quality of Life Research*, 20, 287-300.

118 Piantadosi, S. (2005). Clinical Trials A Methodologic Perspective 2nd edition, John Wiley & Sons Inc., Hoboken.

119 Piccart-Gebhart, M.J., Procter, M., Leyland-Jones, B. et al. (2005). Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. *New England Journal of Medicine,* 353, 1659-1672.

120 Pivot, X., Romieu, G., Debled, M. et al. (2013). 6 months versus 12 months of adjuvant trastuzumab for patients with HER2-positive early breast cancer (PHARE): a randomized phase 3 trial. *Lancet Oncology*, 14, 741-478.

121 Pocock, S.J., Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in a controlled clinical trial. *Biometrics*, 31, 103-115.

122 Priestman, T.J. and Baum, M. (1976). Evaluation of quality of life in patients receiving treatment for advanced breast cancer. *Lancet*, 307, 899-901.

123 Procter, M., Suter, T.M., de Azambuja, E. et al. (2010). Longer-term assessment of trastuzumab-related cardiac adverse events in the Herceptin Adjuvant (HERA) Trial, *Journal of Clinical Oncology*, 28, 3422-4328.

124 Pugh, M., Robins, J., Lipsitz, S. and Harrington, D. (1993). Inference in the Cox proportional hazards model with missing covariate data. *Technical Report*. Division of Biostatistical Science, Dana–Farber Cancer Institute, Boston.

125 Qi, L., Wang, Y.-F. and He, Y. (2010). A Comparison of Multiple Imputation and Fully Augmented Weighted Estimators for Cox Regression with Missing Covariates, *Statistics in Medicine*, 29, 2592-2604.

126 Rabound, J.M. et al. (1998). Estimating the Effect of Treatment of Quality of Life in the Presence of Missing Data Due to Dropout and Death, *Quality of Life Res*earch, 7; 487-494.

127 Ranstam, J. et al. (2012). Alternative analyses for handling incomplete follow-up in the intention-to-treat analysis: the randomised controlled trial of balloon kyphoplasty versus non-surgical care for vertebral compression fracture (FREE). *BMC Medical Research Methodology*, 12, 35. doi:10.1186/1471-2288-12-35

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

129 Rotnitzky, A., Robins, J.M. and Scharfstein, D.O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*, 93, 1321–1339.

130 Royall, R.M. (1986). The effect of samples size on the meaning of significance tests. *American Statistician*, 40, 313-315.

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

132 Rubin, D.B. (1987). *Multiple Imputation for Non-response in Surveys*. John Wiley & Sons Inc., New York.

133 Rubin, D.B. and Schenker, N. (1986). Multiple Imputation for Interval Estimation from Simple Random Samples With Ignorable Non-response, *Journal of the American Statistical Association*, 81, 366-374.

134 Rubin, D.B. and Schenker, N. (1991). Multiple Imputation in Health-Care Data bases: An Overview and Some Applications, *Statistics in Medicine*, 10, 585-598.

135 Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman and Hall, London.

136 Schafer, J. (1999). Multiple Imputation: A Primer, *Statistical Methods in Medical Research*, 8, 3-15.

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

138 Schafer, J.L. and Olsen, M.K. (1998). Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Persepctive. *Multivariate Behavioral Research*, 33, 545-571.

139 Scharfstein, D.O., Rotnitzky, A. and Robins, J.M. (1999). Adjusting for nonignorable drop-out using semiparameteric nonresponse models. *Journal of the American Statistical Association*, 94, 1096-1120 (with Rejoinder 1135-1146).

140 Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, 69, 239-241.

141 Searle, S.R. (1971). *Linear Models*. John Wiley & Sons Inc., New York.

142 Searle, S. R., Casella, G., and McCulloch, C. E. (1992), *Variance Components*, John Wiley & Sons Inc., New York.

143 Siemes, C., Visser, L.E, Coebergh J.-W. W. et al. (2006). C-reactive protein levels, variation in the C-reactive protein gene, and cancer risk: the Rotterdam study. *Journal of Clinical Oncology*, 24, 5216-5222.

144 Slamon, D.J., Clark, G.M., Wong, S.G. et al. (1987). Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu onogene. *Science*, 235, 177-182.

145 Slamon, D.J., Godolphon, W., Jones, L.A. et al. (1989). Studies of the HER-2/neu proto-onogene in human breast cancer and ovarian cancer. *Science*, 244, 707-712.

146 Slamon, D.J., Leyland-Jones, B., Shak, S. et al. (2001). Concurrent administration of anti-HER2 monoclonal antibody and first line chemotherapy for HER2-overexpressing metastatic breast cancer. A phase III, multinational, randomised controlled trial. *New England Journal of Medicine* 344, 784-792.

147 Smith, I., Procter, M., Gelber, R.D. et al. (2007). 2-year follow-up of trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer: a randomised controlled trial. *Lancet* 369; 29-36.

148 Song, X. and Wang, C.Y. (2013). Time-varying coefficient proportional hazards model with missing covariates. *Statistics in Medicine* 32; 2013-2030.
149 Spiegelhalter, D.J, Best, N.G., Carlin, B.P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society, Series B*, 64, 583–639.
150 Spitzer, W.O., Dobson, A.J., Hall, J. et al. (1981). Measuring the quality of life of cancer patients. *Journal of Chronic Diseases*, 34, 585-597.
151 Sterne, J.A.C, White, R., Carlin, J.B. et al. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *British Medical Journal*, 338, b.2393
152 Stanton, A.L., Ganz, P.A., Kwan, L. et al. (2005). Outcomes from the Moving Beyond Cancer psychoeducational, randomised, controlled trial with breast cancer patients. *Journal of Clinical Oncology*, 23, 6009–6018.
153 Stiff, P.J., Emmanouilides, C., Bensinger, W.I. et al. (2006). Palifermin reduces patient-reported mouth and throat soreness and improves patient functioning in the hematopoietic stem-cell transplantation setting. *Journal of Clinical Oncology*, 24, 5186-5193.
154 Streiner, D.L. and Norman, G.R. (2008). *Health Measurement Scales – A Practical Guide to Their Development and Use 4th edition*, Oxford University Press, Oxford.
155 Suter, T.M., Procter, M., van Veldhuisen, D.J. et al. (2007). Trastuzumab-associated cardiac adverse effects in the Herceptin Adjuvant Trial. *Journal of Clinical Oncology*, 25, 3859-3865.
156 Syrjala, K.L., Artherholt, S.B., Kurland, B.F. (2011). Prospective Neurocognitive Function Over 5 Years After Allogenic Hematopoietic Cell Transplantation for Cancer Survivors Compared With Matched Controls at 5 Years. *Journal of Clinical Oncology*, 29, 2397-2404.
157 Tanner, M.A. and Wong, W.H. (1987). The calculation of posterior distributions by data augmentation (with discussion), *Journal of the American Statistical Association*, 82, 528-550

158 The EuroQoL Group (1990). EuroQoL-a new facility for the measurement of health-related quality of life, *Health Policy*, 16, 199-208.

159 The International Breast Cancer Group. (1996). Duration and reintroduction of adjuvant chemotherapy for node-positive premenopausal breast cancer patients. *Journal of Clinical Oncology*, *14*, 1885-1894

160 The International Breast Cancer Group (1997). Effectiveness of adjuvant chemotherapy in combination with tamoxifen for node-positive postmenopausal breast cancer patients. *Journal of Clinical Oncology*, *15*, 1385-1394.

161 Thomas Reuters (2011). The 2010 Journal Citation Reports® http://thomsonreuters.com/products_services/science/science_products/a-z/journal_citation_reports;
https://www.researchgate.net/journal/1527-7755_Journal_of_Clinical_Oncology (last accessed 17May2016)

162 Urruticoechea, A., Smith, I.E. and Dowsett, M. (2005). Proliferation Marker Ki-67 in Early Breast Cancer, *Journal of Clinical Oncology*, 23, 7212-7220.

163 Van Buuren, S., Boshuizen, H.C. and Knook, D.L. (1999). Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis, *Statistics in Medicine*, 18, 681-694.

164 Vogel, C.L., Cobleigh, M.A., Tripathy, D. et al. (2002). Efficacy and safety of trastuzumab as a single agent in the first-line treatment of HER2-overexpressing metastatic breast cancer. *Journal of Clinical Oncology*, 20:719-726.

165 von Zerssen, D. (1986). Clinical self-rating scales (CSRS) of the Munich Psychiatric Information System (PYSCHIS Muenchen), in Sartorius N., Ban, T.A. (eds.) (1986). *Assessment of depression* (pp. 270-303). Springer-Verlag, Berlin.

166 Wang, C.Y. and Chen, H.Y. (2001). Augmented inverse probability weighted estimator for Cox missing covariate regression. *Biometrics*, 57, 414-419.

167 Ware, J.E. (1987). Standards for validating health measures: definition and content. *Journal of Chronic Diseases*, 40, 473-480.

168 Ware, J.E., Brook, R.H., Davies, A.R. and Lohr, K.N. (1981). Choosing measures of health status for individuals in the general populations. *American Journal of Public Health*, 71, 620-625.

169 White, I.R. and Royston, P. (2009). Imputing missing covariate values for the Cox model, *Statistics in Medicine*, 28, 1982-1998.

170 White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, 30, 377-399.

171 Wiklund, I. Dimenäs, E. and Wahl, M. (1990). Factors of importance when evaluating quality of life in clinical trials, *Controlled Clinical Trials*, 11, 169-179.

172 Wilkinson, D.J. https://darrenjw.wordpress.com/2011/07/16/gibbs-sampler-in-various-languages-revisited/ (last accessed 17May2016)

173 Wilson, I.B. and Cleary, P.D. (1995). Linking Clinical Variables with Health-Related Quality of Life, *Journal of the American Medical Association*, 273, 59-65.

174 World Health Organization (1948), *Constitution of the World Health Organization*, WHO, Geneva.

175 Word Health Organization (1958). *The First Ten Years of the World Health Organization*, WHO, Geneva.

176 Yarden, Y. and Silwkowski, M. (2001). Untangling the ErbB signalling network. *Nature Reviews Moleculat Cell Biology*, 2: 127-137.

177 Zhou, H. and Pepe, M. S. (1995). Auxiliary covariate data in failure time regression. *Biometrika*, 82, 139–149.

178 Zhou, H. and Wang, C.Y. (2000). Failure time regression analysis with measurement error in covariates, *Journal of the Royal Statistical Society, Series B*, 62, 657–665.

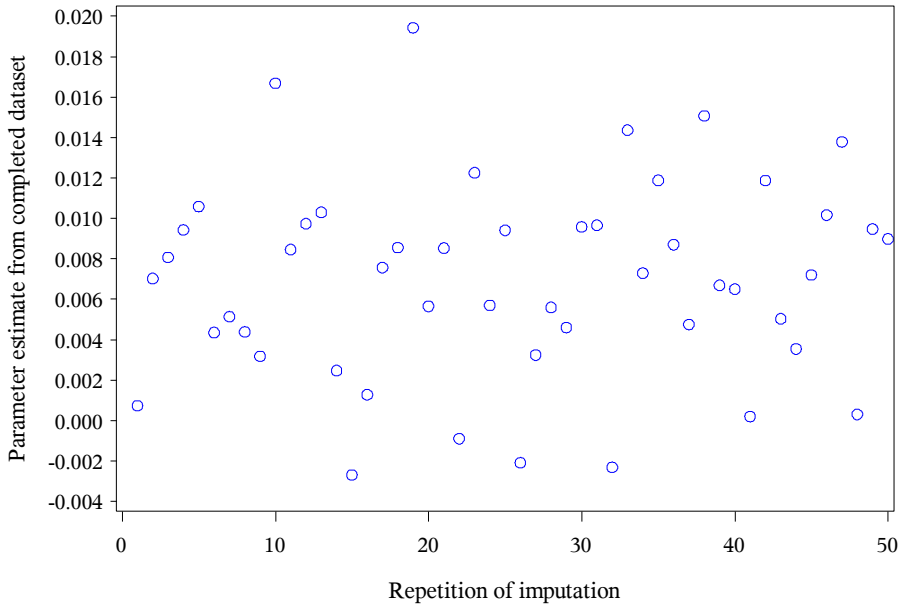
Appendix A Parameter Estimates from the Completed Dataset, the Cumulative Mean Parameter Estimates and the Decomposition of the Variance of the Parameter for Square Root of Coping Score (β_{sp}) for the Remaining Standard Multiple Imputation Methods

The figures for the parameter estimates for the square root of the coping score (S_Pacis) and delayed chemotherapy from the completed dataset and the cumulative mean parameter estimates for the remaining standard multiple imputation methods are shown as noted below:

Method	Detail	Estimate of β_{sp}	Estimate of β_{del}
Bootstrapping	previous coping score	Figure A1.1	Figure A1.2
Nearest neighbour		Figure A2.1	Figure A2.2
Predictive mean matching	initial steps as described for NNI	Figure A3.1	Figure A3.2
Pattern mixture models	Curran's analytic technique	Figure A4.1	Figure A4.2

The decomposition of the variance of β_{sp} for the remaining standard multiple imputation methods is shown in Table A1.1.

A



B

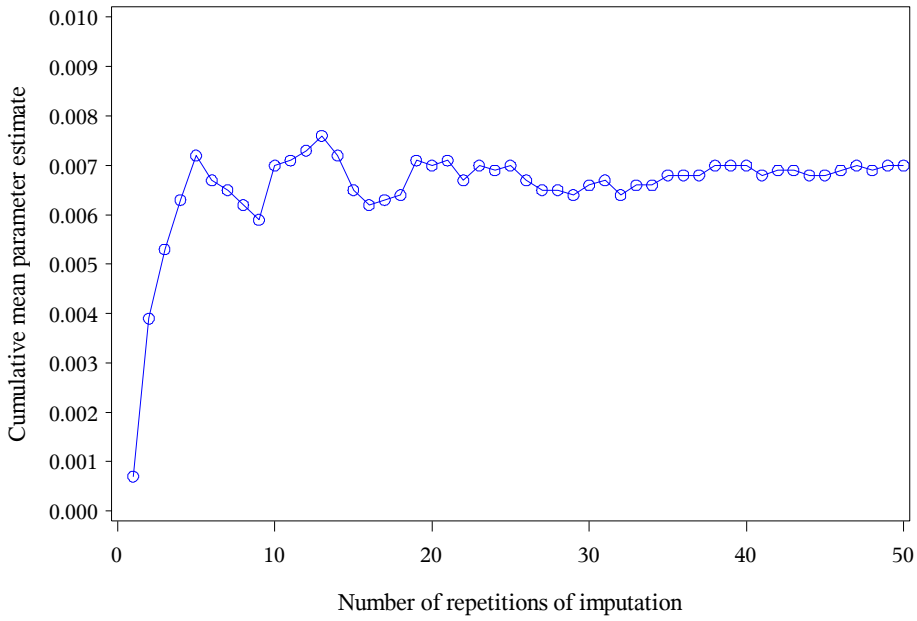
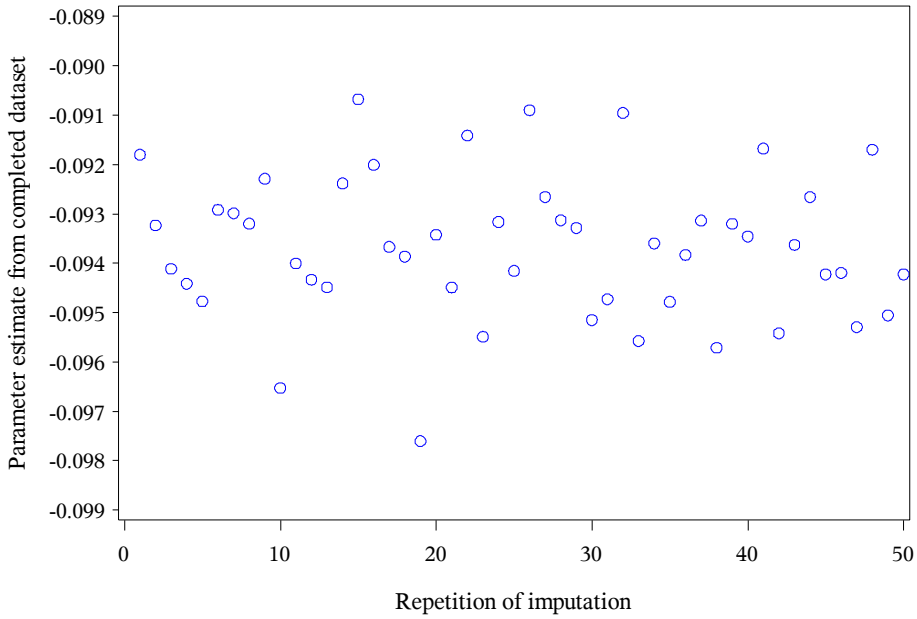


Figure A1.1 Parameter estimate from the completed dataset (A) and the cumulative mean parameter estimate (B) for the square root of the coping score (S_Pacis) from the time-dependent Cox model analysis following bootstrap imputation, subgroups defined by previous coping scores by the number of repetitions of imputation

A



B

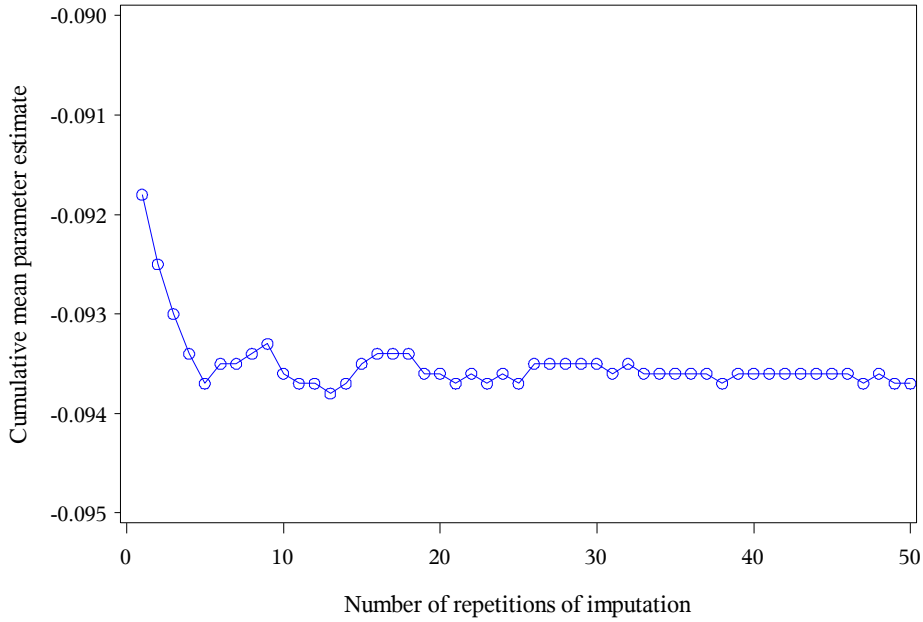
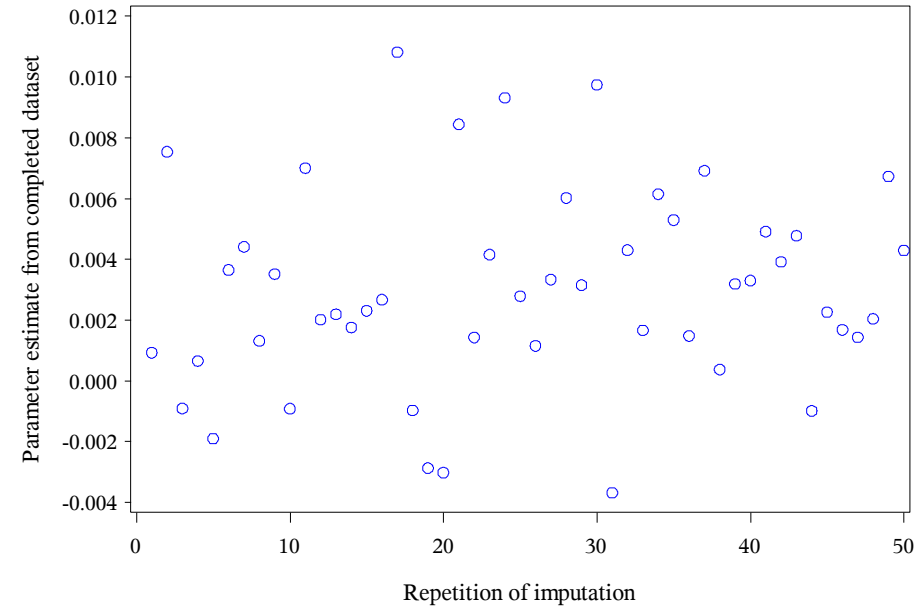


