Modelling event history after allogeneic haematopoietic stem cell transplantation

Elinor Curnow

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy in the Faculty of Health Sciences

Bristol Medical School, October 2021

Word Count: 57, 166

Abstract

For patients with blood diseases, long-term remission is often only possible after haematopoietic stem cell transplantation (HSC). Although transplantation can be life-saving, a number of post-transplant events can increase mortality. This thesis provides the first insight into patient outcomes after HSC transplantation using cord blood donated to the UK National Health Service Cord Blood Bank (NHS-CBB). In the NHS-CBB dataset, event times were incompletely observed for some event types. Missing event times were assumed to have occurred in a known, finite, time-period. Hence, the missing event times were considered interval-censored. Methods for handling interval-censored event times can be categorised as (i) applying multiple imputation (MI) strategies or (ii) taking a full maximum likelihood (FML) approach. I focused on MI strategies, rather than FML methods, because of their flexibility. Using simulation studies, I evaluated MI strategies in competing risks and multi-state model (MSM) analyses, examining the extent to which interval boundaries, the data distribution, and analysis model should be accounted for when data were missing at random (MAR) and missing not at random (MNAR). I found that MI by predictive mean matching (PMM), in which sampling is from a set of observed times without reliance on a specific parametric distribution, resulted in least biased estimates when event times were MAR (conditional on observed data). Furthermore, in MSM analysis, I found that applying PMM separately for each sub-group of patients with a different pathway through the MSM tended to reduce bias and standard error. Finally, I applied the best MI methodology from my simulation studies in an analysis of the NHS-CBB dataset. My results suggest that application of MI methods to the NHS-CBB dataset reduced bias and improved precision in estimates, compared with complete case analysis. My approach ensures that accurate information is available to inform decisions for both clinicians and patients.

Dedication

To Paul, Lowena and Kerensa. All my love.

Acknowledgements

This work would not have been possible without the support of several parties. Firstly, I am indebted to my supervisors Prof. Kate Tilling, Dr. Rachael Hughes and Dr. Kate Birnie, former supervisor Prof. Margaret May, and my collaborator, Dr. Michael Crowther. Thank you for your continued encouragement, inspiration, time, patience, and support. It has been a pleasure and privilege to work with you. I also want to thank Prof. Chris Metcalfe and Dr. Sue Ingle for their helpful comments and advice during my annual reviews.

I am also grateful to NHS Blood and Transplant for funding my research, and to everyone at NHSBT who offered help and guidance along the way. In particular, I want to thank my former colleagues in the Statistics and Clinical Research department: your enthusiasm and dedication to improving patients' lives is admirable (and thanks for all the cakes!).

I am grateful to Eurocord and NHS Blood and Transplant for supplying the UK NHS Cord Blood Bank data, and to all those working in transplant centres who have collected the data used in my research.

Thank you to my friends and family for their continued support. Finally, thank you to my husband, Paul, and my children, Lowena and Kerensa. Paul, I couldn't have done this without you. Knowing that you always have my back means everything. Lowena and Kerensa, thank you for showing suitable reverence for "Mum's PhD", and for always reminding me what's important in life.

Author's Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's *Regulations and Code of Practice for Research Degree Programmes* and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: Elinor Curnow **DATE:** 30 September 2021

Publications arising from this research

Some of the work in this thesis has been published as a paper (primarily, work in Chapter 5, but some elements of the paper are included in other chapters). The reference for this paper is: Curnow E, Hughes RA, Birnie K, Crowther MJ, May MT, Tilling K. Multiple imputation strategies for a bounded outcome variable in a competing risks analysis. *Statistics in Medicine*. 2021;40:1917–1929. My contribution was: designing and analysing the simulation study; cleaning and analysing the data for the applied example; interpreting the results, writing the first draft of the paper and revising in the light of co-author/reviewer comments. Co-authors' contribution was suggestions for the design, analysis, and interpretation of data, as well as commenting on drafts of the manuscript.

SIGNED: Elinor Curnow (First Author) **DATE:** 30 September 2021

SIGNED: Kate Tilling (Final Author) **DATE:** 30 September 2021

Table of Contents

Tables

Figures

References and Appendices

Appendices

List of Abbreviations

ADEMP	aims, data-generating mechanisms, estimands, methods, and performance measures
ANDA	Asymptotic Normal Data Augmentation
BBMR	British Bone Marrow Registry
BM	bone marrow
CB	cord blood
CBB	Cord Blood Bank
CBH	cumulative baseline hazard function
CCA	complete case analysis
CI	confidence interval
CMV+/-	cytomegalovirus-positive/-negative
DGM	data-generating mechanism
EM	expectation-maximisation
FCS	multiple imputation using fully conditional specification
FMI	fraction of missing information
FML	full maximum likelihood
FP	fractional polynomial
FULL-CBB	full data version of the National Health Service Cord Blood Bank dataset
GvHD	graft-versus-host disease
GvL	graft-versus-leukaemia
HIV	human immunodeficiency virus
HLA	human leucocyte antigen
HR	hazard ratio
HSC	haematopoietic stem cell
INTCCR	semi-parametric maximum likelihood approach of Bakoyannis *et al.*
IPW	inverse probability weighting
IQR	inter-quartile range
JAV	just another value, or 'transform then impute'

JM	joint modelling
LOGNORM	log-linear imputation model with post-imputation back-transformation
LRD	multiple imputation using local residual draws
MAR	missing at random
MCAR	missing completely at random
MDM	missing data mechanism
MED	single imputation by replacement with the median of the observed data
MI	multiple imputation
MICI	multiple imputation method proposed by Delord and Genin
MID	single imputation by replacement with the mid-point of the specified boundaries
MITD	multiple imputation, then deletion
MNAR	missing not at random
MSM	multi-state model
NHS	National Health Service
NHSBT	National Health Service Blood and Transplant
NORM	linear imputation model
NORMNOAUX	linear imputation model without auxiliary variables
NORMSUBGP	linear imputation model, applied for separate sub-groups
NPMLE	non-parametric maximum likelihood estimator
OR	odds ratio
PB	peripheral blood
PH	proportional hazards
PMDA	Poor Man's Data Augmentation
PMM	multiple imputation using predictive mean matching
PMM30IMP	multiple imputation using predictive mean matching, with 30 imputations
PMMCOMP	multiple imputation using predictive mean matching, accounting for the ordered nature of the event times

PMMNOAUX	multiple imputation using predictive mean matching without auxiliary variables
PMMSUBGP	multiple imputation using predictive mean matching, applied for separate sub-groups
R&P	extended Royston and Parmar method
RELOS	restricted expected length of stay in state
RESNORM	linear imputation model with restrictions on the imputed values
SE	standard error
SIL	subsample ignorable likelihood
SMC-FCS	substantive model compatible multiple imputation using fully conditional specification
TARMOS	treatment and reporting of missing data in observational studies
TNC	total nucleated cell

CHAPTER 1. INTRODUCTION

1.1. Research motivation

Understanding and improving the treatment of blood disorders is of great clinical importance. Blood disorders include blood cancers, such as leukaemia, lymphoma, and myeloma, as well as non-malignant diseases, such as aplastic anaemia and some benign tumours. On average, 110 people are diagnosed with blood cancers each day in the UK (1). Blood cancer is the most common type of cancer in children and the third most common cause of death due to cancer in the UK (2).

The treatment of blood disorders varies from watchful waiting to chemotherapy or radiotherapy, depending on the aggressiveness of the disorder (1). Long-term remission is often only possible after haematopoietic stem cell (HSC) transplant. Once transplanted into the patient, HSCs are able to proliferate into healthy blood cells, curing, or limiting the progress of the underlying blood disorder. HSCs naturally reside in the bone marrow (BM) but are also present in cord blood (CB) at birth and in circulating (peripheral) blood (PB) (3).

The aim of HSC transplant is either: (i) to restore the patient's own blood cells after intensive chemo- or radiotherapy, known as autologous transplant, or (ii) to replace the patient's abnormal blood cells using donor HSCs, known as allogeneic transplant. Autologous transplant is generally used in the treatment of lymphomas and solid tumours, and allogeneic transplant is generally used in the treatment of leukaemias and non-malignant disorders (4).

In 2019, 4580 HSC transplants were performed in the UK (5), of which 2854 (62%) were autologous transplants and 1726 (38%) were allogeneic transplants. In

allogeneic transplants, matching human leucocyte antigen (HLA) types between the transplant recipient and donor is important to avoid graft rejection (4). If no suitably-matched relative can be found, the transplanting hospital can search a worldwide database of HSC donor registries and banks (6) to identify a suitable unrelated donor. The majority of allogeneic HSC transplants (67% of allogeneic transplants in the UK in 2019) involve unrelated donors, with the source of HSCs being PB, BM or CB (representing 82%, 12% and 6%, respectively, of allogeneic unrelated donor transplants in the UK in 2019) (5).

To date, there is little information available about patient outcomes after HSC transplantation in the UK, due to limited access to patient data. UK patient outcome data could be used to improve clinical practice, and to support evidence-based policymaking around HSC transplantation in the UK and globally. In preparation for this thesis, I was granted access to data about patient outcomes after HSC transplantation using CB donated via a UK CB bank.

The purpose of my research is thus to provide the first insight into patient outcomes after HSC transplantation using CB donated to a UK CB bank. In the following sections, I will describe CB donation and transplantation in more detail (Section 1.2). Here (Section 1.3), and in Chapter 4, I will describe the real dataset that I have used in my research and why this research is important for clinicians and patients. Finally, I will summarise my research aims and objectives, and provide an outline of my thesis (Section 1.4).

1.2. Cord blood donation and transplantation

1.2.1. Cord blood donation

Donations to public CB banks are entirely altruistic acts. The mother is not paid for their donation and generally cannot direct the choice of recipient. Briefly, the process for collection and storage of CB is as follows (7):

- Consent for collection of CB is given by the baby's mother prior to birth.
- The mother receives standard care during pregnancy and birth.
- After the birth, delivery of the placenta, and clamping of the umbilical cord (at the point at which the placenta and umbilical cord would otherwise be discarded), CB is drawn from the placenta and umbilical cord.
- After collection, CB is processed, tested for quality control purposes, and cryopreserved until requested for use in a transplant.

1.2.2. Comparison of cord blood, bone marrow, and peripheral blood transplantation

Compared with BM and PB transplantation, CB transplantation requires less stringent HLA matching (8). Due to the large degree of polymorphism of HLA alleles, it can be difficult to find a well-matched related or unrelated donor (9). This is particularly the case for patients whose ethnicity is under-represented in donor registries, because distributions of HLA alleles vary between ethnic groups (10) (in the UK, the majority of registered adult donors are Northern European) (11). In this case, CB transplantation is a viable alternative. CB also has the advantage that it is immediately available, so can be used for urgent transplantation (8). In addition, CB donation poses no risk to the donor. In contrast, BM and PB donation carry some risk, although donor adverse events are usually mild and short-lived (mainly bone pain and body aches) (12).

1.2.3. Engraftment

The principal measure of success of an allogeneic transplant is full engraftment of the donor HSCs (8). Full engraftment means that the patient's blood cells and immune system are completely replaced by progeny (mature blood and immune cells) of the donor HSCs (10). Compared with BM and PB transplants, CB transplantation is associated with delayed engraftment (9, 13). In previous studies, the median time to myeloid engraftment (defined as absolute neutrophil count > 0.5×10^9/l on three consecutive days) has been reported as between 22-

30 days post-transplant for CB transplants, compared with 14-21 days for BM and PB transplants (8, 14, 15). However, most patients (85-100%) do eventually achieve myeloid engraftment after CB transplantation (8).

1.2.4. Adverse events

After transplantation, patients can experience a number of adverse events that are associated with increased mortality (16), including acute and chronic graft-versus-host disease (GvHD, described in more detail in Section 1.2.5), and relapse (described in more detail in Section 1.2.6). The risks of acute GvHD, relapse, and mortality are similar for CB, BM, and PB transplants using unrelated donors (9, 13). Furthermore, the risk of chronic GvHD (see Section 1.2.5) is lower for CB transplants, compared with BM and PB transplants using unrelated donors (9).