Figure A1.2 Parameter estimate from the completed dataset (A) and the cumulative mean parameter estimate (B) for delayed chemotherapy from the time-dependent Cox model analysis following bootstrap imputation, subgroups defined by previous coping scores by the number of repetitions of imputation

A



B

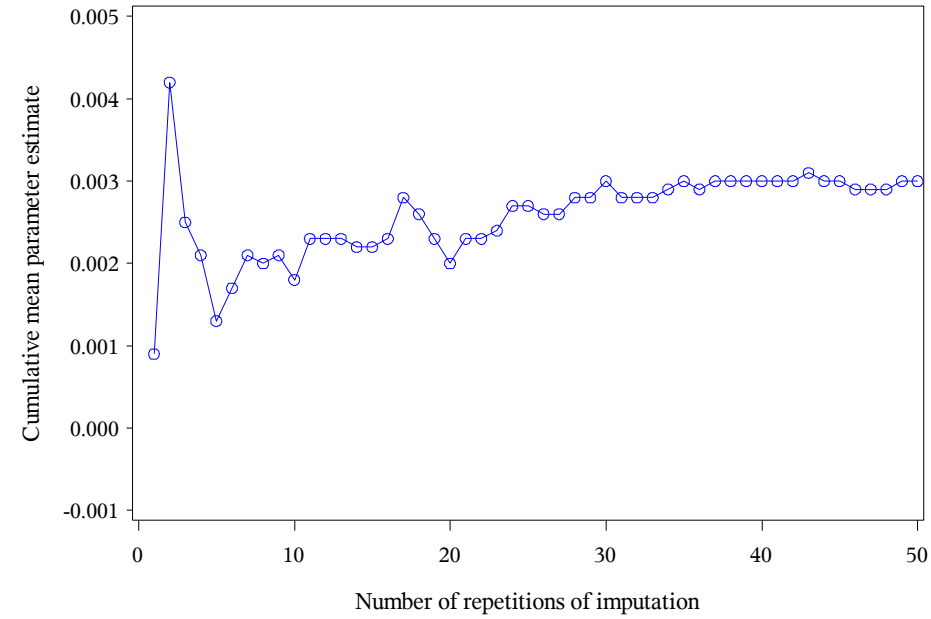
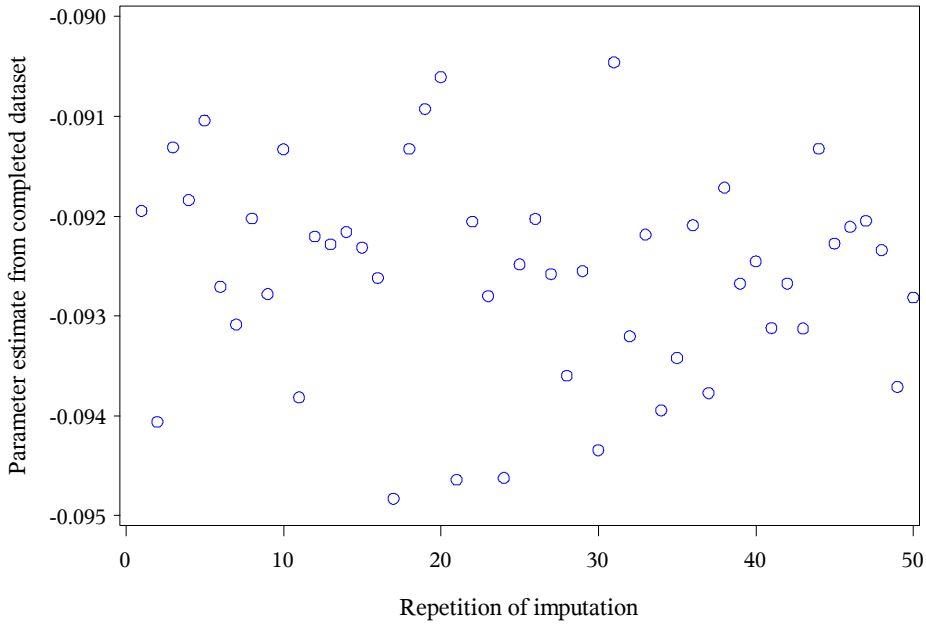


Figure A2.1 Parameter estimate from the completed dataset (A) and the cumulative mean parameter estimate (B) for square root of the coping score (S_Pacis) from the time-dependent Cox model analysis following nearest neighbour imputation by the number of repetitions of imputation

A



B

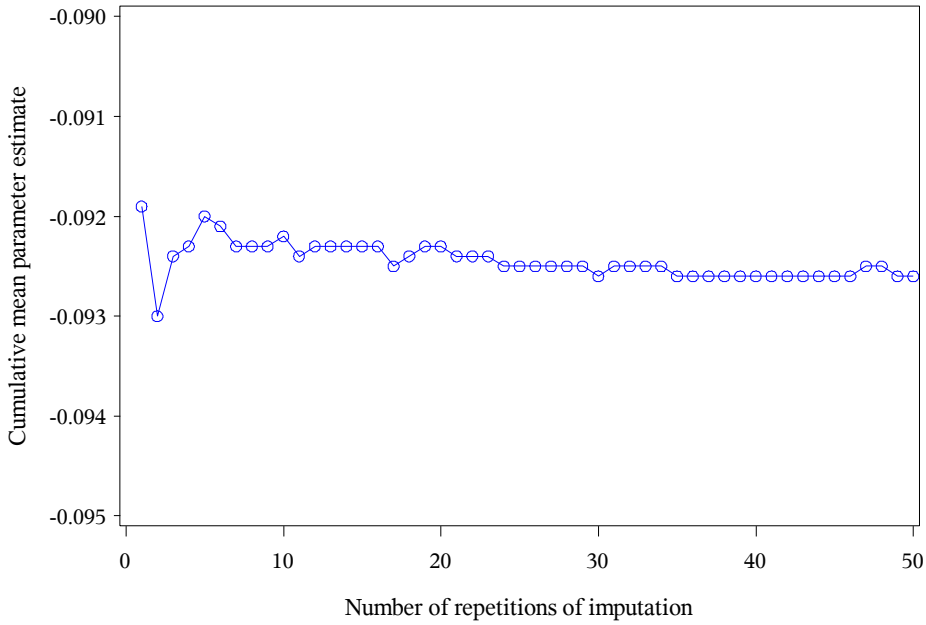
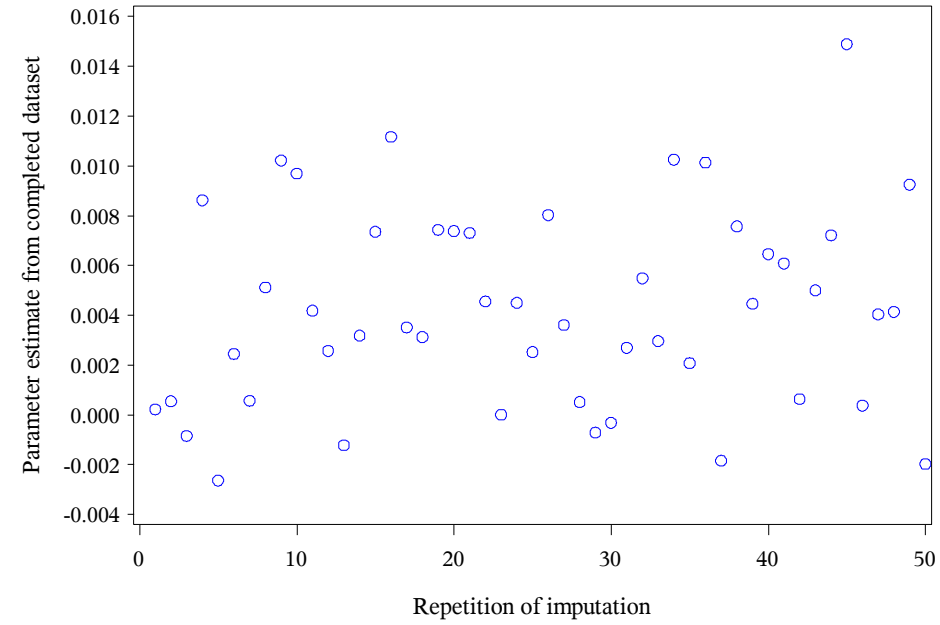


Figure A2.2 Parameter estimate from the completed dataset (A) and the cumulative mean parameter estimate (B) for delayed chemotherapy from the time-dependent Cox model analysis following nearest number imputation by the number of repetitions of imputation

A



B

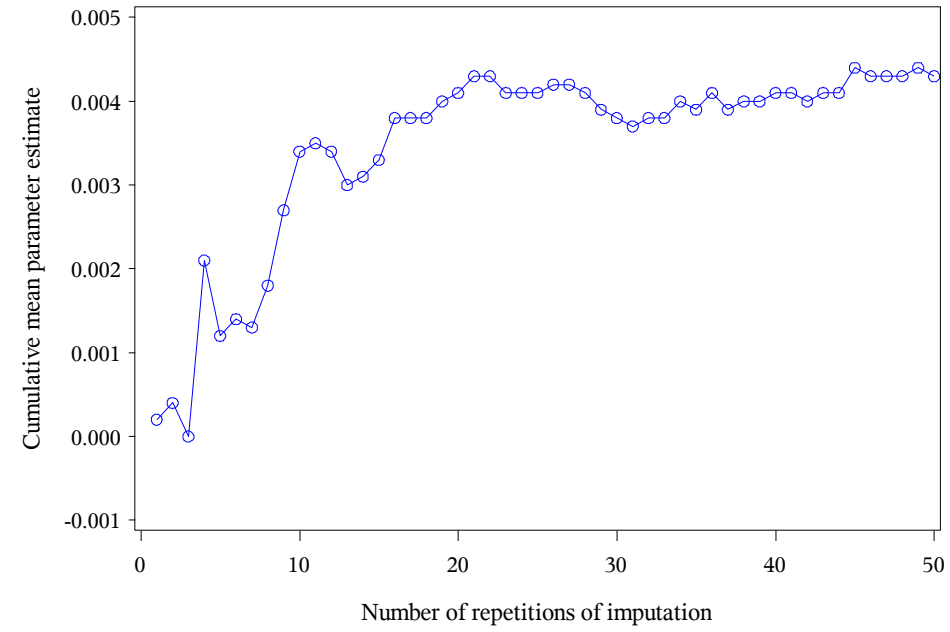
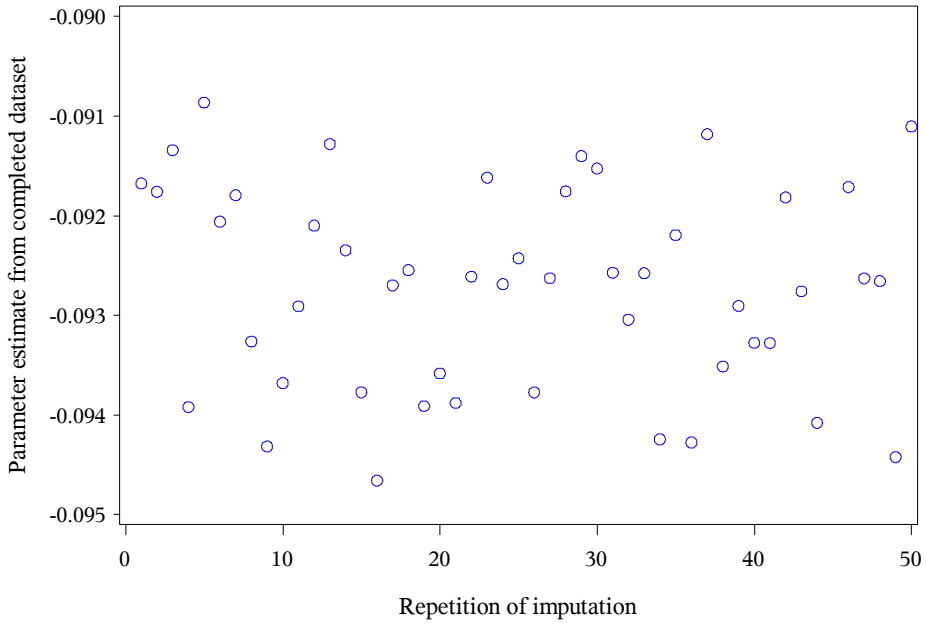


Figure A3.1 Parameter estimate from the completed dataset (A) and the cumulative mean parameter estimate (B) for square root of the coping score (S_Pacis) from the time-dependent Cox model analysis following imputation by predictive mean matching by the number of repetitions of imputation

A



B

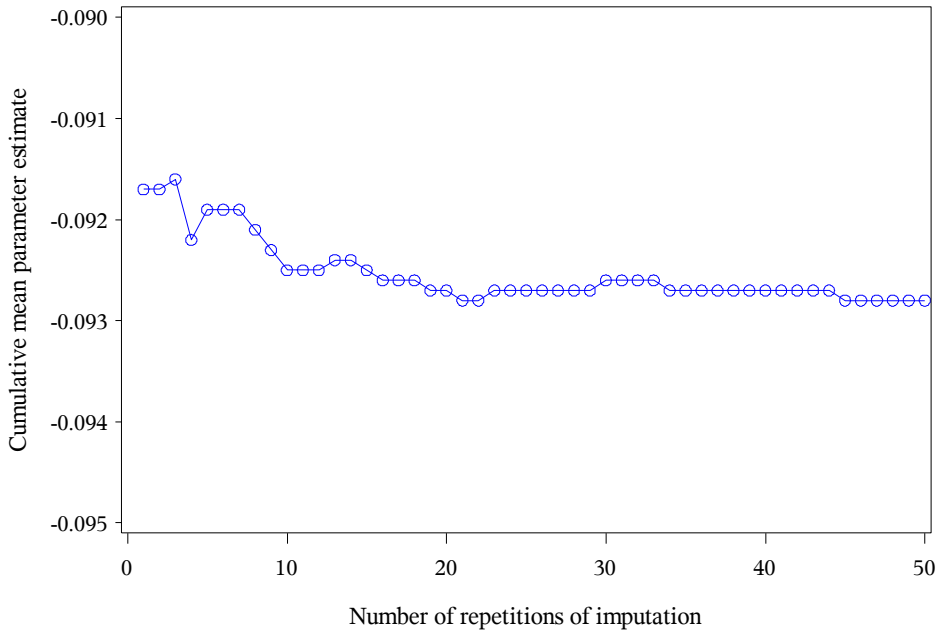
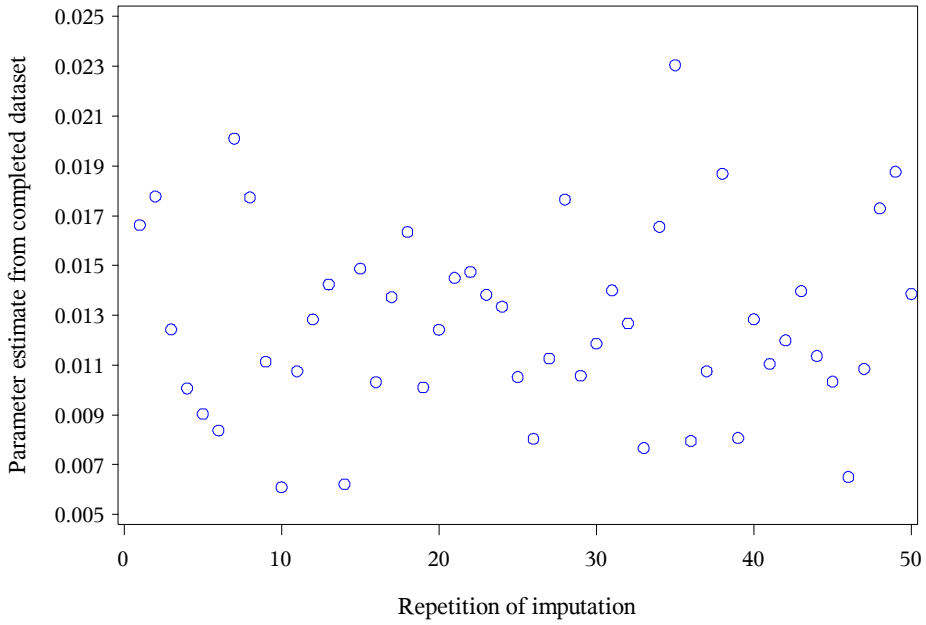


Figure A3.2 Parameter estimate from the completed dataset (A) and the cumulative mean parameter estimate (B) for delayed chemotherapy from the time-dependent Cox model analysis following imputation by predictive mean matching by the number of repetitions of imputation

A



B

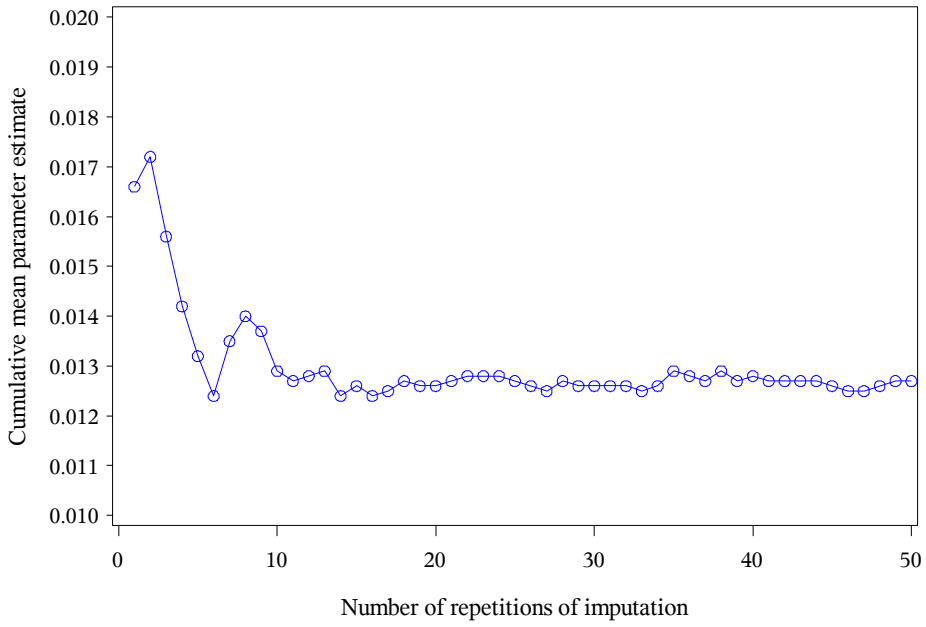
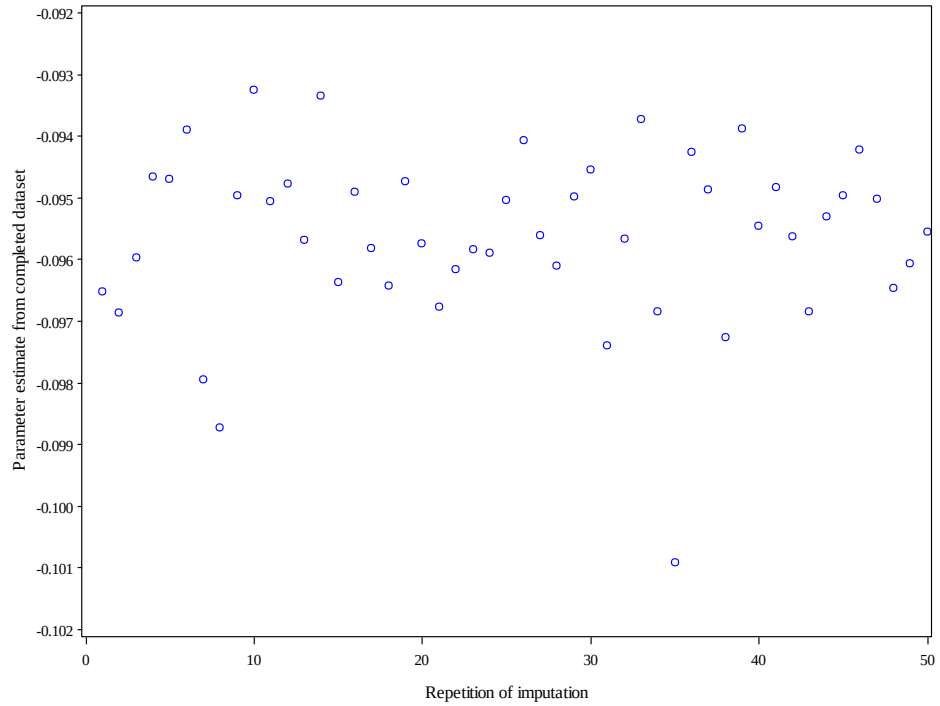