1.2.5. Graft-versus-host disease

GvHD is caused by an immune response of donor cells (the "graft") against the patient's tissues and organs (the "host"). HLAs on the surface of body cells enable our immune system to distinguish between our own and "foreign" cells in the body. However, after HSC transplantation (unless the donor and patient have very similar HLA types), donor immune cells will treat the patient's body cells as foreign and will start to attack them (17).

Acute and chronic graft-versus-host disease

There are two types of GvHD: "acute" and "chronic". The standard classification of GvHD into these two types depends on whether GvHD occurs before (acute) or after (chronic) day 100 post-transplant. More recently, it has been recognised that GvHD with acute-like symptoms can occur after day 100 (17). Acute GvHD usually affects the skin, liver, and gastrointestinal tract. There are four grades of acute GvHD: mild (grade 1), moderate (grade 2), severe (grade 3), and very severe (grade 4). Grade 3-4 acute GvHD has a very poor prognosis and is the

main cause of mortality in the first 100 days post-transplant (17). However, some degree of acute GvHD is desirable in patients with blood cancers in order to eradicate cancerous cells (the "graft-versus-leukaemia" effect, GvL) (10). There is some evidence of this effect in non-malignant disorders too (10).

Chronic GvHD has similar characteristics to auto-immune diseases and immunodeficiencies, and can affect all organs and tissues in the body (17). There are three grades of chronic GvHD: mild, moderate, and severe. Chronic GvHD is the main cause of non-relapse mortality after day 100 post-transplant (17).

1.2.6. Disease relapse

Disease relapse means that there are signs and symptoms that the patient's original blood disease has returned after treatment. If there is no evidence that the blood disease has returned, the patient is said to be "in remission" (18). Most cases of relapse occur within two years of treatment (18). Patients in relapse can still benefit from a subsequent HSC transplant, although engraftment is less likely and the risk of post-transplant relapse or death is increased, compared with patients in remission at the time of transplant (19). Disease relapse is the main cause of death, unrelated to transplant, after allogeneic HSC transplantation (20).

1.2.7. Covariates associated with acute graft-versus-host disease and relapse

In clinical studies of CB transplantation (4, 8, 10, 14, 17, 19, 21-26), the most frequently reported adverse events are acute GvHD and relapse. There is consensus in the clinical literature about the covariates that are associated with acute GvHD and relapse following CB transplantation (4, 8, 10, 14, 17, 19, 21-26). These covariates, and the values associated with increased hazard of acute GvHD and relapse, are summarised in Table 1.1 overleaf.

Table 1.1. Summary of literature review of covariates associated with hazard of acute GvHD and relapse.

Covariate	**Values associated with increased hazard of event**	
	Acute GvHD	**Relapse**
Number of CB units received	Double cord	Single cord
Disease status at time of transplant	In remission at time of transplant	In relapse at time of transplant
Pre-transplant radio- or chemotherapy (conditioning) regimen	Intensive conditioning	Reduced intensity conditioning
Total nucleated cell (TNC) dose at infusion	≥ 3 (vs. < 3×10^7/kg) or ≥ 5 (vs. < 5×10^7/kg)	< 3 (vs. ≥ 3×10^7/kg) or < 5 (vs. ≥ 5×10^7/kg)
GvHD prophylaxis (yes/no – type of prophylaxis varies from study to study)	No prophylaxis	Prophylaxis
Disease type	Analyses generally performed separately for different types of blood cancers, and non-malignant disorders	
Donor-recipient HLA mismatch	2 or more mismatches	
Patient age at transplant	Adult vs. paediatric patient; hazard generally increasing with age	
Donor-recipient sex match	Female donor and male recipient	
Donor-recipient cytomegalovirus (CMV) status	CMV positive patient	

Previous studies have tended to focus on a particular type of blood cancer or non-malignant disorder, restricted to either an adult or paediatric patient group. In my research, I will be investigating adverse events after CB transplantation among patients with a variety of disease types, and both adult and paediatric patients (see Section 1.3 for a description of the real dataset). Therefore, the strength and magnitude of the covariate associations in my analyses may differ from those reported for any one study. Hence, hazard ratios (see Chapter 2, Section 2.2.1) are not reported in Table 1.1.

Double cord transplants (in which the patient receives CB from two different donors) are usually given when a single cord would provide insufficient HSCs for the patient's body weight (14). Double cord transplantation is associated with an increased hazard of acute GvHD, and a decreased hazard of relapse, compared with single cord transplantation (Table 1.1). This may be due to HLA mismatches between the two CB units, inducing a "graft-versus-graft" effect (24). For similar reasons, a higher total nucleated cell (TNC) dose at infusion (described variously as at least 3×10^7/kg or at least 5×10^7/kg) (4, 14, 25, 26) is associated with an increased hazard of acute GvHD and a decreased hazard of relapse.

There are several other covariates for which values associated with increased hazard of acute GvHD are also associated with decreased hazard of relapse (Table 1.1.) (4, 8, 10, 14, 17, 19, 21-24). Specifically, acute GvHD is more likely among patients in remission at time of transplant, among patients receiving intensive pre-transplant radio- or chemotherapy (conditioning), and in transplants in which no GvHD prophylaxis is given. Relapse is less likely in each case. The inverse relationship between hazard of acute GvHD and relapse can be explained by the GvL effect, described previously: acute GvHD promotes the eradication of the underlying blood disease and, hence, leads to a reduction in risk of relapse. In general, some degree of acute GvHD is desirable, although

severe GvHD is often fatal. Achieving the balance between GvHD and relapse is a major challenge for clinicians (14).

Other covariates are generally associated with worse outcomes: hazard of adverse events generally increases with the number of HLA mismatches, age, for a female donor and male recipient, and for a cytomegalovirus-positive (CMV+) patient.

1.3. NHS Cord Blood Bank

The National Health Service (NHS) Cord Blood Bank (CBB) (27) is one of two public CB banks operating in the UK, collecting and storing CB. The NHS CBB is part of NHS Blood and Transplant (NHSBT), which provides services to the NHS related to the donation and transplantation of blood, organs, tissues and stem cells (28).

In 2009, a meeting of senior clinicians acknowledged the lack of UK studies of patient outcomes after CB transplantation, and recommended regular review of UK CB transplant outcomes (29). In 2015, the NHS England Clinical Commissioning Policy for HSC transplantation noted the lack of good quality evidence about HSC transplantation in the UK, which made it difficult to evaluate its effectiveness (30). Furthermore, the Cambridge University Winton Centre for Risk and Evidence Communication (31) notes the importance of communicating the risks and benefits of transplantation to patients in a clear and understandable way. This is exemplified by the patient infographics about characteristics and outcomes of organ transplants using UK donors, which are publicly available on the NHSBT website (32). Despite these recommendations, to date, there have been no in-depth studies of patient outcomes after transplantation using CB donated to UK cord blood banks. In the early 2000s, two preliminary studies of regional cord blood banks in the UK were conducted (33, 34). More recently, a study of the HLA diversity of NHS CBB donations has

been performed (35). In my research, I conduct the first study of engraftment and post-transplant adverse event rates among NHS CBB patients. This important research will benefit NHSBT, patients and clinicians.

CB supplied by the NHS CBB is subject to rigorous quality control, screening, and safety checks, and is compliant with Blood Safety and Quality regulations (27). NHS CBB patient follow-up data are reviewed on a case-by-case basis as part of the Bank's accreditation requirements (36). In addition, regular monitoring of post-transplant event rates (the focus of this thesis), will enable NHSBT to quickly identify, and investigate, any changes. Understanding the patient, donor and transplant characteristics associated with each event will help to identify the cause of such changes. NHS CBB transplants are used to treat patients with a variety of disease types, and both adult and paediatric patients. Changes in event rates over time may be explained by changes in the characteristics of NHS CBB patients. However, large or sudden changes in event rates may indicate a problem with the CB supply and should trigger a rapid investigation by NHSBT. Regular monitoring will also allow NHSBT to assess the impact of any improvements in donation, collection, and storage processes. Furthermore, sharing information about post-transplant event rates will provide reassurance for clinicians and patients that the NHS CBB is a safe and effective source of CB, and will help to inform clinical practice.

Outside the UK, patient outcome studies have been conducted for four CB banks (in Japan, Mexico, Singapore, and Italy) (37-40). In all four studies, post-transplant events were analysed using simple univariate methods (and only three used any survival analysis methods, see Chapter 2, Sections 2.2.1-3 for a description of survival analysis methods). In contrast, in my research, I will consider the sequence of events that patients experience using multi-state model (MSM) analysis (see Chapter 2, Section 2.2.3). The advantage of the MSM approach is that the probability of multiple events can be modelled

simultaneously. This allows more effective communication of the risk of transplant to patients, particularly because the probability of each event at any given time can be illustrated graphically or summarised as the average number of days spent in each state.

1.3.1. NHS Cord Blood Bank data

NHSBT collects data about donors who donate to the NHS CBB. In addition, an international registry, Eurocord, (41) supplies NHSBT with transplant and patient data for transplants using CB donated to the NHS CBB. Transplant and patient data include baseline data (at time of transplant) and data about post-transplant events. Data are described in more detail below and in Chapter 4. For the purposes of this research, I have been given permission by NHSBT and Eurocord to use patient, donor, and transplant data about transplants using CB donated to the NHS CBB, where transplantation occurred between 1996 and 2015.

1.3.2. Data availability

In the NHS CBB dataset, data were available for all covariates listed in Table 1.1. In addition, year and country of transplant were reported. Data were also available about the following post-transplant events: myeloid engraftment, graft failure (the lack of engraftment of donor cells), acute and chronic GvHD, relapse, and death. Graft failure is a "competing risk" for many of the events of interest (see Chapter 2, Section 2.2.2). For each event, whether the event had occurred during the monitoring period for each patient, and the date of onset, were reported. Grades of acute and chronic GvHD were also reported.

In the NHS CBB dataset, several covariates had missing values (see Chapter 4, Section 4.3.2). In addition, the date of onset of an event was sometimes missing. Some information about missing dates of onset could be inferred from clinical criteria (for example, the standard clinical definition of acute GvHD (17) assumes

occurrence between day 0 and 100 post-transplant) or the known length of the monitoring period for each patient (because date of death or last follow-up was always reported, see Chapter 4, Section 4.4). Since each event with a missing event time occurred during a known, finite interval, the missing event times are said to be 'bounded' or 'interval-censored'.

Interval censoring is defined formally as follows: for subject i, event times are known to lie in the interval $(L_i, R_i]$, where L_i represents the last confirmed event-free time and R_i represents the first time at which the event is known to have happened. Left- and right-censoring can be considered special cases of interval-censoring, in which the lower or upper interval boundary takes the value of 0 and ∞, respectively. Studies in dentistry (42) and human immunodeficiency virus (HIV) (43) are typical examples from the literature. For example, Lesaffre and Komarek (42) studied the time to tooth emergence in children. The time of emergence of a particular tooth occurs at some point between the last dental clinic at which no tooth is seen and the first clinic by which the tooth has emerged.

In some HSC transplantation studies, simple strategies have been used to handle missing event times, such as complete case analysis (CCA) (44), or substituting the mean of the observed times for all missing times (45). This can lead to bias and under-coverage (46, 47). Nikolajeva *et al.* (48) performed a logistic regression instead of the usual time-to-event analysis (see Chapter 2, Section 2.2.1-3) due to the large amount of missing acute GvHD times in their study. This approach offers unbiased estimation of the incidence if there is complete follow-up for all patients during the acute post-transplant period but does lead to a loss of information, for example, about the median event time.

In later chapters, I will describe more rigorous methods for handling missing event times and will compare their performance using simulation studies. Subsequently, I will apply the best method in my analysis of the NHS CBB

dataset (see Section 1.4 for a full overview of my thesis). My research represents the first comparison of these rigorous methods for handling missing event times (namely, multiple imputation (MI) and full maximum likelihood (FML) methods), as well as the first time that MI methods have been compared in a MSM analysis. It also represents the first analysis of HSC transplantation outcomes using MI methods.

1.4. Thesis Aims and Objectives

My thesis aims and objectives are summarised below.

1.4.1. Thesis Aims

(i) To describe the incidence of myeloid engraftment, acute and chronic GvHD, and relapse, as well as overall survival, among NHS CBB patients after CB transplantation.

(ii) To identify covariates associated with acute GvHD, relapse, and death following CB transplantation, and to describe the probability of these events for different patient types.