Figure A4.1 Parameter estimate from the completed dataset (A) and the cumulative mean parameter estimate (B) for square root of the coping score (S_Pacis) from the time-dependent Cox model analysis following imputation by pattern mixture models by the number of repetitions of imputation

A



B

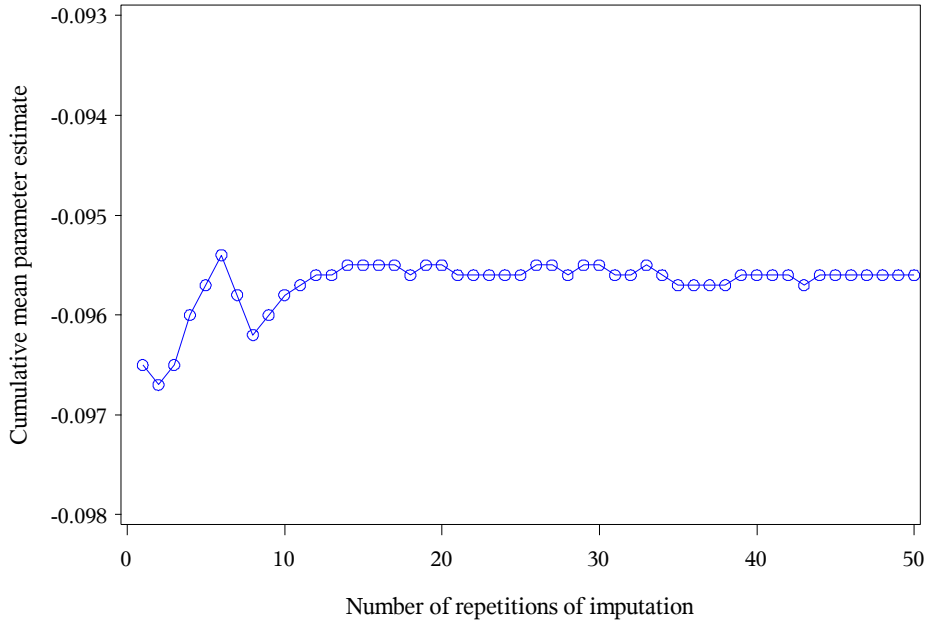


Figure A4.2 Parameter estimate from the completed dataset (A) and the cumulative mean parameter estimate (B) for delayed chemotherapy from the time-dependent Cox model analysis following imputation by pattern mixture models by the number of repetitions of imputation

Table A1.1 Variance Decomposition of the Parameter for Square Root of Coping Score (β_{sp}) from Time-Dependent Cox Model Analysis Following Standard Imputation Methods

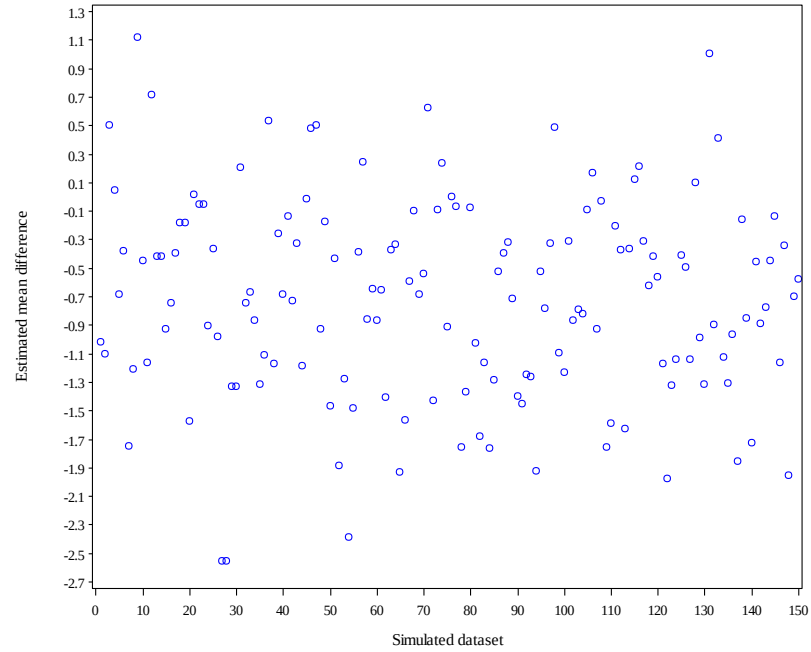
Bootstrap - previous coping scores			
	K=5	*K*=40	*K*=50
Within-imputation (W)	0.000132	0.000132	0.000131
Between-imputation (B)	0.000015	0.000025	0.000024
Total variance	0.000150	0.000157	0.000156
Fraction of missing information	0.1067	0.1603	0.1557
Efficiency of estimate	0.9791	0.9689	0.9969
Nearest neighbour			
	K=5	*K*=40	*K*=50
Within-imputation (W)	0.000128	0.000128	0.000128
Between-imputation (B)	0.000013	0.000012	0.000011
Total variance	0.000143	0.000140	0.000139
Fraction of missing information	0.0960	0.0861	0.0794
Efficiency of estimate	0.9981	0.9983	0.9984
Predictive mean matching			
	K=5	*K*=40	*K*=50
Within-imputation (W)	0.00013	0.00013	0.000131
Between-imputation (B)	0.00002	0.000014	0.000016
Total variance	0.000154	0.000145	0.000147
Fraction of missing information	0.1409	0.0977	0.1093
Efficiency of estimate	0.9972	0.9981	0.9978
Pattern mixture models			
	K=5	*K*=40	*K*=50
Within-imputation (W)	0.000136	0.000136	0.000136
Between-imputation (B)	0.000015	0.000015	0.000014
Total variance	0.000154	0.000152	0.000151
Fraction of missing information	0.1037	0.0998	0.0937
Efficiency of estimate	0.9979	0.9980	0.9981

Appendix B Estimated Mean Difference Between the Imputed Coping Score and the Missing Coping Score and Estimated Standard Deviation of the Difference Following the Standard Imputation Methods in Simulated Datasets

The estimated mean difference between the imputed coping and the missing coping score following standard imputation methods in the IBCSG dataset is considered in this appendix. The estimated difference between the imputed coping score and the missing coping score is the estimate of the real value of the missing coping score – the imputed coping score.

The figures for i) the estimated mean difference from each simulated dataset and ii) the cumulative estimated mean difference following LOCF is shown in Figure B1.1. This estimated mean difference is summarised for the standard simple imputation methods in Table B1.1. This estimated mean difference estimated following the standard multiple imputation methods varying the number of repetitions of multiple imputation and varying the number of simulated datasets is described in Table B2.1 and Table B2.2 respectively.

A



B

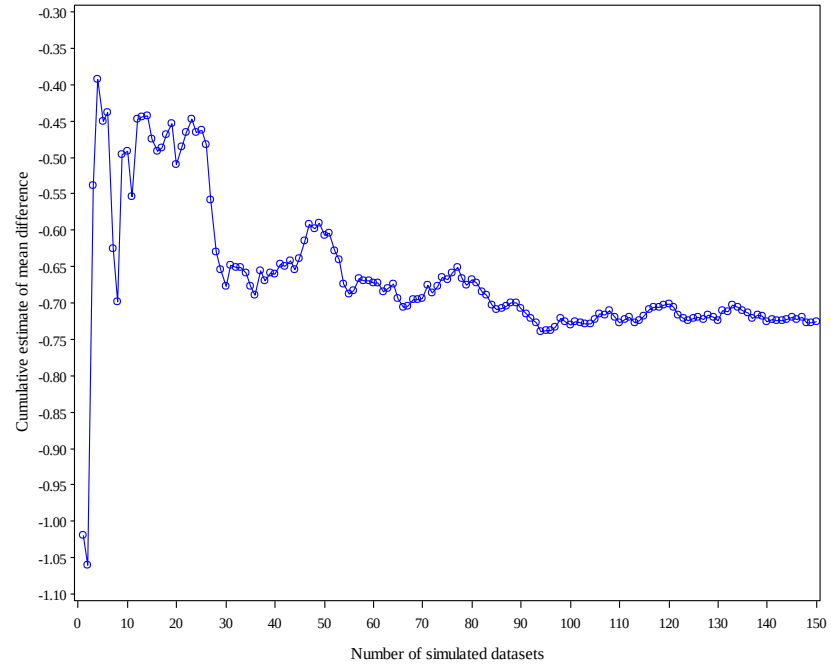


Figure B1.1 Estimated mean difference (A) and the cumulative estimated mean difference (B) between the imputed coping score and the missing coping score following imputation by last observation carried forward by the number of simulated datasets

Table B1.1 Range of Estimated Mean Difference and Estimated Standard Deviation of Estimated Mean Difference Between Imputed Coping Score and Missing Coping Score Following Simple Imputation in Simulated Datasets

Method	Number of Simulated Datasets	Estimated Mean Difference	Range of Estimated Difference	Estimated SD of Difference
LOCF	100	-0.73	-2.55 to 1.12	20.64
	150	-0.73	-2.55 to 1.12	20.64
Median	100	2.09	0.44 to 3.72	18.01
by patient	150	2.09	0.44 to 3.72	18.01
Median	100	11.22	7.99 to 13.19	25.17
by time period	150	11.22	7.99 to 13.19	25.17
Median	100	10.19	7.67 to 12.38	25.00
time period and trt arm	150	10.19	7.67 to 12.40	25.00
Linear regression	100	5.36	3.73 to 6.90	18.35
previous coping scores	150	5.36	3.73 to 6.90	18.35
Linear regression	100	9.83	6.19 to 12.34	26.03
concurrent variables	150	9.83	6.19 to 13.91	26.03

trt = treatment

Estimated difference between the imputed coping score and the missing coping score is the estimate of the real value of the missing coping score – the imputed coping score

Table B2.1 Estimated Mean Difference Between Imputed Coping Score and Missing Coping Score Following a Varying Number of Repetitions of Multiple Imputation in Simulated Datasets

Imputation Method	Simulated Dataset	Estimated Mean Difference (10)	Estimated Mean Difference (20)	Estimated Mean Difference (30)
Bootstrap:	1	3.32	3.47	3.33
baseline coping score	2	4.49	4.51	4.43
	3	4.16	4.03	3.99
	4	2.86	3.10	3.09
	5	4.98	4.87	4.76
Bootstrap:	1	2.97	2.81	2.76
previous coping score	2	3.31	3.31	3.24
	3	3.50	3.47	3.50
	4	2.45	2.59	2.59
	5	3.13	3.38	3.36
Nearest neighbour	1	-0.80	-0.75	-0.76
	2	-0.82	-0.89	-0.84
	3	0.45	0.49	0.51
	4	0.41	0.32	0.31
	5	-0.69	-0.65	-0.60
Predictive mean matching	1	-0.62	-0.72	-0.75
(Initial steps as described	2	-0.61	-0.68	-0.73
for NNI)	3	0.46	0.41	0.41
	4	0.24	0.30	0.32
	5	-0.59	-0.57	-0.56
Pattern mixture models	1	-0.76	-0.72	-0.74
	2	-0.43	-0.52	-0.53
	3	0.41	0.48	0.49
	4	0.40	0.39	0.37
	5	-0.56	-0.51	-0.48

nni = nearest neighbour imputation
Estimated difference between the imputed coping score and the missing coping score is the estimate of the real value of the missing coping score – the imputed coping score
Estimated difference following 10, 20 and 30 repetitions of multiple imputation respectively

Table B2.2 Estimated Mean Difference Between Imputed Coping Score and Missing Coping Score Following Multiple Imputation in a Varying Number of Simulated Datasets

Imputation Method	Number of Simulated Datasets	Estimated Mean Difference	Range of Estimated Mean Difference from Individual Simulated Datasets
Bootstrap:	10	4.01	2.86 to 4.98
baseline coping score	50	3.72	2.02 to 5.19
	75	3.63	1.59 to 5.19
	100	3.60	1.59 to 5.19
Bootstrap:	10	3.09	2.12 to 4.65
previous coping score	50	2.94	1.47 to 4.77
	75	2.93	1.40 to 4.77
	100	2.87	1.07 to 4.77
Nearest neighbour	10	-0.32	-1.43 to 1.16
	50	-0.43	-2.44 to 1.16
	75	-0.49	-2.44 to 1.16
	100	-0.55	-2.44 to 1.16
Predictive mean matching	10	-0.27	-1.51 to 1.20
(Initial steps as described	50	-0.40	-2.33 to 1.20
for NNI)	75	-0.47	-2.33 to 1.20
	100	-0.51	-2.33 to 1.20
Pattern mixture models	10	-0.29	-1.54 to 1.41
	50	-0.42	-2.43 to 1.41
	75	-0.48	-2.43 to 1.41
	100	-0.52	-2.43 to 1.41

nni = nearest neighbour imputation
Estimated difference between the imputed coping score and the missing coping score is the estimate of the real value of the missing coping score – the imputed coping score
Estimated difference following 10 repetitions of multiple imputation respectively

Appendix C Technical Details of Simulated Datasets with a Positive Relationship Between Quality of Life and Disease-Free Survival and a Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival

The technical details of generating the simulated datasets with a positive relationship between quality of life and DFS and a positive relationship between delayed chemotherapy and DFS are described in this appendix. Simulating time to event data (section 4.2.1) is described in Part 1 and artificially removing coping scores from the complete simulated datasets (section 4.2.2) is described in Part 2.

Part 1: Simulating Time to Event Data

The method for simulating a positive relationship between quality of life and DFS and a positive relationship between delayed chemotherapy and DFS according to the algorithm described by MacKenzie and Abrahamowicz (section 4.2.1) is described below:

Algorithm for Randomly Generating Time to Event Data

1a) Start with the actual observed coping scores for the patients in the IBCSG dataset with an observed baseline coping score

2231 patients with an observed baseline coping score are considered

1b) Create a matrix of 9 coping scores for 2231 patients. This is achieved by replacing any missing coping scores or any coping scores when a quality of life assessment was no longer expected by the last known coping score for the patient

1c) Centre the square root of the coping score, S_Pacis, by subtracting the median of square root of the coping scores for the time point

2a) Generate 1338 times to DFS event from a Weibull distribution with shape parameter 1.199 and scale parameter 1519
2b) Generate 893 follow-up times in the trial (censored DFS time) from a Weibull distribution with shape parameter 5.5014 and scale parameter 4997

3) Sort the n (n=2331) simulated DFS times (event or censored) in ascending order
The simulated DFS times were rounded up to the nearest day to avoid a DFS time of 0

4) Identify the risk set of patients who have yet to be matched to a DFS time (at the beginning this is all 2231 patients)

5) For i=1 to n (starting with lowest DFS time, event or censored),
a) for times to DFS event, select the patient to match to the DFS time based on the value of the centred S_Pacis and indicator for delayed chemotherapy from the risk set. For each individual patient in the risk set, the probability of selection is:
(ratio of the hazard of an event at time t for an individual with covariate matrix $\boldsymbol{X}$ vs an individual for whom the covariate matrix $\boldsymbol{\beta}^{T}$ = (0, 0)) / (the sum of the hazard ratio for all patients in the risk set)
b) for the follow-up times in the trial (censored DFS), the ratio of the hazard of a censored DFS event at time t for an individual with covariate matrix $\boldsymbol{X}$ versus an individual for whom the covariate matrix $\boldsymbol{X}^{T}$ = (0, 0) is 1 for all values of the covariate matrix $\boldsymbol{X}$
Therefore, for each individual patient in the risk set, the probability of selection is: 1 / (the number of patients in the risk set)
c) Include this patient and matched DFS time (event or censored) in the simulated dataset of outcome in the IBCSG trial
d) Remove the patient selected from the set of patients at risk

Hazard Function for the Weibull Distribution

The probability density function for the Weibull distribution with shape parameter p and scale parameter q is:

$$f(t) = \left(\frac{p}{q^p}\right) t^{p-1} \exp\left\{-\left(\frac{t}{q}\right)^p\right\} \quad \text{(C1)}$$

The corresponding hazard function at time t for a patient with covariate matrix $\boldsymbol{X}$ is:

$$h(t|\boldsymbol{X}) = \left(\frac{1}{q^p}\right) \exp\{\boldsymbol{\beta}^{\mathrm{T}} X\} p(t^{p-1}) \quad \text{(C2)}$$

where

β_{sp} is the parameter for square root of the coping scores (S_Pacis) in the time-dependent Cox model

β_{del} is the parameter for the delayed chemotherapy in the time-dependent Cox model

$\boldsymbol{X}^{\mathrm{T}}$ = (centred square root of S_Pacis, del_ind)

$\boldsymbol{\beta}^{\mathrm{T}} = (\beta_{sp}, \beta_{del})$

The indicator variable del_ind is 1 if the patient has delayed chemotherapy and 0 if a patient has no delayed chemotherapy

Example C1 Simulated DFS times According to Algorithm from MacKenzie and Abrahamowicz (2002)

For illustration, consider β_{sp} = 0.1 and β_{del} = -0.165. Let the 4 patients in the risk set be:

Table C1 Example Risk Set in Algorithm for Simulating Disease-Free Survival Times

Patient	Coping score	Centred S_Pacis	Indicator for Delayed Chemotherapy
395	13	-0.8666	1
1322	2	-3.0579	1
1467	3	-2.7401	1
1928	15	-0.5992	1

The indicator variable is 1 if the patient has delayed chemotherapy

and let the 4 simulated disease free survival times remaining to be matched (the 4 longest) be:

Table C2 Example Simulated Disease-Free Survival Times to be Matched in Step 5 of Algorithm for Simulating Disease-Free Survival Times

DFS (days)	DFS (censored)
7133	1
7149	1
7151	1
9325	1

Censoring indicator for DFS is 0; DFS= disease-free survival

Consider the time to DFS event of 7133 days. Then the probability of selecting each of the 4 patients to match to the time to DFS event of 7133 days is calculated as in step 5 of the algorithm and was as follows:

Table C3 Example Probability of Selection of Patient in the Risk Set in Step 5 of Algorithm for Simulating Disease-Free Survival Times

Patient	Hazard of an event at time t with covariates $\mathbf{0}$	Hazard of an event at time t with covariates $\boldsymbol{X}^{\mathrm{T}}$	Ratio of hazards at time t	Probability of selection
395	0.00107	0.00083	0.7757	0.2721
1322	0.00107	0.00067	0.6262	0.2197
1467	0.00107	0.00069	0.6449	0.2262
1928	0.00107	0.00086	0.8037	0.2820

Suppose patient 1467 was selected. The simulated disease-free survival for patient 1467 became 7133 days, with censoring indicator 1 for event. Patient 1467 is removed from the risk set.