1.4.2. Thesis Objectives

(i) Compare methods for handling missing event times in a competing risks analysis, using simulation studies.

(ii) Compare methods for handling missing event times in a MSM analysis, using simulation studies.

(iii) Apply the best method(s) for handling missing covariate data and event times to the NHS CBB dataset.

(iv) Calculate the incidence of each event of interest in the NHS CBB dataset, using methods that allow for the presence of competing risks.

(v) Identify covariates associated with the events of interest by conducting a MSM analysis of the NHS CBB dataset.

(vi) Calculate state occupation probabilities and the expected length of stay in each state (transplant, acute GvHD, relapse/death) in the first-year post-transplant, for different patient types in the NHS CBB patient cohort.

1.4.3. Thesis Outline

The outline of my thesis is as follows:

In Chapter 2, I describe the statistical methods used in this thesis. In Chapter 3, I review the existing methods and concepts for handling missing data in detail. In Chapter 4, I give an overview of the NHS CBB dataset and outline the possible missingness mechanisms for event times. In Chapters 5 and 6, I describe simulation studies comparing different methods for handling missing event times in a competing risks analysis (Chapter 5) and an illness-death MSM (Chapter 6). In Chapter 7, I apply methods for handling missing data in an analysis of the NHS CBB dataset. I conclude with general discussion in Chapter 8.

CHAPTER 2. STATISTICAL METHODS

2.1. Introduction

In this chapter, I describe the statistical methods used in my research. I provide an overview of survival analysis, and a description of the survival models and associated estimands used in my analyses. Here (and in more detail in Chapters 5 and 6), I briefly describe my approach to simulating event time data, the benefits of my simulation approach, and the challenges of simulating survival data.

2.2. Survival Models

In my analysis of the events experienced by patients in the NHS CBB dataset, I consider two types of survival model:

(i) A competing risks model, to estimate the cumulative incidence of each event of interest.

(ii) An "illness-death" MSM, to identify covariates associated with the hazard of the events of interest, to calculate state occupation probabilities, and to calculate the expected number of days spent in each state.

In all analyses, I assume that event time distributions are continuous and that the censoring distribution is independent of the event time distribution (see Section 2.2.1). In the following sections, I describe basic concepts of survival data (Section 2.2.1), as well as the two types of survival model I have used and methods for their analysis (Sections 2.2.2 and 2.2.3).

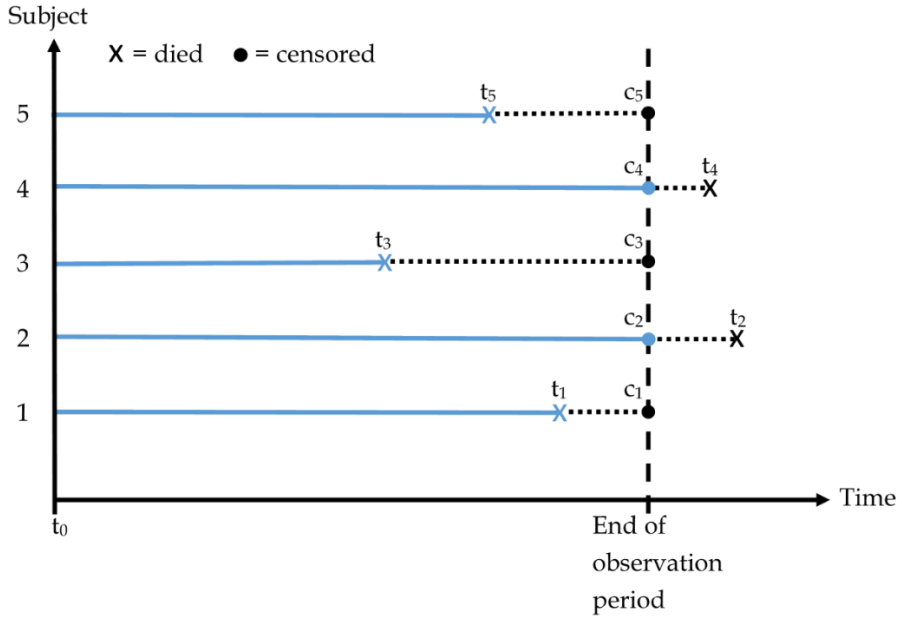
2.2.1. Measuring the time to an event

Measuring the time until an event occurs is of interest in many contexts. The

classic application of time to event methods is the mortality model, measuring time to death. In this case, all study subjects will experience the event of interest at some point, although, typically, not all subjects in the study will have died during the observation period. Subjects still alive at study end are said to be "right-censored". For brevity, the term "censored" will be used hereafter to refer to right-censoring. Other types of censoring are left-censoring (when the time origin is not observed) and interval-censoring (described previously in Section 1.3.2). Formally, the model for censored data is as follows (49):
For each subject i, with an event time t_i and a censoring time c_i, the observed data are time $x_i = \min(t_i, c_i)$ and censoring indicator δ_i, with $\delta_i = 1$ if the event time is observed and 0 otherwise. The event and censoring times are considered random samples from a survival distribution $T_i \sim S$, where $S(t) = P(T > t)$, and a censoring distribution $C_i \sim G$.

To illustrate the concept of censoring, a set of censored data are shown in Figure 2.1, below.

Figure 2.1. Illustration of censored data.



Solid lines and points represent the observed data. Dotted lines represent unobserved data.

In Figure 2.1, subjects 1, 3 and 5 died before the end of the observation period and hence, the event time is observed for these subjects: $x_i = t_i$ and $\delta_i = 1$ in each

case. Subjects 2 and 4 died after the end of the observation period and hence, the event time is not observed for these subjects: $x_i = c_i$ and $\delta_i = 0$ in each case. Note that no left- or interval-censoring occurred (the time origin, t_0, is observed for all subjects and event times are observed exactly). In addition, in this example, the end of the observation period is the same for all subjects; in reality, the censoring time may vary between subjects.

Generally, methods that can accommodate censoring assume independent censoring and the absence of left- or interval-censoring. Independent censoring means that the distribution of censoring times is independent of the distribution of event times (this may be conditional on covariates) (50-52). In other words, at each time-point, the subjects who are still under observation are representative of the censored subjects.

Estimands of interest

In survival analysis, describing the survivor function $S(t)$ (the probability of surviving until at least time t) and the associated hazard of the event $\lambda(t)$ (the instantaneous probability of an event at time t for a subject who is event-free until that time) is often of interest. When all subjects will experience the event of interest at some point (as in the mortality model, for example), the survivor function and hazard are directly related, as follows:
Given $\lambda(t)$, the cumulative hazard $\Lambda(t) = \int_0^t \lambda(s)ds$ and the survivor function $S(t) = \exp(-\Lambda(t))$ (53). The standard estimator of the survivor function is Kaplan and Meier's non-parametric maximum likelihood estimator (NPMLE) (50). Kaplan-Meier estimates of the survivor function can be compared between independent groups using the log-rank test (54).

Modelling the association between covariates and the hazard

Frequently used models for exploring the association between covariates and the hazard of an event are Cox's semi-parametric proportional hazards (PH)

regression model (51) and Aalen's additive hazards model (52). More recently, flexible parametric hazards models have been proposed (55, 56).

In my MSM analyses, I use both Cox and parametric (Weibull) PH models. The structure of a PH model is as follows:
Given time-fixed covariates $\boldsymbol{x}$, then $\lambda(t \mid \boldsymbol{x}) = \lambda_0(t) \exp\{\boldsymbol{\beta' x}\}$. $\lambda_0(t)$ is referred to as the "baseline" hazard function, that is, the hazard function for a subject with baseline, or reference, values of all covariates (so $\boldsymbol{x} = \mathbf{0}$).

The appeal of the PH model is the ease of interpretation of $\boldsymbol{\beta}$: consider a model with a single binary covariate ($x = 0$ or 1). Then the ratio of the hazards for a subject with $x = 1$ compared with a subject with $x = 0$ is $\frac{\lambda_0(t) \exp\{\beta\}}{\lambda_0(t)} = \exp\{\beta\}$. In a PH model, $\exp\{\beta\}$ is usually the principal estimand. It is referred to as the "hazard ratio" (HR).

In a Cox model, the vector of regression coefficients $\boldsymbol{\beta}$ is estimated by maximising a partial likelihood function (51). The advantage of Cox's approach is that estimation of $\boldsymbol{\beta}$ does not require estimation of $\lambda_0(t)$, although $\lambda_0(t)$ can be calculated non-parametrically if desired (57). Alternatively, a parametric model for $\lambda_0(t)$ can be explicitly defined (in my analysis, using a Weibull model). In this case, baseline model parameters and the vector of regression coefficients $\boldsymbol{\beta}$ are estimated together using full maximum likelihood estimation (58).

2.2.2. Competing risks

In the previous section, it was assumed that all subjects would experience the event of interest at some point. However, "competing risks" can occur when the occurrence of a "competing" type of event precludes or changes the probability of the event of interest. This could be because only one of a set of mutually exclusive events can be experienced, *e.g.* death due to cancer is a competing risk for transplant-related death: prior to death, a patient is at risk of both causes of

death, but ultimately can only experience one or the other. Competing risks can also occur when only the time of the first event experienced is of interest, *e.g.* if acute GvHD is the event of interest, then death prior to acute GvHD is a competing risk.

Consistent with the clinical literature (45, 59), I define death prior to the event of interest as a competing risk for myeloid engraftment, acute and chronic GvHD, and relapse. In addition, I define graft failure prior to the event of interest as a competing risk for myeloid engraftment, acute GvHD, and chronic GvHD. Again, for consistency with the clinical literature (45, 59), I do not consider relapse a competing risk for engraftment nor GvHD. However, engraftment and GvHD are unlikely after relapse (and never occurred after relapse for subjects in the NHS CBB dataset, see Chapter 4, Section 4.4.1). Hence, in my MSM analysis, I treat both relapse and death as terminal events (see Section 2.2.3 below, Chapter 6, Section 6.2.2, and Chapter 7, Section 7.3).

Estimands of interest

In the competing risks framework, the cumulative incidence function $I_j(t)$ describes the probability of experiencing an event of type j by time t. Formally, it is defined as follows:

For the j^{th} of n_j event types, the cumulative incidence $I_j(t) = \int_0^t \lambda_j(s)\, S(s) ds$

where $\lambda_j(t)$ is the cause-specific hazard for the j^{th} event type,

$S(t) = \exp\left(- \sum_{j=1}^{n_j} \Lambda_j(t) \right)$ represents the probability of not having experienced any event by time t and $\Lambda_j(t) = \int_0^t \lambda_j(s) ds$.

In my analysis, I use the unbiased NPMLE proposed by Aalen and Johansen (60) to estimate the cumulative incidence. This estimator is defined as follows:

$\hat{I}_j(t) = \sum_{m:t_m \leq t} \hat{\lambda}_j(t_m)\, \hat{S}(t_{m-1})$

where $\hat{\lambda}_j(t_m) = \frac{d_{jm}}{n_m}$ and $\hat{S}(t_m) = \prod_{m:t_m \leq t} \left(1 - \frac{d_m}{n_m}\right)$,

$0 < t_1 < \ldots < t_M$ are the set of ordered event times of any event type,

d_{jm} denotes the number of subjects experiencing an event of type j at time t_m,

d_m denotes the number of subjects experiencing an event of any type at time t_m,

and n_m denotes the number of subjects at risk (*i.e.* still in follow-up and event-free) at time t_m. Note that the estimator $\hat{S}(t_m)$ defined here is the Kaplan-Meier NPMLE of the survivor function. Aalen-Johansen estimates of the cumulative incidence function can be compared between independent groups using Gray's test (61).

In an early CB transplantation study (34), the cumulative incidence of acute GvHD was estimated as $1 - \hat{S}_j(t_m)$ (where $\hat{S}_j(t_m)$ denotes the "cause-specific" Kaplan-Meier survival estimate, treating only cases of acute GvHD as events and censoring any cases of graft failure or death prior to acute GvHD). This approach will give a biased estimate of the cumulative incidence, because the independent censoring assumption of the Kaplan-Meier method no longer holds (49). This is because subjects who have experienced graft failure or death prior to acute GvHD have zero probability of experiencing acute GvHD at any future time-point. Hence, subjects still event-free at this future point are not representative of censored subjects (apart from in the special case when all cases of acute GvHD occur before any cases of graft failure or death).