Creating a Time-Dependent Process

As noted, to create a time-dependent process, the centred square root of the coping score at the appropriate time period in the calculation of the selection probability (section 4.2.1). The appropriate time period is as shown in Table C4:

Table C4 Time Period of Centred Square Root Coping Score Considered when Calculating the Hazard at Time t for a Patient with Covariates X

DFS (days)	Time Period
0-91	Baseline (Time 1)
92-182	Month 3 (Time 2)
183-273	Month 6 (Time 3)
274-364	Month 9 (Time 4)
365-455	Month 12 (Time 5)
456-546	Month 15 (Time 6)
547-637	Month 18 (Time 7)
638-728	Month 21 (Time 8)
$\geq$ 729	Month 24 (Time 9)

Therefore in Example C1 the centred square root of the coping score at Month 24 (Time 9) was considered.

Part 2: Artificially Removing Data from the Complete Simulated Datasets

There were 600 complete simulated datasets generated (150 x 4 combinations of β_{sp} and β_{del}) (Figure 4.1). As noted, for each of the complete simulated datasets 5 different methods of artificially removing coping scores were considered. The details of the 5 methods of artificially removing coping scores were as follows:

Method 1: Higher coping scores (lower quality of life) have a higher chance of being missing

For each coping score, generate a random number, RTerm, from the Uniform(0, 1) distribution. As the quality of life increases (coping score decreases) the probability of the coping score being observed increases. Patients with the highest quality of life (coping score $\leq$ 10) have at least a 95% probability of coping score being observed. In contrast, patients with poor quality of life (coping score > 60) have a 50% probability of a coping score being observed.

For 0 $\leq$ coping scores $\leq$ 2, the coping score is set to missing if RTerm is $\leq$ 0.025
For 3 $\leq$ coping scores $\leq$ 10, the coping score is set to missing if RTerm is $\leq$ 0.05

For 11 ≤ coping scores ≤ 20, the coping score is set to missing if RTerm is ≤ 0.125

For 21 ≤ coping scores ≤ 39, the coping score is set to missing if RTerm is ≤ 0.2

For 40 ≤ coping scores ≤ 49, the coping score is set to missing if RTerm is ≤ 0.35

For 50 ≤ coping scores ≤ 60, the coping score is set to missing if RTerm is ≤ 0.425

For 61 ≤ coping scores ≤ 100, the coping score is set to missing if RTerm is ≤ 0.5

Method 2: Lower coping scores (higher quality of life) have a higher chance of being missing

For each coping score, generate an random number, RTerm, from the Uniform(0, 1) distribution. As the quality of life increases (coping score decreases) the probability of the coping score being observed decreases. Patients with the highest quality of life (coping score ≤ 10) have at most a 50% probability of coping score being observed. In contrast, patients with poor quality of life (coping score > 60) have a 99% probability of a coping score being observed.

For 0 ≤ coping scores ≤ 2, the coping score is set to missing if RTerm is ≤ 0.65

For 3 ≤ coping scores ≤ 5, the coping score is set to missing if RTerm is ≤ 0.575

For 6 ≤ coping scores ≤ 10, the coping score is set to missing if RTerm is ≤ 0.5

For 11 ≤ coping scores ≤ 20, the coping score is set to missing if RTerm is ≤ 0.35

For 21 ≤ coping scores ≤ 39, the coping score is set to missing if RTerm is ≤ 0.275

For 40 ≤ coping scores ≤ 60, the coping score is set to missing if RTerm is ≤ 0.2

For 61 ≤ coping scores ≤ 100, the coping score is set to missing if RTerm is ≤ 0.01

Method 3: Later time periods have a higher chance of being missing

For each coping score, generate an random number, RTerm, from the Uniform(0, 1) distribution. As the time in study increases the probability of the coping score being observed decreases. At baseline, patients have an 80% probability of coping

score being observed. In contrast, at Month 24 patients have a 50% probability of a coping score being observed.

For baseline (approx randomisation), the coping score is set to missing if RTerm is ≤ 0.2
For Month 3, the coping score is set to missing if RTerm is ≤ 0.225
For Month 6, the coping score is set to missing if RTerm is ≤ 0.25
For Month 9, the coping score is set to missing if RTerm is ≤ 0.30
For Month 12, the coping score is set to missing if RTerm is ≤ 0.35
For Month 15, the coping score is set to missing if RTerm is ≤ 0.4
For Month 18, the coping score is set to missing if RTerm is ≤ 0.45
For Month 21, the coping score is set to missing if RTerm is ≤ 0.475
For Month 24, the coping score is set to missing if RTerm is ≤ 0.5

Method 4: Artificially remove approximately 30% of coping scores at random

For each coping score, generate an random number, RTerm, from the Uniform(0, 1) distribution

If RTerm is ≤ 0.3, set the coping score to missing

Method 5: Higher coping scores have a higher chance of being missing

As in Method 1, the quality of life increases (coping score decreases) the probability of the coping score being observed increases. However in Method 5, the coping scores are considered on a continuous scale rather than grouped into categories.

The initial step is, for each coping score, generate an random number, RTerm, from the Uniform(0, 1) distribution.

Let the variable Formula = ((Cope_status + 10) / 200) + RTerm, where Cope_status is the coping score

The ratio term will be between 10/200 = 0.05 and 110/200 = 0.55. When adding a random term between 0 and 1, the value of the formula will be above 0.925 approximately 30% of the time. Therefore the rule was:

If Formula > 0.925, set the coping score to missing

Appendix D Schoenfeld Residuals from Time-Dependent Cox Model Analysis for Different Combinations of a Positive Relationship Between Quality of Life and Disease-Free Survival and a Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival

The plots of Schoenfeld residuals against time for the square root of coping score (S_Pacis) and delayed chemotherapy for the 4 for complete simulated datasets with different combinations of a positive relationship between quality of life are shown in this appendix. The Schoenfeld residuals were calculated from the time-dependent Cox model analysis on S_Pacis and delayed chemotherapy stratified by trial. The first complete simulated dataset for each of the different combinations of a positive relationship between quality of life and DFS and a positive relationship between delayed chemotherapy and DFS is considered, as shown below:

Combination of β_{sp} and β_{del}	S_Pacis	Delayed Chemotherapy
Weak, weak	Figure D1.1	Figure D1.2
Weak, strong	Figure D2.1	Figure D2.2
Strong, weak	Figure D3.1	Figure D3.2
Strong, strong	Figure D4.1	Figure D4.2

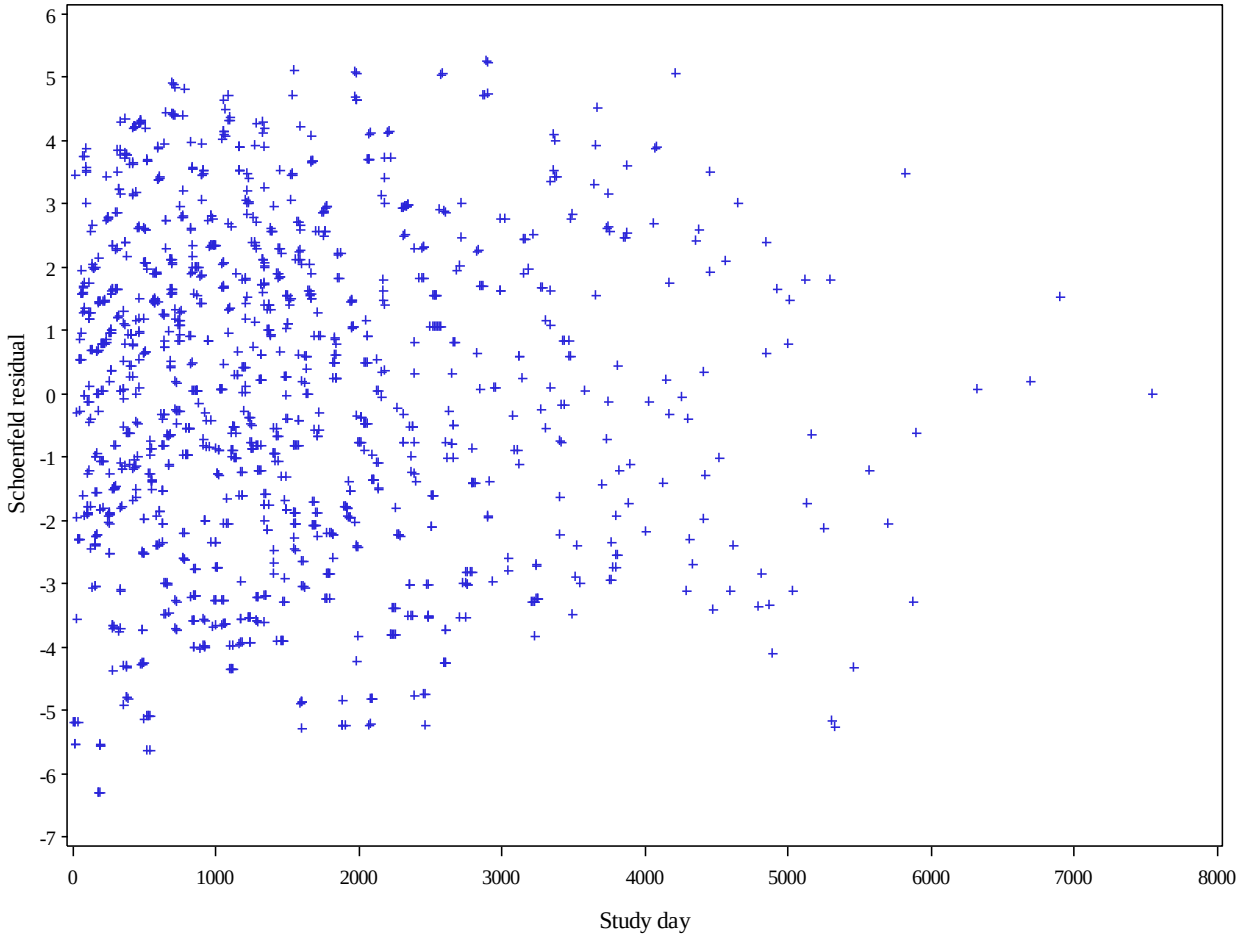


Figure D1.1 Schoenfeld residuals against time for the square root of the coping score (S_Pacis) from the time-dependent Cox model analysis for the first complete simulated dataset with weak relationship between quality of life and disease-free survival and weak relationship between delayed chemotherapy and disease-free survival

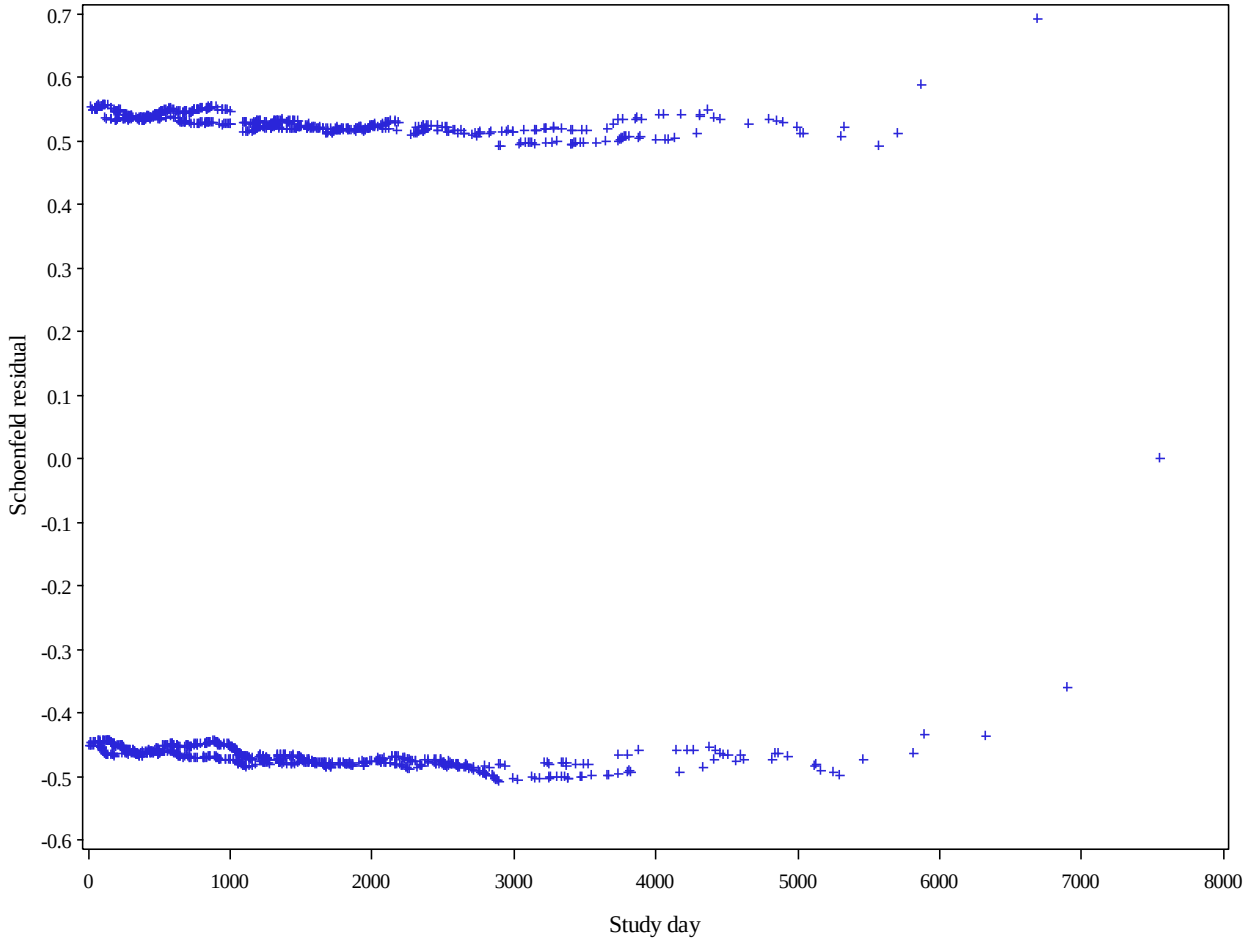


Figure D1.2 Schoenfeld residuals against time for delayed chemotherapy from the time-dependent Cox model analysis for the first complete simulated dataset with weak relationship between quality of life and disease-free survival and weak relationship between delayed chemotherapy and disease-free survival

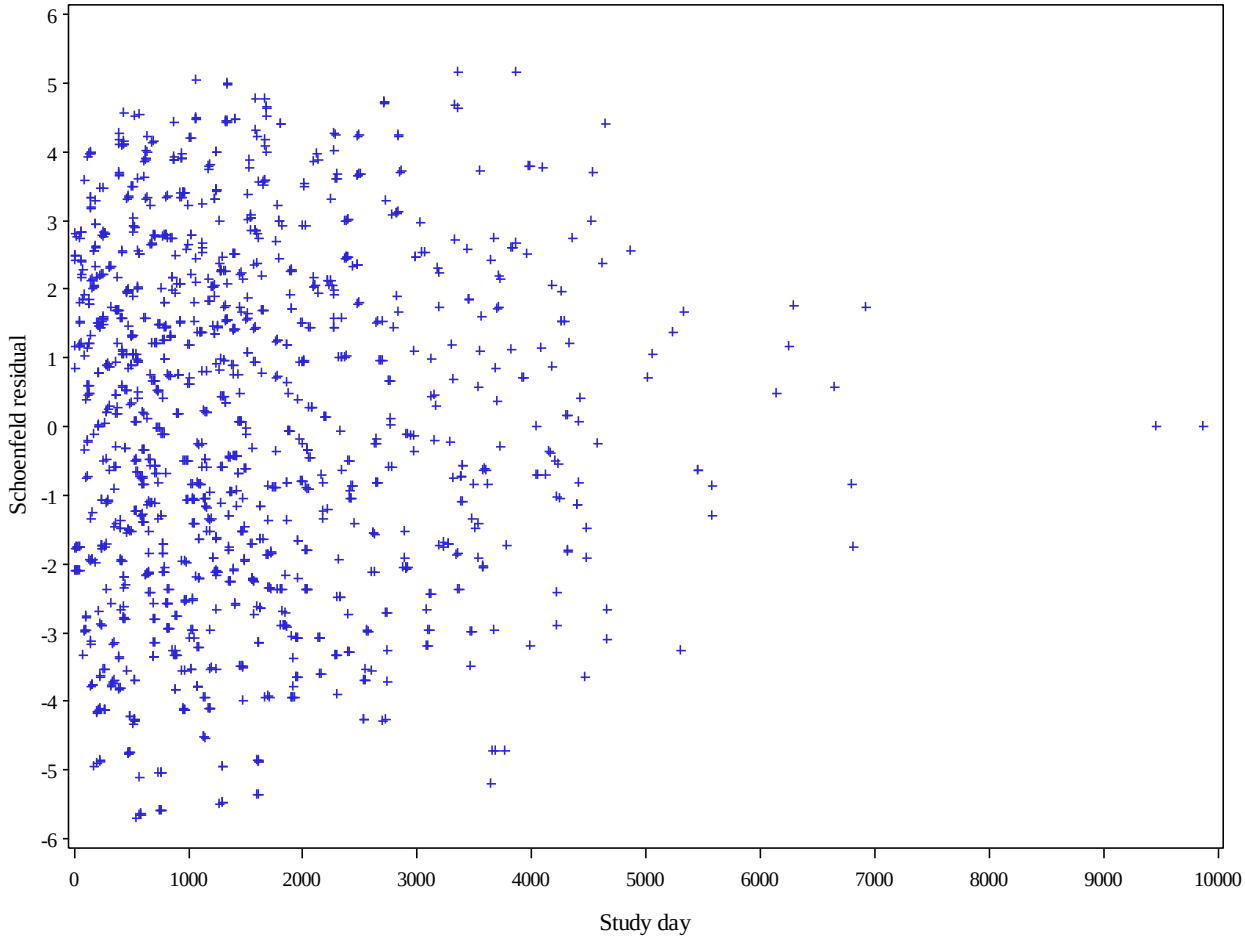


Figure D2.1 Schoenfeld residuals against time for the square root of the coping score (S_Pacis) from the time-dependent Cox model analysis for the first complete simulated dataset with weak relationship between quality of life and disease-free survival and strong relationship between delayed chemotherapy and disease-free survival

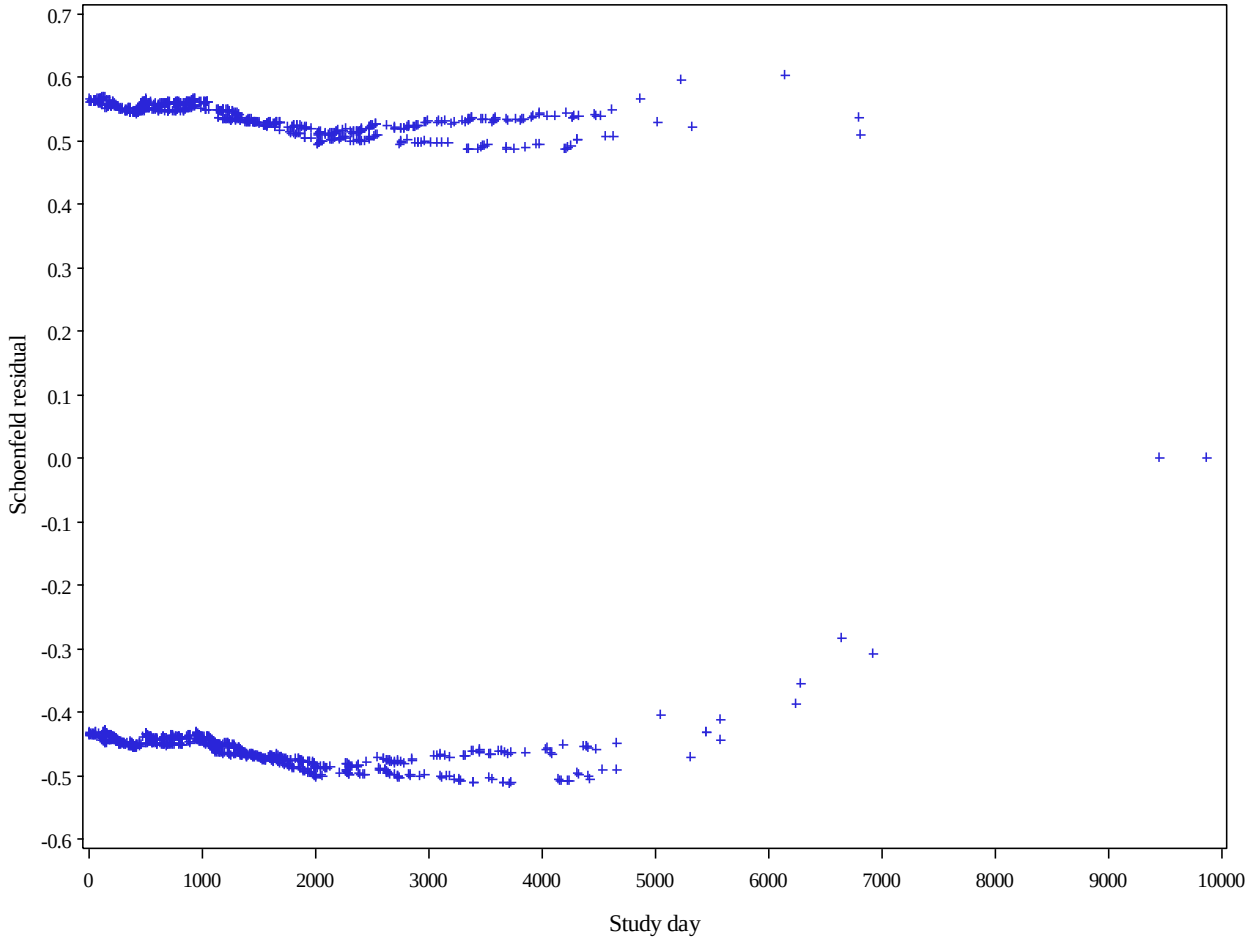


Figure D2.2 Schoenfeld residuals against time for delayed chemotherapy from the time-dependent Cox model analysis for the first complete simulated dataset with weak relationship between quality of life and disease-free survival and strong relationship between delayed chemotherapy and disease-free survival

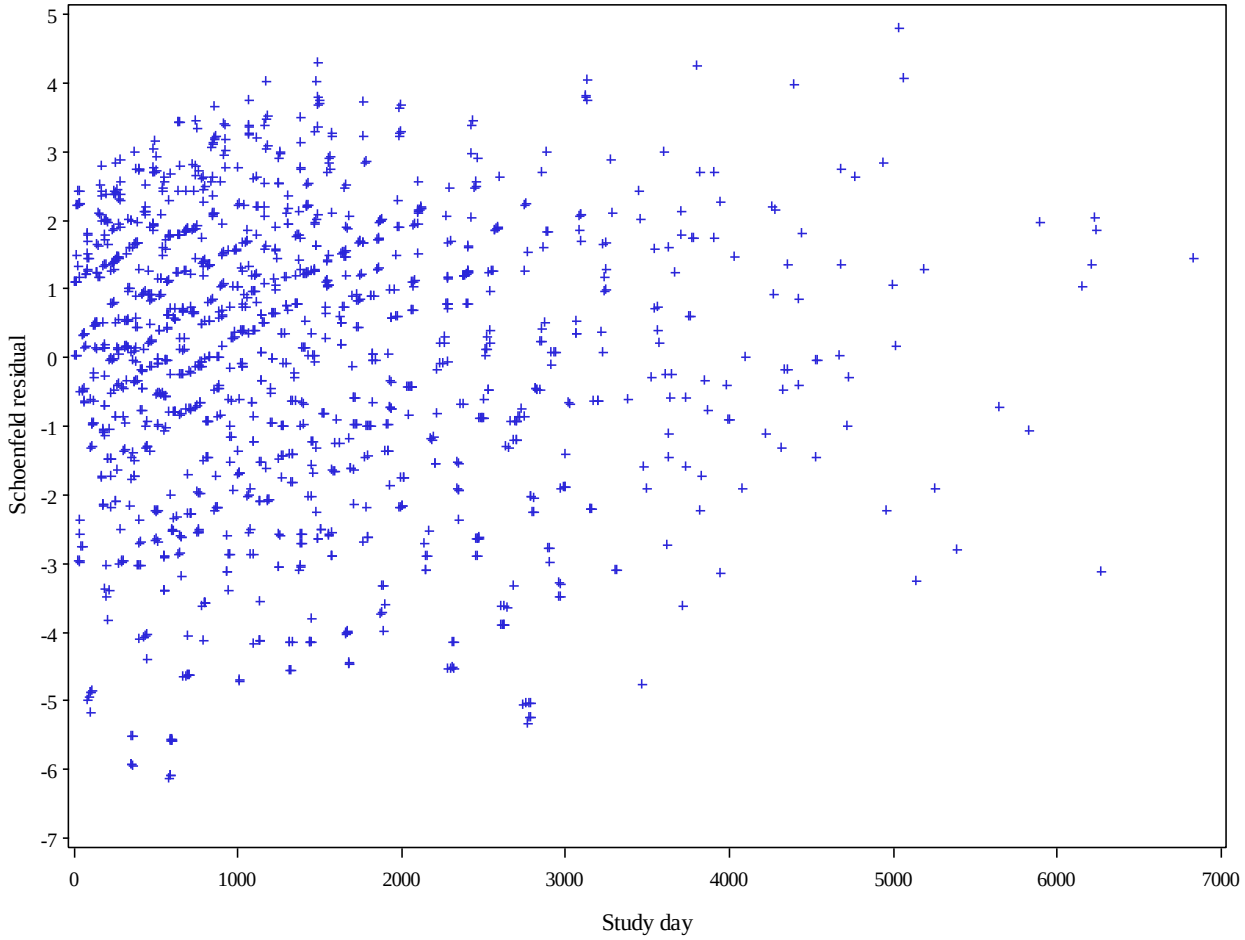


Figure D3.1 Schoenfeld residuals against time for the square root of the coping score (S_Pacis) from the time-dependent Cox model analysis for the first complete simulated dataset with strong relationship between quality of life and disease-free survival and weak relationship between delayed chemotherapy and disease-free survival

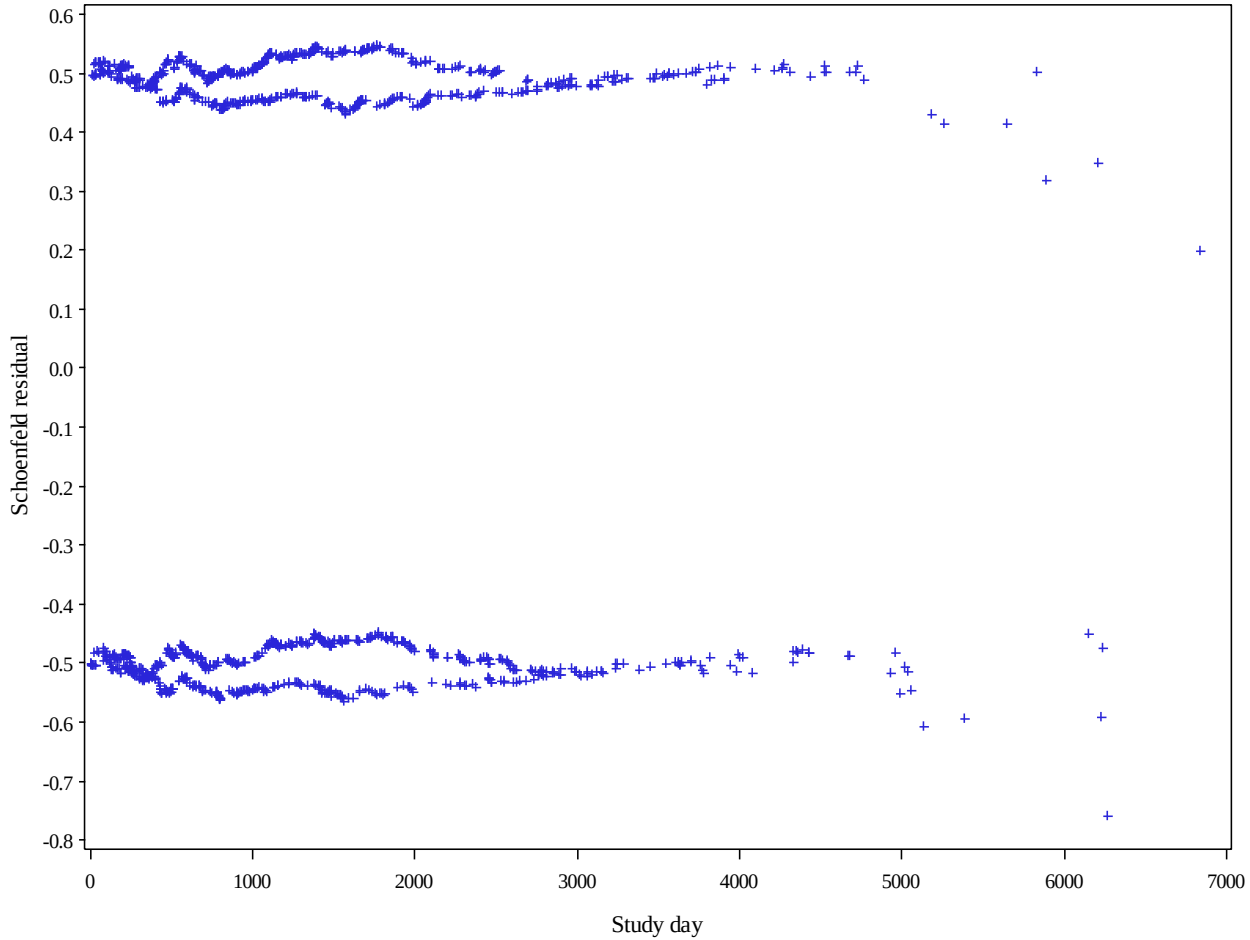


Figure D3.2 Schoenfeld residuals against time for delayed chemotherapy from the time-dependent Cox model analysis for the first complete simulated dataset with strong relationship between quality of life and disease-free survival and weak relationship between delayed chemotherapy and disease-free survival

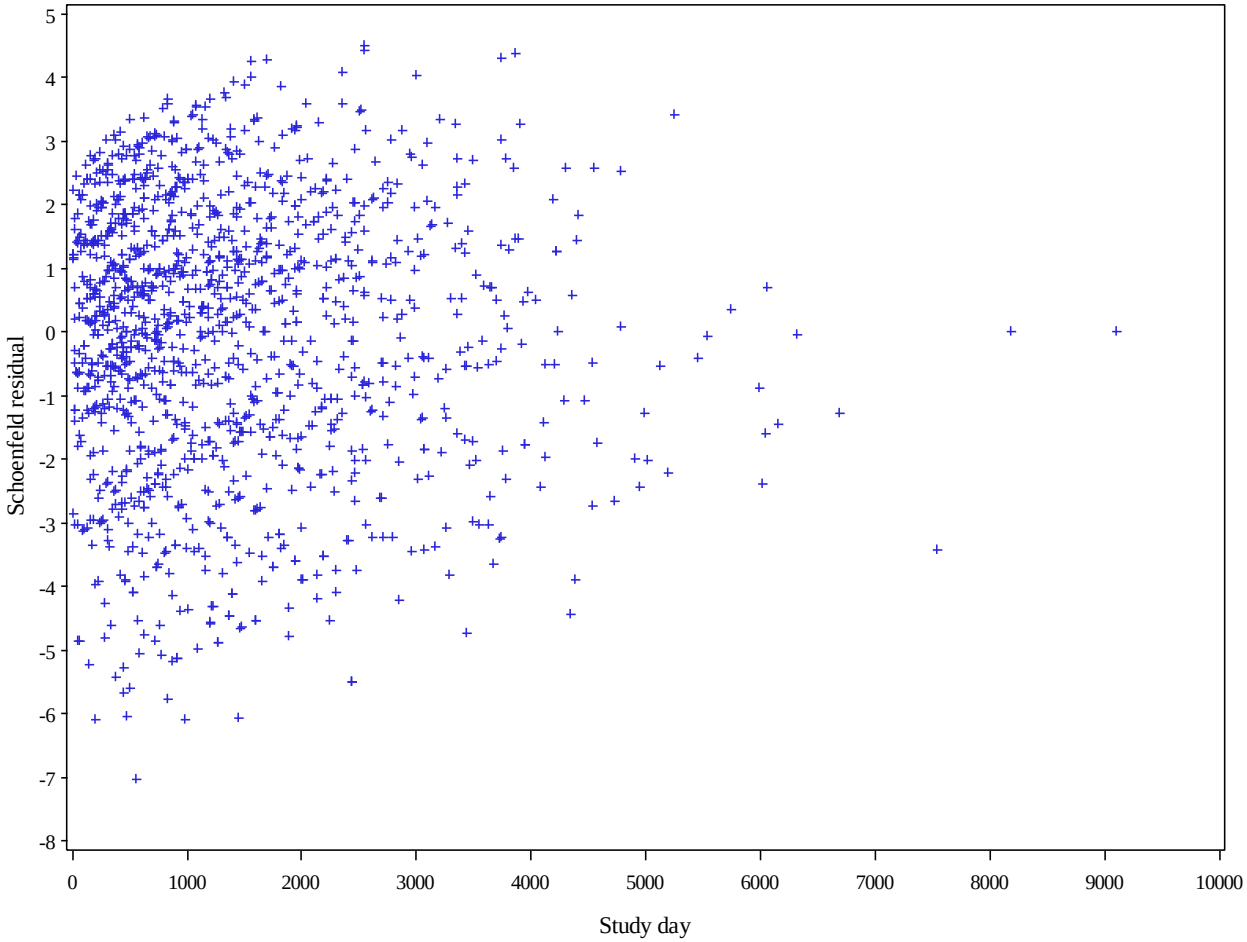


Figure D4.1 Schoenfeld residuals against time for the square root of the coping score (S_Pacis) from the time-dependent Cox model analysis for the first complete simulated dataset with strong relationship between quality of life and disease-free survival and strong relationship between delayed chemotherapy and disease-free survival

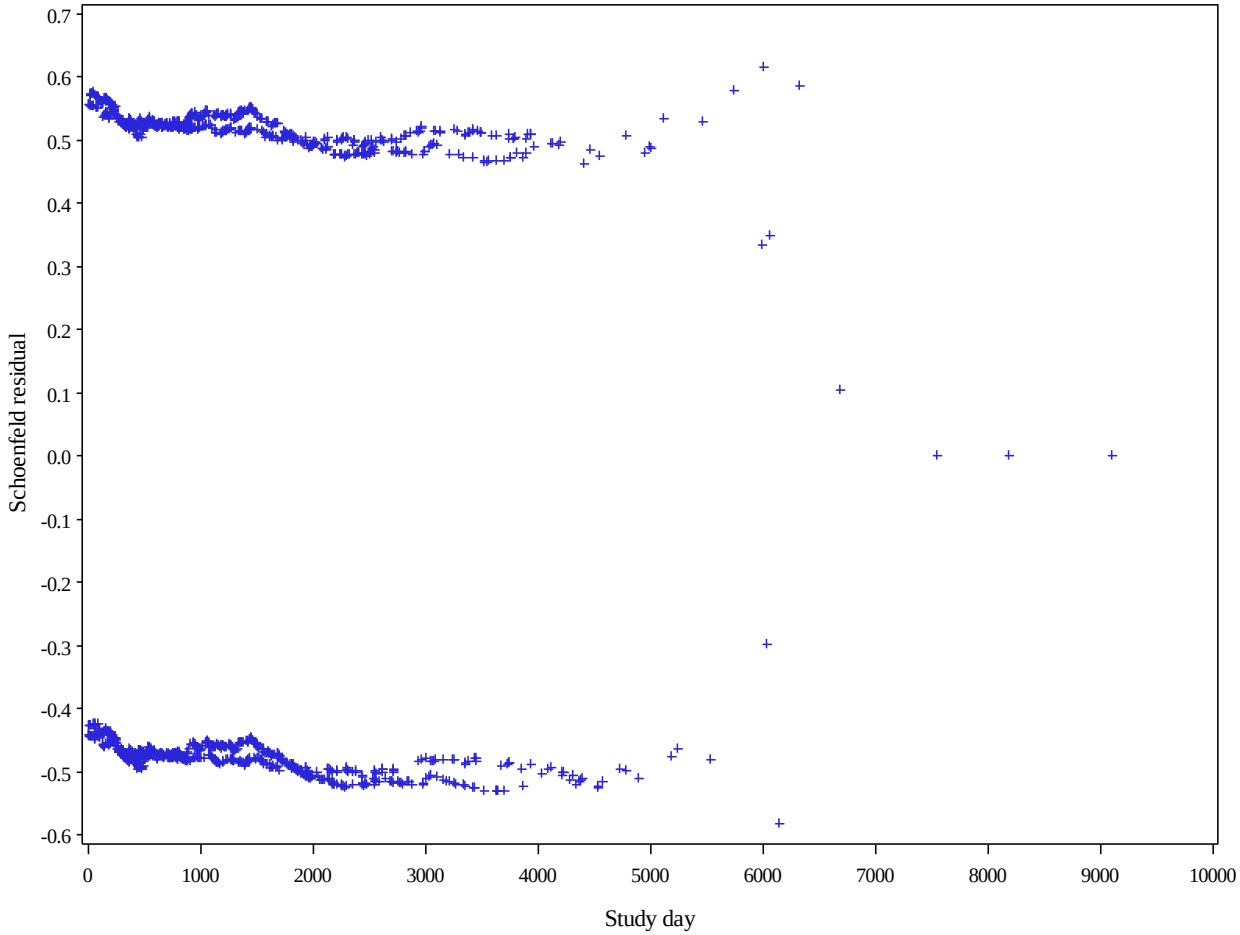


Figure D4.2 Schoenfeld residuals against time for delayed chemotherapy from the time-dependent Cox model analysis for the first complete simulated dataset with strong relationship between quality of life and disease-free survival and strong relationship between delayed chemotherapy and disease-free survival

Appendix E Results from Time-Dependent Cox Model Analysis for Different Combinations of a Positive Relationship Between Quality of Life and Disease-Free Survival and a Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival Following Simple Imputation of 150 Simulated Datasets with Coping Scores Artificially Removed