Modelling the association between covariates and the cumulative incidence

The association between covariates and the cumulative incidence can be explored indirectly using a cause-specific hazards model (51). An advantage of this approach is that standard hazards models (such as those mentioned in Section

2.2.1) can be applied. In addition, estimates from hazards models for each event type can be combined, using a MSM approach (49) (see Section 2.2.3), to calculate cumulative incidence functions. However, there is not a direct relationship between the cause-specific hazards and the cumulative incidence function (in contrast to the mortality model described in Section 2.2.1). This is because the cumulative incidence of any one event type depends on *all* the cause-specific hazards functions (see definition on previous page). Therefore, a covariate value associated with an increased cause-specific hazard will not necessarily be associated with an increased cumulative incidence (49). This makes it difficult to interpret the cause-specific HRs.

Alternatively, the association between covariates and the cumulative incidence can be explored directly using models such as the Fine and Gray proportional "sub-distribution hazards" model (62). The sub-distribution hazard for an event of type j is defined by Fine and Gray (62) as $-\frac{d \log\left(1 - I_j(t)\right)}{dt}$. A criticism of Fine and Gray's model is that subjects are still included in the risk set for the event of interest even after they have experienced a competing event (63). In addition, the Fine and Gray model is designed to estimate the cumulative incidence of only one event of interest (with other event types treated as nuisance factors). Hence, Fine and Gray models for the cumulative incidence functions of all possible event types are not constrained to sum to one (64, 65).

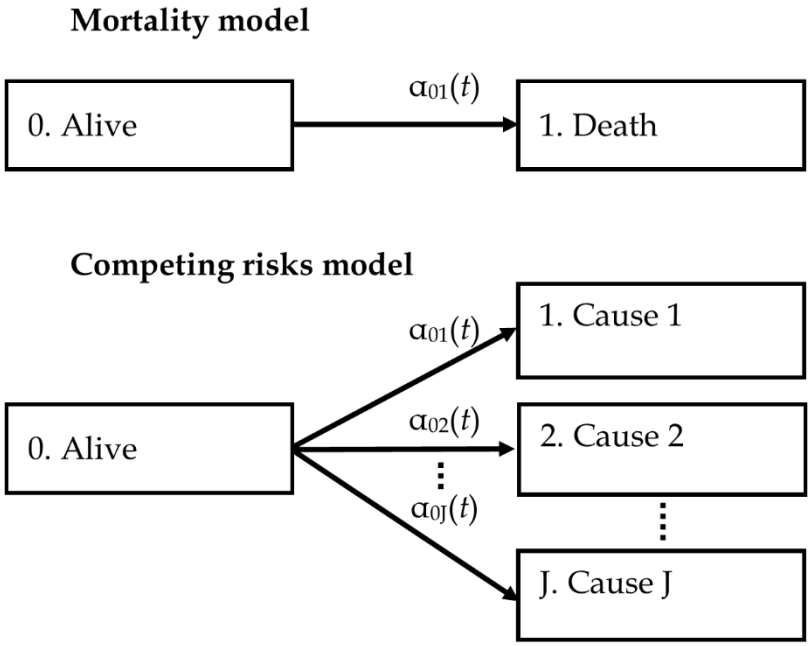
To overcome these disadvantages, Putter *et al.* (63) describe an approach for linking the cause-specific hazards and sub-distribution hazards functions using a "reduction factor". Their approach is designed to correct deficiencies in the Fine and Gray model, rather than in the MSM approach. In addition, it relies on correct specification of the reduction factor. In my comparison of methods for handling missing event times in a competing risks analysis (Chapter 5), I include a FML method based on the Fine and Gray model (64). However, due to the limitations of the Fine and Gray approach, I focus on MSM methodology (see

Section 2.2.3) when exploring the association between covariates and the events of interest, and when predicting event probabilities for subjects with different sets of characteristics (using a simulation study in Chapter 6, and in analysis of my real data in Chapter 7).

2.2.3. Multi-state models

Both mortality and competing risks models consider time until the first event experienced. A MSM is a generalisation of these models, allowing the sequence of events (the "event history") experienced by each subject to be analysed. In the MSM framework, experiencing an event can be thought of as a move ("transition") from one "state" to another. Figure 2.2 shows MSMs equivalent to the mortality and competing risks models described previously. States are represented by rectangles and possible transitions by arrows. The transition intensity, $\alpha_{hj}(t)$ is shown for each transition. Analogous to the hazard rate, described previously, $\alpha_{hj}(t)$ is defined as the instantaneous probability of moving from state h to state j at time t.

Figure 2.2. Multi-state models for mortality and competing risks data

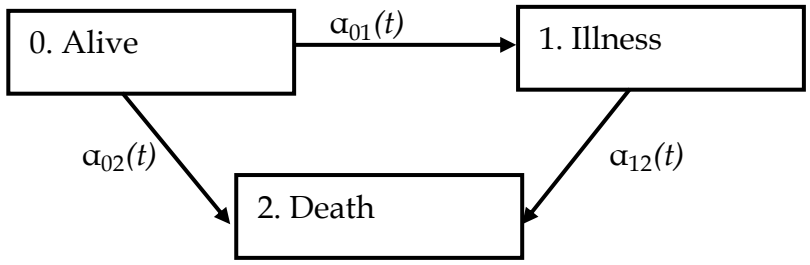


In both models shown in Figure 2.2, there is a single initial state, and all subjects are in this state at the time origin, t_0. In the mortality model, there is a single

"terminal" or "absorbing" state, that is, a state from which further transitions cannot occur. In the competing risks model, there is more than one absorbing state.

In more complex MSMs, there are also "intermediate" states between initial and absorbing states. The simplest MSM including an intermediate state is the unidirectional three-state "illness-death" model, shown in Figure 2.3 below. Klein *et al.* (66) advise that although intermediate events can be included as time-varying covariates in time to first event models, interpretation of the associated parameter estimates and comparison of survival probabilities can be difficult. They argue that if the main question of clinical interest is how the probabilities of death or other events depend on the patient's history, then the MSM approach is more useful.