The results of the time-dependent Cox model analysis on the square root of coping score (S_Pacis) and delayed chemotherapy stratified by trial for different combinations of a positive relationship between quality of life and DFS and a positive relationship between delayed chemotherapy and DFS following simple imputation of 150 simulated datasets with coping scores artificially removed are shown in this appendix. The mean hazard ratio is calculated from the hazard ratios of the simulated completed datasets and is the exponential of the mean parameter estimate. These results for the different combinations of a positive relationship between quality of life and DFS and a positive relationship between delayed chemotherapy and DFS are shown as noted below:

Combination of β_{sp} and β_{del}	Estimate of β_{sp}	Estimate of β_{del}
Weak, weak	Table E1.1	Table E1.2
Weak, strong	Table E2.1	Table E2.2
Strong, weak	Table E3.1	Table E3.2
Strong, strong	Table E4.1	Table E4.2

Table E1.1 Summary of Square Root of Coping Score (S_Pacis) from Time-Dependent Cox Model Analysis Stratified by Trial

Weak Positive Relationship Between Quality of Life and Disease-Free Survival and Weak Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival

Imputation Method	Method of Artificially Removing Coping Scores	Mean Parameter Estimate	Mean Standard Error	Bias (%): Theoretical Parameter Value = 0.1	Number of 95% CIs for hazard ratio containing 1	Number of 95% CIs for Parameter Estimate Containing Simulated Value
LOCF	Method 1	0.0969	0.0133	3.1	0 (0.0%)	150 (100.0%)
	Method 2	0.0950	0.0134	5.0	0 (0.0%)	149 (99.3%)
	Method 3	0.0914	0.0124	8.6	0 (0.0%)	147 (98.0%)
	Method 4	0.0958	0.0132	4.2	0 (0.0%)	150 (100.0%)
	Method 5	0.0948	0.0138	5.2	0 (0.0%)	149 (99.3%)
Median imputation:	Method 1	0.0993	0.0114	0.7	0 (0.0%)	150 (100.0%)
by patient	Method 2	0.1066	0.0120	6.6	0 (0.0%)	150 (100.0%)
	Method 3	0.0988	0.0116	1.2	0 (0.0%)	150 (100.0%)
	Method 4	0.1030	0.0114	3.0	0 (0.0%)	150 (100.0%)
	Method 5	0.1000	0.0115	0.0	0 (0.0%)	150 (100.0%)
Linear regression:	Method 1	0.1025	0.0140	2.5	0 (0.0%)	150 (100.0%)
previous coping score	Method 2	0.1025	0.0141	2.5	0 (0.0%)	150 (100.0%)
	Method 3	0.1009	0.0134	0.9	0 (0.0%)	150 (100.0%)
	Method 4	0.1019	0.0139	1.9	0 (0.0%)	150 (100.0%)
	Method 5	0.1020	0.0148	2.0	0 (0.0%)	149 (99.3%)

Method 1: Higher coping scores have a higher chance of being missing.
Method 2: Lower coping scores have a higher chance of being missing.
Method 3: Later time period have a higher chance of being missing
Method 4: 30% of coping scores missing at random
Method 5: Higher coping scores have a higher chance of being missing

Table E1.2 Summary of Delayed Chemotherapy from Time-Dependent Cox Model Analysis Stratified by Trial

Weak Positive Relationship Between Quality of Life and Disease-Free Survival and Weak Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival

Imputation Method	Method of Artificially Removing Coping Scores	Mean Parameter Estimate	Mean Standard Error	Bias (%): Theoretical Parameter Value = -0.165	Number of 95% CIs for hazard ratio containing 1	Number of 95% CIs for Parameter Estimate Containing Simulated Value
LOCF	Method 1	-0.1641	0.0655	0.5	51 (34.0%)	150 (100.0%)
	Method 2	-0.1689	0.0631	2.3	49 (32.7%)	150 (100.0%)
	Method 3	-0.1714	0.0614	3.9	49 (32.7%)	150 (100.0%)
	Method 4	-0.1730	0.0656	4.9	47 (31.3%)	150 (100.0%)
	Method 5	-0.1722	0.0675	4.3	57 (38.0%)	150 (100.0%)
Median imputation: by patient	Method 1	-0.1700	0.0554	3.0	43 (28.7%)	150 (100.0%)
	Method 2	-0.1732	0.0552	5.0	42 (28.0%)	150 (100.0%)
	Method 3	-0.1743	0.0551	5.6	40 (26.7%)	150 (100.0%)
	Method 4	-0.1741	0.0553	5.5	41 (27.3%)	150 (100.0%)
	Method 5	-0.1725	0.0555	4.5	41 (27.3%)	150 (100.0%)
Linear regression: previous coping score	Method 1	-0.1655	0.0655	0.3	49 (32.7%)	150 (100.0%)
	Method 2	-0.1689	0.0631	2.4	49 (32.7%)	150 (100.0%)
	Method 3	-0.1722	0.0614	4.4	42 (28.0%)	150 (100.0%)
	Method 4	-0.1742	0.0656	5.6	47 (31.3%)	150 (100.0%)
	Method 5	-0.1738	0.0675	5.3	55 (36.7%)	150 (100.0%)

Method 1: Higher coping scores have a higher chance of being missing.
Method 2: Lower coping scores have a higher chance of being missing.
Method 3: Later time period have a higher chance of being missing
Method 4: 30% of coping scores missing at random
Method 5: Higher coping scores have a higher chance of being missing

Table E2.1 Summary of Square Root of Coping Score (S_Pacis) from Time-Dependent Cox Model Analysis Stratified by Trial

Weak Positive Relationship Between Quality of Life and Disease-Free Survival and Strong Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival

Imputation Method	Method of Artificially Removing Coping Scores	Mean Parameter Estimate	Mean Standard Error	Bias (%): Theoretical Parameter Value = 0.1	Number of 95% CIs for hazard ratio containing 1	Number of 95% CIs for Parameter Estimate Containing Simulated Value
LOCF	Method 1	0.0992	0.0133	0.8	0 (0.0%)	150 (100.0%)
	Method 2	0.0968	0.0134	3.2	0 (0.0%)	150 (100.0%)
	Method 3	0.0932	0.0124	6.8	0 (0.0%)	147 (98.0%)
	Method 4	0.0969	0.0133	3.1	0 (0.0%)	149 (99.3%)
	Method 5	0.0980	0.0138	2.0	0 (0.0%)	149 (99.3%)
Median imputation:	Method 1	0.1006	0.0114	0.6	0 (0.0%)	150 (100.0%)
by patient	Method 2	0.1081	0.0120	8.1	0 (0.0%)	150 (100.0%)
	Method 3	0.1000	0.0117	0.0	0 (0.0%)	150 (100.0%)
	Method 4	0.1046	0.0115	4.6	0 (0.0%)	150 (100.0%)
	Method 5	0.1011	0.0115	1.1	0 (0.0%)	150 (100.0%)
Linear regression:	Method 1	0.1051	0.0141	5.1	0 (0.0%)	150 (100.0%)
previous coping score	Method 2	0.1039	0.0141	3.9	0 (0.0%)	149 (99.3%)
	Method 3	0.1025	0.0135	2.5	0 (0.0%)	150 (100.0%)
	Method 4	0.1028	0.0139	2.8	0 (0.0%)	149 (99.3%)
	Method 5	0.1048	0.0147	4.8	0 (0.0%)	149 (99.3%)

Method 1: Higher coping scores have a higher chance of being missing.
Method 2: Lower coping scores have a higher chance of being missing.
Method 3: Later time period have a higher chance of being missing
Method 4: 30% of coping scores missing at random
Method 5: Higher coping scores have a higher chance of being missing

Table E2.2 Summary of Delayed Chemotherapy from Time-Dependent Cox Model Analysis Stratified by Trial

Weak Positive Relationship Between Quality of Life and Disease-Free Survival and Strong Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival

Imputation Method	Method of Artificially Removing Coping Scores	Mean Parameter Estimate	Mean Standard Error	Bias (%): Theoretical Parameter Value = -0.195	Number of 95% CIs for hazard ratio containing 1	Number of 95% CIs for Parameter Estimate Containing Simulated Value
LOCF	Method 1	-0.1781	0.0656	8.6	41 (27.3%)	150 (100.0%)
	Method 2	-0.1864	0.0631	4.4	31 (20.7%)	150 (100.0%)
	Method 3	-0.1899	0.0615	2.6	26 (17.3%)	150 (100.0%)
	Method 4	-0.1891	0.0657	3.0	35 (23.3%)	150 (100.0%)
	Method 5	-0.1877	0.0674	3.7	36 (24.0%)	149 (99.3%)
Median imputation: by patient	Method 1	-0.1850	0.0554	5.2	20 (13.3%)	150 (100.0%)
	Method 2	-0.1884	0.0553	3.4	21 (14.0%)	150 (100.0%)
	Method 3	-0.1888	0.0552	3.2	20 (13.3%)	150 (100.0%)
	Method 4	-0.1892	0.0553	3.0	20 (13.3%)	150 (100.0%)
	Method 5	-0.1852	0.0555	5.0	23 (15.3%)	150 (100.0%)
Linear regression: previous coping score	Method 1	-0.1795	0.0656	8.0	40 (26.7%)	150 (100.0%)
	Method 2	-0.1862	0.0631	4.5	31 (20.7%)	150 (100.0%)
	Method 3	-0.1907	0.0615	2.2	26 (17.3%)	150 (100.0%)
	Method 4	-0.1899	0.0657	2.6	34 (22.7%)	150 (100.0%)
	Method 5	-0.1893	0.0674	2.9	36 (24.0%)	149 (99.3%)

Method 1: Higher coping scores have a higher chance of being missing.
Method 2: Lower coping scores have a higher chance of being missing.
Method 3: Later time period have a higher chance of being missing
Method 4: 30% of coping scores missing at random
Method 5: Higher coping scores have a higher chance of being missing

Table E3.1 Summary of Square Root of Coping Score (S_Pacis) from Time-Dependent Cox Model Analysis Stratified by Trial

Strong Positive Relationship Between Quality of Life and Disease-Free Survival and Weak Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival

Imputation Method	Method of Artificially Removing Coping Scores	Mean Parameter Estimate	Mean Standard Error	Bias (%): Theoretical Parameter Value = 0.4	Number of 95% CIs for hazard ratio containing 1	Number of 95% CIs for Parameter Estimate Containing Simulated Value
LOCF	Method 1	0.3726	0.0153	6.9	0 (0.0%)	78 (52.0%)
	Method 2	0.3758	0.0154	6.0	0 (0.0%)	93 (62.0%)
	Method 3	0.3512	0.0142	12.2	0 (0.0%)	5 (3.3%)
	Method 4	0.3684	0.0154	7.9	0 (0.0%)	56 (37.3%)
	Method 5	0.3681	0.0159	8.0	0 (0.0%)	63 (42.0%)
Median imputation: by patient	Method 1	0.3792	0.0132	5.2	0 (0.0%)	101 (67.3%)
	Method 2	0.3881	0.0139	3.0	0 (0.0%)	142 (94.7%)
	Method 3	0.3667	0.0135	8.3	0 (0.0%)	27 (18.0%)
	Method 4	0.3825	0.0134	4.4	0 (0.0%)	114 (76.0%)
	Method 5	0.3775	0.0134	5.6	0 (0.0%)	89 (59.3%)
Linear regression: previous coping score	Method 1	0.3930	0.0162	1.7	0 (0.0%)	149 (99.3%)
	Method 2	0.4082	0.0162	2.1	0 (0.0%)	147 (98.0%)
	Method 3	0.3910	0.0155	2.3	0 (0.0%)	145 (96.7%)
	Method 4	0.3942	0.0162	1.5	0 (0.0%)	150 (100.0%)
	Method 5	0.3929	0.0171	1.8	0 (0.0%)	146 (97.3%)

Method 1: Higher coping scores have a higher chance of being missing.
Method 2: Lower coping scores have a higher chance of being missing.
Method 3: Later time period have a higher chance of being missing
Method 4: 30% of coping scores missing at random
Method 5: Higher coping scores have a higher chance of being missing

Table E3.2 Summary of Delayed Chemotherapy from Time-Dependent Cox Model Analysis Stratified by Trial

Strong Positive Relationship Between Quality of Life and Disease-Free Survival and Weak Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival

Imputation Method	Method of Artificially Removing Coping Scores	Mean Parameter Estimate	Mean Standard Error	Bias (%): Theoretical Parameter Value = -0.165	Number of 95% CIs for hazard ratio containing 1	Number of 95% CIs for Parameter Estimate Containing Simulated Value
LOCF	Method 1	-0.1345	0.0667	18.5	67 (44.7%)	149 (99.3%)
	Method 2	-0.1549	0.0620	6.1	51 (34.0%)	150 (100.0%)
	Method 3	-0.1514	0.0614	8.2	52 (34.7%)	149 (99.3%)
	Method 4	-0.1476	0.0657	10.5	64 (42.7%)	150 (100.0%)
	Method 5	-0.1405	0.0686	14.9	67 (44.7%)	149 (99.3%)
Median imputation: by patient	Method 1	-0.1460	0.0556	11.5	46 (30.7%)	150 (100.0%)
	Method 2	-0.1517	0.0551	8.1	43 (28.7%)	150 (100.0%)
	Method 3	-0.1574	0.0551	4.6	38 (25.3%)	150 (100.0%)
	Method 4	-0.1559	0.0553	5.5	39 (26.0%)	150 (100.0%)
	Method 5	-0.1482	0.0557	10.2	44 (29.3%)	150 (100.0%)
Linear regression: previous coping score	Method 1	-0.1393	0.0667	15.6	61 (40.7%)	149 (99.3%)
	Method 2	-0.1547	0.0620	6.2	51 (34.0%)	150 (100.0%)
	Method 3	-0.1544	0.0614	6.4	51 (34.0%)	149 (99.3%)
	Method 4	-0.1510	0.0657	8.5	63 (42.0%)	150 (100.0%)
	Method 5	-0.1466	0.0686	11.2	63 (42.0%)	149 (99.3%)

Method 1: Higher coping scores have a higher chance of being missing.
Method 2: Lower coping scores have a higher chance of being missing.
Method 3: Later time period have a higher chance of being missing
Method 4: 30% of coping scores missing at random
Method 5: Higher coping scores have a higher chance of being missing

Table E4.1 Summary of Square Root of Coping Score (S_Pacis) from Time-Dependent Cox Model Analysis Stratified by Trial

Strong Positive Relationship Between Quality of Life and Disease-Free Survival and Strong Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival

Imputation Method	Method of Artificially Removing Coping Scores	Mean Parameter Estimate	Mean Standard Error	Bias (%): Theoretical Parameter Value = 0.4	Number of 95% CIs for hazard ratio containing 1	Number of 95% CIs for Parameter Estimate Containing Simulated Value
LOCF	Method 1	0.3740	0.0153	6.5	0 (0.0%)	76 (50.7%)
	Method 2	0.3749	0.0154	6.3	0 (0.0%)	85 (56.7%)
	Method 3	0.3512	0.0142	12.2	0 (0.0%)	1 (0.7%)
	Method 4	0.3709	0.0154	7.3	0 (0.0%)	62 (41.3%)
	Method 5	0.3698	0.016	7.5	0 (0.0%)	67 (44.7%)
Median imputation:	Method 1	0.3785	0.0132	5.4	0 (0.0%)	87 (58.0%)
by patient	Method 2	0.3879	0.0139	3.0	0 (0.0%)	142 (94.7%)
	Method 3	0.3657	0.0135	8.6	0 (0.0%)	17 (11.3%)
	Method 4	0.3828	0.0136	4.3	0 (0.0%)	125 (83.3%)
	Method 5	0.3773	0.0134	5.7	0 (0.0%)	80 (53.3%)
Linear regression:	Method 1	0.3947	0.0163	1.3	0 (0.0%)	147 (98.0%)
previous coping score	Method 2	0.4070	0.0162	1.7	0 (0.0%)	150 (100.0%)
	Method 3	0.3914	0.0155	2.1	0 (0.0%)	144 (96.0%)
	Method 4	0.3953	0.0162	1.2	0 (0.0%)	149 (99.3%)
	Method 5	0.3948	0.0171	1.3	0 (0.0%)	149 (99.3%)

Method 1: Higher coping scores have a higher chance of being missing.
Method 2: Lower coping scores have a higher chance of being missing.
Method 3: Later time period have a higher chance of being missing
Method 4: 30% of coping scores missing at random
Method 5: Higher coping scores have a higher chance of being missing

Table E4.2 Summary of Delayed Chemotherapy from Time-Dependent Cox Model Analysis Stratified by Trial

Strong Positive Relationship Between Quality of Life and Disease-Free Survival and Strong Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival

Imputation Method	Method of Artificially Removing Coping Scores	Mean Parameter Estimate	Mean Standard Error	Bias (%): Theoretical Parameter Value = -0.195	Number of 95% CIs for hazard ratio containing 1	Number of 95% CIs for Parameter Estimate Containing Simulated Value
LOCF	Method 1	-0.1803	0.0668	7.5	40 (26.7%)	150 (100.0%)
	Method 2	-0.1881	0.0620	3.6	36 (24.0%)	150 (100.0%)
	Method 3	-0.1853	0.0614	5.0	32 (21.3%)	150 (100.0%)
	Method 4	-0.1863	0.0657	4.5	35 (23.3%)	149 (99.3%)
	Method 5	-0.1775	0.0686	9.0	44 (29.3%)	150 (100.0%)
Median imputation:	Method 1	-0.1830	0.0556	6.2	25 (16.7%)	150 (100.0%)
by patient	Method 2	-0.1862	0.0551	4.5	23 (15.3%)	150 (100.0%)
	Method 3	-0.1930	0.0552	1.0	20 (13.3%)	150 (100.0%)
	Method 4	-0.1956	0.0562	0.3	24 (16.0%)	150 (100.0%)
	Method 5	-0.1852	0.0557	5.0	22 (14.7%)	150 (100.0%)
Linear regression:	Method 1	-0.1845	0.0668	5.4	36 (24.0%)	150 (100.0%)
previous coping score	Method 2	-0.1895	0.0620	2.8	32 (21.3%)	150 (100.0%)
	Method 3	-0.1905	0.0614	2.3	27 (18.0%)	150 (100.0%)
	Method 4	-0.1895	0.0657	2.8	35 (23.3%)	150 (100.0%)
	Method 5	-0.1824	0.0686	6.5	39 (26.0%)	150 (100.0%)

Method 1: Higher coping scores have a higher chance of being missing.
Method 2: Lower coping scores have a higher chance of being missing.
Method 3: Later time period have a higher chance of being missing
Method 4: 30% of coping scores missing at random
Method 5: Higher coping scores have a higher chance of being missing

Table E5.1 Range of Parameter Estimate for Delayed Chemotherapy from Time-Dependent Cox Model Analysis Stratified by Trial Following Last Observation Carried Forward

Positive Relationship Between Quality of Life and Disease-Free Survival and Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival

Combination of βsp and βdel	Method of Artificially Removing Coping Scores	Theoretical Value	Mean Parameter Estimate	Range of Parameter Estimate
Weak, weak	Method 1	-0.165	-0.1641	(-0.4322, 0.0846)
	Method 2	-0.165	-0.1689	(-0.3983, 0.0344)
	Method 3	-0.165	-0.1714	(-0.4036, 0.0862)
	Method 4	-0.165	-0.1730	(-0.4396, 0.0901)
	Method 5	-0.165	-0.1722	(-0.5142, 0.0854)
Weak, strong	Method 1	-0.195	-0.1781	(-0.3721, 0.0937)
	Method 2	-0.195	-0.1864	(-0.4252, 0.0598)
	Method 3	-0.195	-0.1899	(-0.3840, 0.0673)
	Method 4	-0.195	-0.1891	(-0.4418, 0.0897)
	Method 5	-0.195	-0.1877	(-0.4662, 0.0667)
Strong, weak	Method 1	-0.165	-0.1345	(-0.3371, 0.1395)
	Method 2	-0.165	-0.1549	(-0.3548, 0.0904)
	Method 3	-0.165	-0.1514	(-0.3657, 0.1527)
	Method 4	-0.165	-0.1476	(-0.3423, 0.1498)
	Method 5	-0.165	-0.1405	(-0.4166, 0.1447)
Strong, strong	Method 1	-0.195	-0.1803	(-0.3749, 0.0434)
	Method 2	-0.195	-0.1881	(-0.3820, 0.0479)
	Method 3	-0.195	-0.1853	(-0.4068, 0.0258)
	Method 4	-0.195	-0.1863	(-0.4041, 0.0184)
	Method 5	-0.195	-0.1775	(-0.4131, 0.0658)

Method 1: Higher coping scores have a higher chance of being missing.
Method 2: Lower coping scores have a higher chance of being missing.
Method 3: Later time period have a higher chance of being missing
Method 4: 30% of coping scores missing at random
Method 5: Higher coping scores have a higher chance of being missing

Appendix F Results from Time-Dependent Cox Model Analysis for Different Combinations of a Positive Relationship Between Quality of Life and Disease-Free Survival and a Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival Following Multiple Imputation of Simulated Datasets with Coping Scores Artificially Removed

The results of the time-depedent Cox model analysis on the square root of coping score (S_Pacis) and delayed chemotherapy stratified by trial for different combinations of a positive relationship between quality of life and DFS and a positive relationship between delayed chemotherapy and DFS following 10 repetitions of multiple imputation in 50 simulated datasets with coping scores artificially removed are shown in this appendix.

The parameter estimate from each simulated dataset is the mean of the parameter estimate from 10 repetitions of multiple imputation. The variance of the parameter estimate from each simulated dataset is calculated based on the 10 repetitions of multiple imputation according to (2.13). The standard error from each simulated dataset is the square root of the variance of the parameter estimate. These estimates are used to calculate the 95% confidence interval for the parameter estimate from each simulated dataset. The exponential of the lower and upper 95% confidence limits for the parameter estimate gives the lower and upper 95% confidence limits for the hazard ratio from the simulated dataset. It is then determined for how many of the 95% confidence intervals for the hazard ratio from the 50 simulated datasets contains the value 1. Similarly, it is determined for how many of the 95% confidence intervals for the parameter estimated from the

50 simulated datasets contained the parameter estimate from the complete simulated dataset with no missing observations.

The results for the parameter estimates from the above time-dependent Cox model analysis for the different combinations of a positive relationship between quality of life and DFS and a positive relationship between delayed chemotherapy and DFS are shown as noted below:

Combination of β_{sp} and β_{del}	Estimate of β_{sp}	Estimate of β_{del}
Weak, weak	Table F1.1	Table F1.2
Weak, strong	Table F2.1	Table F2.2
Strong, weak	Table F3.1	Table F3.2
Strong, strong	Table F4.1	Table F4.2

Table F1.1 Summary of Square Root of Coping Score (S_Pacis) from Time-Dependent Cox Model Analysis Stratified by Trial Weak Positive Relationship Between Quality of Life and Disease-Free Survival and Weak Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival

Imputation Method	Detail	Method of Artificially Removing Coping Scores	Parameter Estimate: mean [range]	Bias (%): Theoretical Parameter Value = 0.1	Mean Standard Error	n (%) of the 50x95% CIs for Hazard Ratio Containing 1	n (%) of the 50x95% CIs for Parameter Estimate Containing Simulated Value
Bootstrap	baseline coping score	Method 1	0.0867 [0.0634 to 0.1204]	13.3	0.0147	0 (0%)	47 (94%)
		Method 2	0.0780 [0.0568 to 0.0977]	22.0	0.0152	0 (0%)	39 (78%)
		Method 3	0.0645 [0.0493 to 0.0831]	35.5	0.0141	0 (0%)	11 (22%)
		Method 4	0.0783 [0.0602 to 0.1036]	21.7	0.0147	0 (0%)	37 (74%)
		Method 5	0.0820 [0.0572 to 0.1076]	18.0	0.0156	0 (0%)	43 (86%)
Bootstrap	previous coping score	Method 1	0.0925 [0.0727 to 0.1245]	7.5	0.0143	0 (0%)	49 (98%)
		Method 2	0.0865 [0.0639 to 0.1082]	13.5	0.0151	0 (0%)	49 (98%)
		Method 3	0.0734 [0.0459 to 0.0945]	26.6	0.0142	0 (0%)	29 (58%)
		Method 4	0.0858 [0.0685 to 0.1113]	14.2	0.0144	0 (0%)	47 (94%)
		Method 5	0.0889 [0.0601 to 0.1147]	11.1	0.0153	0 (0%)	49 (98%)
Nearest neighbour		Method 1	0.0893 [0.0719 to 0.1261]	10.7	0.0138	0 (0%)	49 (98%)
		Method 2	0.0910 [0.0710 to 0.1142]	9.0	0.0142	0 (0%)	50 (100%)
		Method 3	0.0797 [0.0491 to 0.1051]	20.3	0.0134	0 (0%)	35 (70%)
		Method 4	0.0880 [0.0705 to 0.1109]	12.0	0.0138	0 (0%)	50 (100%)
		Method 5	0.0875 [0.0590 to 0.1129]	12.5	0.0145	0 (0%)	48 (96%)

Table F1.1 Summary of Square Root of Coping Score (S_Pacis) from Time-Dependent Cox Model Analysis Stratified by Trial Weak Positive Relationship Between Quality of Life and Disease-Free Survival and Weak Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival (continued)

Imputation Method	Detail	Method of Artificially Removing Coping Scores	Parameter Estimate: mean [range]	Bias (%): Theoretical Parameter Value = 0.1	Mean Standard Error	n (%) of the 50x95% CIs for Hazard Ratio Containing 1	n (%) of the 50x95% CIs for Parameter Estimate Containing Simulated Value
Predictive mean matching	initial steps as described for NNI	Method 1	0.0895 [0.0728 to 0.1259]	10.5	0.0139	0 (0%)	49 (98%)
		Method 2	0.0905 [0.0701 to 0.1141]	9.5	0.0143	0 (0%)	50 (100%)
		Method 3	0.0800 [0.0514 to 0.1024]	20.0	0.0134	0 (0%)	38 (76%)
		Method 4	0.0879 [0.0705 to 0.1126]	12.1	0.0138	0 (0%)	50 (100%)
		Method 5	0.0870 [0.0586 to 0.1124]	13.0	0.0145	0 (0%)	47 (94%)
Pattern mixture models		Method 1	0.0890 [0.0703 to 0.1221]	11.0	0.0141	0 (0%)	49 (98%)
		Method 2	0.0925 [0.0699 to 0.1148]	7.5	0.0143	0 (0%)	49 (98%)
		Method 3	0.0809 [0.0509 to 0.1036]	19.1	0.0137	0 (0%)	45 (90%)
		Method 4	0.0877 [0.0703 to 0.1110]	12.3	0.0141	0 (0%)	49 (98%)
		Method 5	0.0867 [0.0605 to 0.1123]	13.3	0.0149	0 (0%)	46 (92%)

Method 1: Higher coping scores have a higher chance of being missing; Method 2: Lower coping scores have a higher chance of being missing;
Method 3: Later time period have a higher chance of being missing; Method 4: 30% of coping scores missing at random;
Method 5: Higher coping scores have a higher chance of being missing

Table F1.2 Summary of Delayed Chemotherapy from Time-Dependent Cox Model Analysis Stratified by Trial
Weak Positive Relationship Between Quality of Life and Disease-Free Survival and Weak Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival

Imputation Method	Detail	Method of Artificially Removing Coping Scores	Parameter Estimate: mean [range]	Bias (%): Theoretical Parameter Value = -0.165	Mean Standard Error	n (%) of the 50x95% CIs for Hazard Ratio Containing 1	n (%) of the 50x95% CIs for Parameter Estimate Containing Simulated Value
Bootstrap	baseline coping score	Method 1	-0.1623 [-0.4270 to 0.0865]	1.6	0.0656	15 (30%)	50 (100%)
		Method 2	-0.1712 [-0.3892 to -0.0074]	3.8	0.0631	16 (32%)	50 (100%)
		Method 3	-0.1699 [-0.4002 to 0.0177]	3.0	0.0615	15 (30%)	50 (100%)
		Method 4	-0.1814 [-0.4260 to -0.0065]	9.9	0.0658	13 (26%)	50 (100%)
		Method 5	-0.1759 [-0.5083 to 0.0433]	6.6	0.0676	18 (36%)	50 (100%)
Bootstrap	previous coping score	Method 1	-0.1650 [-0.4300 to 0.0870]	0.0	0.0656	14 (28%)	50 (100%)
		Method 2	-0.1744 [-0.3944 to -0.0076]	5.7	0.0631	15 (30%)	50 (100%)
		Method 3	-0.1742 [-0.4014 to 0.0100]	5.6	0.0616	14 (28%)	50 (100%)
		Method 4	-0.1847 [-0.4336 to -0.0086]	11.9	0.0658	12 (24%)	50 (100%)
		Method 5	-0.1790 [-0.5120 to 0.0397]	8.5	0.0676	16 (32%)	50 (100%)
Nearest neighbour		Method 1	-0.1645 [-0.4325 to 0.0830]	0.3	0.0655	14 (28%)	50 (100%)
		Method 2	-0.1780 [-0.3941 to -0.0120]	7.9	0.0631	15 (30%)	50 (100%)
		Method 3	-0.1787 [-0.4046 to 0.0028]	8.3	0.0615	12 (24%)	50 (100%)
		Method 4	-0.1870 [-0.4364 to -0.0091]	13.4	0.0657	12 (24%)	50 (100%)
		Method 5	-0.1807 [-0.5128 to 0.0415]	9.5	0.0675	16 (32%)	50 (100%)

Table F1.2 Summary of Delayed Chemotherapy from Time-Dependent Cox Model Analysis Stratified by Trial
Weak Positive Relationship Between Quality of Life and Disease-Free Survival and Weak Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival (continued)

Imputation Method	Detail	Method of Artificially Removing Coping Scores	Parameter Estimate: mean [range]	Bias (%): Theoretical Parameter Value = -0.165	Mean Standard Error	n (%) of the 50x95% CIs for Hazard Ratio Containing 1	n (%) of the 50x95% CIs for Parameter Estimate Containing Simulated Value
Predictive mean matching	initial steps as described for NNI	Method 1	-0.1646 [-0.4295 to 0.0844]	0.2	0.0655	14 (28%)	50 (100%)
		Method 2	-0.1778 [-0.3926 to -0.0113]	7.7	0.0631	15 (30%)	50 (100%)
		Method 3	-0.1789 [-0.4048 to 0.0029]	8.4	0.0615	14 (28%)	50 (100%)
		Method 4	-0.1870 [-0.4345 to -0.0111]	13.3	0.0658	12 (24%)	50 (100%)
		Method 5	-0.1805 [-0.5135 to 0.0433]	9.4	0.0675	16 (32%)	50 (100%)
Pattern mixture models		Method 1	-0.1637 [-0.4284 to 0.0863]	0.8	0.0655	14 (28%)	50 (100%)
		Method 2	-0.1799 [-0.3979 to -0.0152]	9.1	0.0632	15 (30%)	50 (100%)
		Method 3	-0.1772 [-0.4015 to 0.0029]	7.4	0.0616	13 (26%)	50 (100%)
		Method 4	-0.1870 [-0.4265 to -0.0072]	13.0	0.0676	12 (24%)	50 (100%)
		Method 5	-0.1796 [-0.5121 to 0.0411]	8.9	0.0675	16 (32%)	50 (100%)

Method 1: Higher coping scores have a higher chance of being missing; Method 2: Lower coping scores have a higher chance of being missing;
Method 3: Later time period have a higher chance of being missing; Method 4: 30% of coping scores missing at random;
Method 5: Higher coping scores have a higher chance of being missing

Table F2.1 Summary of Square Root of Coping Score (S_Pacis) from Time-Dependent Cox Model Analysis Stratified by Trial Weak Positive Relationship Between Quality of Life and Disease-Free Survival and Strong Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival

Imputation Method	Detail	Method of Artificially Removing Coping Scores	Parameter Estimate: mean [range]	Bias (%): Theoretical Parameter Value = 0.1	Mean Standard Error	n (%) of the 50x95% CIs for Hazard Ratio Containing 1	n (%) of the 50x95% CIs for Parameter Estimate Containing Simulated Value
Bootstrap	baseline coping score	Method 1	0.0869 [0.0602 to 0.1166]	13.1	0.0148	0 (0%)	48 (96%)
		Method 2	0.0776 [0.0626 to 0.1022]	22.4	0.0153	0 (0%)	38 (76%)
		Method 3	0.0679 [0.0433 to 0.0932]	32.1	0.0145	0 (0%)	12 (24%)
		Method 4	0.0800 [0.0592 to 0.1090]	20.0	0.0147	0 (0%)	36 (72%)
		Method 5	0.0823 [0.0562 to 0.1123]	17.7	0.0155	0 (0%)	46 (92%)
Bootstrap	previous coping score	Method 1	0.0933 [0.0625 to 0.1185]	6.7	0.0143	0 (0%)	50 (100%)
		Method 2	0.0868 [0.0683 to 0.1157]	13.2	0.0151	0 (0%)	49 (98%)
		Method 3	0.0752 [0.0455 to 0.0945]	24.8	0.0143	0 (0%)	30 (60%)
		Method 4	0.0879 [0.0585 to 0.1174]	12.1	0.0144	0 (0%)	47 (94%)
		Method 5	0.0900 [0.0580 to 0.1215]	10.0	0.0151	0 (0%)	50 (100%)
Nearest neighbour		Method 1	0.0904 [0.0576 to 0.1159]	9.6	0.0138	0 (0%)	49 (98%)
		Method 2	0.0914 [0.0721 to 0.1205]	8.6	0.0143	0 (0%)	50 (100%)
		Method 3	0.0830 [0.0555 to 0.1024]	17.0	0.0133	0 (0%)	40 (80%)
		Method 4	0.0895 [0.0640 to 0.1154]	10.5	0.0138	0 (0%)	47 (94%)
		Method 5	0.0890 [0.0610 to 0.1146]	11.0	0.0145	0 (0%)	48 (96%)

Table F2.1 Summary of Square Root of Coping Score (S_Pacis) from Time-Dependent Cox Model Analysis Stratified by Trial Weak Positive Relationship Between Quality of Life and Disease-Free Survival and Strong Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival (continued)

Imputation Method	Detail	Method of Artificially Removing Coping Scores	Parameter Estimate: mean [range]	Bias (%): Theoretical Parameter Value = 0.1	Mean Standard Error	n (%) of the 50x95% CIs for Hazard Ratio Containing 1	n (%) of the 50x95% CIs for Parameter Estimate Containing Simulated Value
Predictive mean matching	initial steps as described for NNI	Method 1	0.0904 [0.0581 to 0.1189]	9.6	0.0138	0 (0%)	49 (96%)
		Method 2	0.0916 [0.0724 to 0.1230]	8.4	0.0142	0 (0%)	50 (100%)
		Method 3	0.0830 [0.0555 to 0.1022]	17.0	0.0135	0 (0%)	44 (88%)
		Method 4	0.0895 [0.0615 to 0.1168]	10.5	0.0139	0 (0%)	47 (94%)
		Method 5	0.0895 [0.0608 to 0.1153]	10.5	0.0145	0 (0%)	48 (96%)
Pattern mixture models		Method 1	0.0894 [0.0583 to 0.1176]	10.6	0.0141	0 (0%)	49 (98%)
		Method 2	0.0935 [0.0763 to 0.1233]	6.5	0.0144	0 (0%)	50 (100%)
		Method 3	0.0830 [0.0571 to 0.1048]	17.0	0.0137	0 (0%)	43 (86%)
		Method 4	0.0894 [0.0625 to 0.1142]	10.6	0.0142	0 (0%)	46 (92%)
		Method 5	0.0879 [0.0574 to 0.1159]	12.1	0.0149	0 (0%)	49 (98%)

Method 1: Higher coping scores have a higher chance of being missing; Method 2: Lower coping scores have a higher chance of being missing;
Method 3: Later time period have a higher chance of being missing; Method 4: 30% of coping scores missing at random;
Method 5: Higher coping scores have a higher chance of being missing

Table F2.2 Summary of Delayed Chemotherapy from Time-Dependent Cox Model Analysis Stratified by Trial
Weak Positive Relationship Between Quality of Life and Disease-Free Survival and Strong Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival

Imputation Method	Detail	Method of Artificially Removing Coping Scores	Parameter Estimate: mean [range]	Bias (%): Theoretical Parameter Value = -0.195	Mean Standard Error	n (%) of the 50x95% CIs for Hazard Ratio Containing 1	n (%) of the 50x95% CIs for Parameter Estimate Containing Simulated Value
Bootstrap	baseline coping score	Method 1	-0.1667 [-0.3697 to -0.0170]	14.5	0.0656	19 (38%)	50 (100%)
		Method 2	-0.1724 [-0.4141 to 0.0070]	11.6	0.0633	15 (30%)	50 (100%)
		Method 3	-0.1770 [-0.3732 to -0.0089]	9.2	0.0615	13 (26%)	50 (100%)
		Method 4	-0.1844 [-0.4419 to 0.0009]	5.5	0.0659	15 (30%)	50 (100%)
		Method 5	-0.1767 [-0.4688 to -0.0206]	9.4	0.0676	15 (30%)	49 (100%)
Bootstrap	previous coping score	Method 1	-0.1704 [-0.3728 to -0.0231]	12.6	0.0656	18 (36%)	50 (100%)
		Method 2	-0.1767 [-0.4220 to -0.0018]	9.4	0.0633	15 (30%)	50 (100%)
		Method 3	-0.1807 [-0.3789 to -0.0172]	7.4	0.0616	11 (22%)	50 (100%)
		Method 4	-0.1882 [-0.4379 to -0.0037]	3.5	0.0659	16 (32%)	50 (100%)
		Method 5	-0.1813 [-0.4677 to -0.0246]	7.0	0.0676	14 (28%)	50 (100%)
Nearest neighbour		Method 1	-0.1697 [-0.3698 to -0.0230]	13.0	0.0656	18 (36%)	50 (100%)
		Method 2	-0.1802 [-0.4205 to -0.0054]	7.6	0.0633	14 (28%)	50 (100%)
		Method 3	-0.1856 [-0.3792 to -0.0200]	4.8	0.0615	9 (18%)	50 (100%)
		Method 4	-0.1898 [-0.4414 to -0.0055]	2.7	0.0658	14 (28%)	50 (100%)
		Method 5	-0.1817 [-0.4686 to -0.0247]	6.8	0.0675	14 (28%)	50 (100%)

Table F2.2 Summary of Delayed Chemotherapy from Time-Dependent Cox Model Analysis Stratified by Trial
Weak Positive Relationship Between Quality of Life and Disease-Free Survival and Strong Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival (continued)

Imputation Method	Detail	Method of Artificially Removing Coping Scores	Parameter Estimate: mean [range]	Bias (%): Theoretical Parameter Value = -0.195	Mean Standard Error	n (%) of the 50x95% CIs for Hazard Ratio Containing 1	n (%) of the 50x95% CIs for Parameter Estimate Containing Simulated Value
Predictive mean matching	initial steps as described for NNI	Method 1	-0.1696 [-0.3713 to -0.0245]	13.0	0.0656	18 (36%)	50 (100%)
		Method 2	-0.1803 [-0.4222 to -0.0034]	7.5	0.0633	14 (28%)	50 (100%)
		Method 3	-0.1857 [-0.3802 to -0.0186]	4.8	0.0615	10 (20%)	50 (100%)
		Method 4	-0.1899 [-0.4415 to -0.0054]	2.6	0.0658	13 (26%)	50 (100%)
		Method 5	-0.1818 [-0.4693 to -0.0251]	6.8	0.0675	14 (28%)	50 (100%)
Pattern mixture models		Method 1	-0.1687 [-0.3733 to -0.0204]	13.5	0.0656	19 (38%)	50 (100%)
		Method 2	-0.1819 [-0.4307 to -0.0015]	6.7	0.0633	14 (28%)	50 (100%)
		Method 3	-0.1863 [-0.3979 to -0.0299]	4.5	0.0617	10 (20%)	50 (100%)
		Method 4	-0.1893 [-0.4431 to -0.0040]	2.9	0.0659	14 (28%)	50 (100%)
		Method 5	-0.1801 [-0.0261 to 0.0675]	7.6	0.0675	14 (28%)	50 (100%)

Method 1: Higher coping scores have a higher chance of being missing; Method 2: Lower coping scores have a higher chance of being missing;
Method 3: Later time period have a higher chance of being missing; Method 4: 30% of coping scores missing at random;
Method 5: Higher coping scores have a higher chance of being missing

Table F3.1 Summary of Square Root of Coping Score (S_Pacis) from Time-Dependent Cox Model Analysis Stratified by Trial Strong Positive Relationship Between Quality of Life and Disease-Free Survival and Weak Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival

Imputation Method	Detail	Method of Artificially Removing Coping Scores	Parameter Estimate: mean [range]	Bias (%): Theoretical Parameter Value = 0.4	Mean Standard Error	n (%) of the 50x95% CIs for Hazard Ratio Containing 1	n (%) of the 50x95% CIs for Parameter Estimate Containing Simulated Value
Bootstrap	baseline coping score	Method 1	0.3074 [0.2811 to 0.3419]	23.1	0.0169	0 (0%)	0 (0%)
		Method 2	0.2956 [0.2655 to 0.3223]	26.1	0.0183	0 (0%)	0 (0%)
		Method 3	0.2383 [0.2211 to 0.2563]	40.4	0.0161	0 (0%)	0 (0%)
		Method 4	0.2864 [0.2645 to 0.3092]	28.4	0.0174	0 (0%)	0 (0%)
		Method 5	0.2878 [0.2597 to 0.3150]	28.1	0.0180	0 (0%)	0 (0%)
Bootstrap	previous coping score	Method 1	0.3375 [0.3074 to 0.3730]	15.6	0.0169	0 (0%)	0 (0%)
		Method 2	0.3346 [0.3052 to 0.3674]	16.4	0.0184	0 (0%)	0 (0%)
		Method 3	0.2712 [0.2525 to 0.2986]	32.2	0.0172	0 (0%)	0 (0%)
		Method 4	0.3247 [0.2947 to 0.3505]	18.8	0.0177	0 (0%)	0 (0%)
		Method 5	0.3226 [0.2968 to 0.3527]	19.4	0.0181	0 (0%)	0 (0%)
Nearest neighbour		Method 1	0.3474 [0.3198 to 0.3765]	13.2	0.0164	0 (0%)	0 (0%)
		Method 2	0.3540 [0.3272 to 0.3899]	11.5	0.0180	0 (0%)	7 (14%)
		Method 3	0.3138 [0.2854 to 0.3327]	21.6	0.0166	0 (0%)	0 (0%)
		Method 4	0.3448 [0.3097 to 0.3695]	13.8	0.0169	0 (0%)	1 (2%)
		Method 5	0.3391 [0.3121 to 0.3739]	15.2	0.0171	0 (0%)	0 (0%)

Table F3.1 Summary of Square Root of Coping Score (S_Pacis) from Time-Dependent Cox Model Analysis Stratified by Trial Strong Positive Relationship Between Quality of Life and Disease-Free Survival and Weak Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival (continued)

Imputation Method	Detail	Method of Artificially Removing Coping Scores	Parameter Estimate: mean [range]	Bias (%): Theoretical Parameter Value = 0.4	Mean Standard Error	n (%) of the 50x95% CIs for Hazard Ratio Containing 1	n (%) of the 50x95% CIs for Parameter Estimate Containing Simulated Value
Predictive mean matching	initial steps as described for NNI	Method 1	0.3472 [0.3209 to 0.3771]	13.2	0.0164	0 [0 to 0]	0 (0%)
		Method 2	0.3541 [0.3268 to 0.3888]	11.5	0.0178	0 [0 to 0]	6 (12%)
		Method 3	0.3135 [0.2879 to 0.3429]	21.6	0.0165	0 [0 to 0]	0 (0%)
		Method 4	0.3441 [0.3098 to 0.3668]	14.0	0.0169	0 [0 to 0]	1 (2%)
		Method 5	0.3389 [0.3100 to 0.3714]	15.3	0.0173	0 [0 to 0]	0 (0%)
Pattern mixture models		Method 1	0.3452 [0.3197 to 0.3711]	13.7	0.0165	0 [0 to 0]	0 (0%)
		Method 2	0.3621 [0.3335 to 0.3980]	9.5	0.0173	0 [0 to 0]	14 (28%)
		Method 3	0.3151 [0.2952 to 0.3371]	21.2	0.0166	0 [0 to 0]	0 (0%)
		Method 4	0.3442 [0.3177 to 0.3702]	13.9	0.0171	0 [0 to 0]	2 (4%)
		Method 5	0.3375 [0.3077 to 0.3719]	15.6	0.0173	0 [0 to 0]	0 (0%)

Method 1: Higher coping scores have a higher chance of being missing; Method 2: Lower coping scores have a higher chance of being missing;
Method 3: Later time period have a higher chance of being missing; Method 4: 30% of coping scores missing at random;
Method 5: Higher coping scores have a higher chance of being missing

Table F3.2 Summary of Delayed Chemotherapy from Time-Dependent Cox Model Analysis Stratified by Trial
Strong Positive Relationship Between Quality of Life and Disease-Free Survival and Weak Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival

Imputation Method	Detail	Method of Artificially Removing Coping Scores	Parameter Estimate: mean [range]	Bias (%): Theoretical Parameter Value = -0.165	Mean Standard Error	n (%) of the 50x95% CIs for Hazard Ratio Containing 1	n (%) of the 50x95% CIs for Parameter Estimate Containing Simulated Value
Bootstrap	baseline coping score	Method 1	-0.1003 [-0.2934 to 0.1143]	39.2	0.0682	33 (66%)	48 (96%)
		Method 2	-0.0966 [-0.2877 to 0.1249]	41.5	0.0645	30 (60%)	50 (100%)
		Method 3	-0.0840 [-0.3137 to 0.1213]	49.1	0.0635	34 (68%)	48 (96%)
		Method 4	-0.1061 [-0.3077 to 0.0772]	35.7	0.0682	31 (62%)	50 (100%)
		Method 5	-0.0943 [-0.3439 to 0.0981]	42.9	0.0705	34 (68%)	49 (98%)
Bootstrap	previous coping score	Method 1	-0.1126 [-0.2860 to 0.0997]	31.7	0.0683	28 (56%)	49 (98%)
		Method 2	-0.1141 [-0.3034 to 0.0912]	32.5	0.0650	30 (60%)	50 (100%)
		Method 3	-0.1021 [-0.3369 to 0.0980]	38.1	0.0646	32 (64%)	49 (98%)
		Method 4	-0.1217 [-0.3284 to 0.0838]	26.2	0.0681	26 (52%)	50 (100%)
		Method 5	-0.1091 [-0.3558 to 0.0751]	33.9	0.0707	32 (64%)	50 (100%)
Nearest neighbour		Method 1	-0.1154 [-0.2902 to 0.1016]	30.0	0.0678	27 (54%)	49 (98%)
		Method 2	-0.1257 [-0.3338 to 0.0979]	23.8	0.0645	26 (52%)	50 (100%)
		Method 3	-0.1230 [-0.3687 to 0.0768]	25.5	0.0638	28 (56%)	49 (98%)
		Method 4	-0.1327 [-0.3374 to 0.0699]	19.6	0.0674	22 (44%)	50 (100%)
		Method 5	-0.1160 [-0.3857 to 0.0951]	29.7	0.0698	32 (64%)	50 (100%)

Table F3.2 Summary of Delayed Chemotherapy from Time-Dependent Cox Model Analysis Stratified by Trial
Strong Positive Relationship Between Quality of Life and Disease-Free Survival and Weak Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival (continued)

Imputation Method	Detail	Method of Artificially Removing Coping Scores	Parameter Estimate: mean [range]	Bias (%): Theoretical Parameter Value = -0.165	Mean Standard Error	n (%) of the 50x95% CIs for Hazard Ratio Containing 1	n (%) of the 50x95% CIs for Parameter Estimate Containing Simulated Value
Predictive mean matching	initial steps as described for NNI	Method 1	-0.1152 [-0.2988 to 0.1027]	30.2	0.0678	28 (56%)	49 (98%)
		Method 2	-0.1253 [-0.3171 to 0.1014]	24.1	0.0645	25 (50%)	50 (100%)
		Method 3	-0.1243 [-0.3620 to 0.0815]	24.6	0.0639	25 (50%)	49 (98%)
		Method 4	-0.1332 [-0.3365 to 0.0586]	19.3	0.0677	24 (48%)	50 (100%)
		Method 5	-0.1160 [-0.3923 to 0.0936]	29.7	0.0700	32 (64%)	50 (100%)
Pattern mixture models		Method 1	-0.1136 [-0.2957 to 0.0961]	31.1	0.0678	28 (56%)	49 (98%)
		Method 2	-0.1367 [-0.3643 to 0.0752]	17.1	0.0640	22 (44%)	50 (100%)
		Method 3	-0.1257 [-0.3671 to 0.0416]	23.8	0.0653	26 (52%)	49 (98%)
		Method 4	-0.1357 [-0.3361 to 0.0641]	17.8	0.0677	23 (46%)	50 (100%)
		Method 5	-0.1149 [-0.3898 to 0.0891]	30.4	0.0702	34 (68%)	50 (100%)

Method 1: Higher coping scores have a higher chance of being missing; Method 2: Lower coping scores have a higher chance of being missing;
Method 3: Later time period have a higher chance of being missing; Method 4: 30% of coping scores missing at random;
Method 5: Higher coping scores have a higher chance of being missing

Table F4.1 Summary of Square Root of Coping Score (S_Pacis) from Time-Dependent Cox Model Analysis Stratified by Trial Strong Positive Relationship Between Quality of Life and Disease-Free Survival and Strong Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival

Imputation Method	Detail	Method of Artificially Removing Coping Scores	Parameter Estimate: mean [range]	Bias (%): Theoretical Parameter Value = 0.4	Mean Standard Error	n (%) of the 50x95% CIs for Hazard Ratio Containing 1	n (%) of the 50x95% CIs for Parameter Estimate Containing Simulated Value
Bootstrap	baseline coping score	Method 1	0.3090 [0.2875 to 0.3284]	22.7	0.0171	0 (0%)	0 (0%)
		Method 2	0.2934 [0.2646 to 0.3138]	26.7	0.0182	0 (0%)	0 (0%)
		Method 3	0.2347 [0.2100 to 0.2689]	41.3	0.0162	0 (0%)	0 (0%)
		Method 4	0.2858 [0.2578 to 0.3125]	28.5	0.0177	0 (0%)	0 (0%)
		Method 5	0.2928 [0.2684 to 0.3240]	26.8	0.0182	0 (0%)	0 (0%)
Bootstrap	previous coping score	Method 1	0.3412 [0.3189 to 0.3716]	14.7	0.0169	0 (0%)	0 (0%)
		Method 2	0.3341 [0.3095 to 0.3632]	16.5	0.0182	0 (0%)	0 (0%)
		Method 3	0.2707 [0.2558 to 0.2981]	32.3	0.0165	0 (0%)	0 (0%)
		Method 4	0.3257 [0.3074 to 0.3526]	18.6	0.0175	0 (0%)	0 (0%)
		Method 5	0.3278 [0.3042 to 0.3531]	18.0	0.0184	0 (0%)	0 (0%)
Nearest neighbour		Method 1	0.3511 [0.3238 to 0.3864]	12.2	0.0164	0 (0%)	0 (0%)
		Method 2	0.3517 [0.3306 to 0.3795]	12.1	0.0177	0 (0%)	2 (4%)
		Method 3	0.3123 [0.2862 to 0.3558]	21.9	0.0165	0 (0%)	0 (0%)
		Method 4	0.3445 [0.3172 to 0.3770]	13.9	0.0169	0 (0%)	1 (2%)
		Method 5	0.3443 [0.3219 to 0.3810]	13.9	0.0174	0 (0%)	2 (4%)

Table F4.1 Summary of Square Root of Coping Score (S_Pacis) from Time-Dependent Cox Model Analysis Stratified by Trial Strong Positive Relationship Between Quality of Life and Disease-Free Survival and Strong Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival (continued)

Imputation Method	Detail	Method of Artificially Removing Coping Scores	Parameter Estimate: mean [range]	Bias (%): Theoretical Parameter Value = 0.4	Mean Standard Error	n (%) of the 50x95% CIs for Hazard Ratio Containing 1	n (%) of the 50x95% CIs for Parameter Estimate Containing Simulated Value
Predictive mean matching	initial steps as described for NNI	Method 1	0.3507 [0.3238 to 0.3868]	12.3	0.0164	0 (0%)	0 (0%)
		Method 2	0.3531 [0.3254 to 0.3794]	11.7	0.0177	0 (0%)	3 (6%)
		Method 3	0.3126 [0.2900 to 0.3477]	21.9	0.0164	0 (0%)	0 (0%)
		Method 4	0.3449 [0.3177 to 0.3738]	13.8	0.0170	0 (0%)	1 (2%)
		Method 5	0.3441 [0.3187 to 0.3738]	14.0	0.0174	0 (0%)	0 (0%)
Pattern mixture models		Method 1	0.3487 [0.3261 to 0.3824]	12.8	0.0165	0 (0%)	0 (0%)
		Method 2	0.3603 [0.3262 to 0.3953]	9.9	0.0175	0 (0%)	9 (18%)
		Method 3	0.3119 [0.2773 to 0.3584]	22.0	0.0166	0 (0%)	0 (0%)
		Method 4	0.3444 [0.3207 to 0.3678]	13.9	0.0169	0 (0%)	0 (0%)
		Method 5	0.3422 [0.3233 to 0.3709]	14.5	0.0176	0 (0%)	0 (0%)

Method 1: Higher coping scores have a higher chance of being missing; Method 2: Lower coping scores have a higher chance of being missing;
Method 3: Later time period have a higher chance of being missing; Method 4: 30% of coping scores missing at random;
Method 5: Higher coping scores have a higher chance of being missing

Table F4.2 Summary of Delayed Chemotherapy from Time-Dependent Cox Model Analysis Stratified by Trial
Strong Positive Relationship Between Quality of Life and Disease-Free Survival and Strong Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival

Imputation Method	Detail	Method of Artificially Removing Coping Scores	Parameter Estimate: mean [range]	Bias (%): Theoretical Parameter Value = -0.195	Mean Standard Error	n (%) of the 50x95% CIs for Hazard Ratio Containing 1	n (%) of the 50x95% CIs for Parameter Estimate Containing Simulated Value
Bootstrap	baseline coping score	Method 1	-0.1628 [-0.3430 to 0.0122]	16.5	0.0683	19 (38%)	50 (100%)
		Method 2	-0.1491 [-0.3212 to 0.0342]	23.5	0.0648	16 (32%)	49 (98%)
		Method 3	-0.1383 [-0.3036 to 0.0028]	29.1	0.0639	20 (40%)	47 (94%)
		Method 4	-0.1604 [-0.3533 to 0.0266]	17.8	0.0686	16 (32%)	50 (100%)
		Method 5	-0.1651 [-0.3286 to -0.0228]	15.3	0.0708	20 (40%)	50 (100%)
Bootstrap	previous coping score	Method 1	-0.1780 [-0.3533 to 0.0029]	8.7	0.0684	12 (24%)	50 (100%)
		Method 2	-0.1671 [-0.3232 to -0.0010]	14.3	0.0653	15 (30%)	50 (100%)
		Method 3	-0.1545 [-0.3162 to -0.0311]	20.8	0.0643	16 (32%)	50 (100%)
		Method 4	-0.1774 [-0.3768 to 0.0120]	9.0	0.0682	13 (26%)	50 (100%)
		Method 5	-0.1821 [-0.3600 to -0.0401]	6.6	0.0708	14 (28%)	50 (100%)
Nearest neighbour		Method 1	-0.1806 [-0.3533 to -0.0044]	7.4	0.0678	12 (24%)	50 (100%)
		Method 2	-0.1816 [-0.3568 to 0.0048]	6.9	0.0645	11 (22%)	50 (100%)
		Method 3	-0.1762 [-0.3288 to -0.0321]	9.7	0.0639	13 (26%)	50 (100%)
		Method 4	-0.1892 [-0.4057 to -0.0040]	3.0	0.0673	10 (20%)	50 (100%)
		Method 5	-0.1891 [-0.3946 to -0.0469]	3.0	0.0699	12 (24%)	50 (100%)

Table F4.2 Summary of Delayed Chemotherapy from Time-Dependent Cox Model Analysis Stratified by Trial
Strong Positive Relationship Between Quality of Life and Disease-Free Survival and Strong Positive Relationship Between Delayed Chemotherapy and Disease-Free Survival (continued)

Imputation Method	Detail	Method of Artificially Removing Coping Scores	Parameter Estimate: mean [range]	Bias (%): Theoretical Parameter Value = -0.195	Mean Standard Error	n (%) of the 50x95% CIs for Hazard Ratio Containing 1	n (%) of the 50x95% CIs for Parameter Estimate Containing Simulated Value
Predictive mean matching	initial steps as described for NNI	Method 1	-0.1807 [-0.3550 to -0.0104]	7.3	0.0677	13 (26%)	50 (100%)
		Method 2	-0.1836 [-0.3701 to -0.0016]	5.9	0.0641	11 (22%)	50 (100%)
		Method 3	-0.1780 [-0.3309 to -0.0441]	8.7	0.0638	11 (22%)	50 (100%)
		Method 4	-0.1892 [-0.4055 to -0.0054]	3.0	0.0678	10 (20%)	50 (100%)
		Method 5	-0.1887 [-0.3946 to -0.0469]	3.2	0.0701	11 (22%)	50 (100%)
Pattern mixture models		Method 1	-0.1779 [-0.3550 to -0.0070]	8.8	0.0678	13 (26%)	50 (100%)
		Method 2	-0.1953 [-0.3663 to 0.0015]	0.1	0.0654	9 (18%)	50 (100%)
		Method 3	-0.1745 [-0.3147 to -0.0324]	10.5	0.0645	14 (28%)	50 (100%)
		Method 4	-0.1925 [-0.3956 to 0.0011]	1.3	0.0678	9 (18%)	50 (100%)
		Method 5	-0.1874 [-0.3800 to -0.0468]	3.9	0.0701	11 (22%)	50 (100%)

Method 1: Higher coping scores have a higher chance of being missing; Method 2: Lower coping scores have a higher chance of being missing;
Method 3: Later time period have a higher chance of being missing; Method 4: 30% of coping scores missing at random;
Method 5: Higher coping scores have a higher chance of being missing