Figure 2.3. The illness-death model



Markov models

The calculation of transition intensities and related probabilities is most straight-forward for MSMs with the Markov property. This property states that the transition intensity depends only on the current state occupied but not the amount of time spent in the current state nor the past history prior to entry into the current state. The Markov property implies a "clock-forward" structure (time *t* is measured from entry into the initial state) for all transitions (49). Assuming there is no left-censoring, the mortality model and competing risks model are always Markovian, because there is no event history (67). For simplicity, and ease of calculation, only Markov models are considered in my analyses.

I use a simple test of the Markov property in my analyses. In this test, a term for the time until entry into the current state is included in the transition intensity model (68). Recently, Titman and Putter (69) compared this simple test with a new class of tests of the Markov property. They found that the simple test maintained good power in a range of situations.

Estimands of interest

Estimation of transition intensities is usually of interest in MSM analysis. These can be estimated using standard hazards models (67). These models can also be used to explore the association between covariates and the transition intensities. In my analyses, as described in Section 2.2.1, I use Cox and Weibull PH models to explore these associations.

Transition probabilities are also of interest in MSM analysis. The transition probability, $P_{hj}(s,t)$, is defined as the probability of being in state j at any time t, conditional on having previously been in state h at time s. For a Markov model, the matrix of transition probabilities is calculated from the transition intensities (70) as follows:
$\mathbf{P}(s,t) = \prod_{(s,t]}(\mathbf{I} + d\boldsymbol{H}(u))$ where $\boldsymbol{H}(u)$ is the matrix of cumulative transition intensities (71). Note that, for a competing risks model, the transition probabilities are equivalent to the cumulative incidence functions.

Related estimands that I consider in my analyses (see Chapter 6, Section 6.2.10 for further details) are:

(i) the "state occupation" probability, $P_j(t)$, the probability of being in state j at time t (72), calculated as follows: $P_j(t) = \sum_h P_h(0)P_{hj}(0,t)$.
If all subjects are in state 0 initially, *i.e.* $P_0(0) = 1$, $P_j(t)$ is equivalent to the transition probability from state 0 to state j at time t (73).

(ii) The restricted expected length of stay (RELOS) in each state (72). RELOS

from time 0 to time s for state j is defined as:

$$e_j(s) = \int_0^s P_j(t)\ dt$$

Calculation of transition probabilities becomes more difficult for semi-Markov (where the transition intensity depends on the time spent in the current state) and non-Markov models (where the transition intensity depends on the amount of time spent in the current state and the past history prior to entry into the current state) (74). Hence, these models are out of the scope of my research.

2.3. Simulating event time data

When simulating event time data, valid inference is only possible when the data-generating mechanism (DGM) is compatible with the desired analysis model (75). Early methods for simulating event time data, given multiple event types, were based on a multivariate model (76). In this model, also called the "latent failure time" model (77), each subject is assumed to have an event time for each possible event type. The observed data are defined as the minimum of all potential event times (with all unobserved event times regarded as latent variables), plus an indicator of which event type happened first. However, this model is not identifiable (49) because the covariance structure of the latent failure times cannot be determined using the observed data (77). This model is also implausible, because, in a competing risks framework, a subject can only experience one of a set of mutually exclusive events. In other words, in a competing risks framework, the event time represents a single random variable, rather than a set of potential random variables. Therefore, the latent failure time approach is incompatible with a competing risks framework. In my simulation studies, I use approaches for generating event time data that are compatible with my chosen analysis models. These are summarised below.

2.3.1. Model 1: Non-parametric estimation of the cumulative incidence function

For this simple analysis model, simulated event times were generated using two different methods (see Chapter 6, Section 6.2.3 for more details):

(i) Sampling with replacement from a full data version (*i.e.* a version with no missing event times) of the real NHS CBB dataset.

(ii) Direct draws from parametric distributions (with the choice of distribution dependent on the event type experienced).

Table 2.1 overleaf summarises the advantages and disadvantages of each of these DGMs. By using two different methods, I was able to assess whether my simulation study results were sensitive to the choice of DGM.

Table 2.1. Comparison of two different sampling methods when generating data

Characteristics of the generated data	**Sampling method**	
	Sampling from real data	**Sampling from parametric distributions**
Representative of real data	Yes	May include elements of the real data *e.g.* parameter choice guided by real data, but usually fewer covariates and simpler covariance structure than in real dataset. May be difficult to identify a parametric distribution with characteristics similar to those of the real data *e.g.* observed survivor function may not be well fitted by any standard parametric distribution.
Simulated datasets are independent samples	Cannot be assumed, unless the real dataset is very large	Yes
Results are reproducible/generalisable	Results, *e.g.* size and direction of bias, may be data-specific	Results may be data-specific, but easy to explore the sensitivity of results to parameter choice
Size of the simulated dataset	Increasing beyond the size of the real dataset may lead to under-estimation of sample variance	Any size
Estimands of interest	May be limited by the characteristics of the real dataset *e.g.* cannot choose the median time of an event as an estimands if experienced by less than 50% of subjects in the real dataset	Any, within reason

2.3.2. Model 2: MSM analysis including an intermediate event and covariates

In my MSM analysis, to ensure my DGM was compatible with a MSM structure, I used the approach of Beyersmann *et al.* (77) (see Chapter 7, Section 7.2.3-8 for more details). This approach is based on specification of the cause-specific hazards, which can have any desired structure (I used PH models including covariates). Simulated event times can then be generated from the all-cause hazards function (the sum of the cause-specific hazards). The associated event type is then generated by drawing from a binomial distribution (or multinomial if there are more than two competing event types), with probability of each event type dependent on the cause-specific hazards at the simulated event time.

By applying this method in two sequential competing risks experiments, I was able to generate an event history for each subject that was consistent with a three-state illness-death model (as in Figure 2.3). An additional challenge of my model was that, for each subject experiencing the "illness" event, their simulated "death" time had to be greater than their simulated "illness" time. To achieve this, I specified a conditional survival function for the "death" time.

It should be noted that this DGM will generally not be compatible with a proportional sub-distribution hazards model (due to the indirect relationship between hazards and sub-distribution hazards, discussed previously). However, Beyersmann *et al.* (77) suggest that the sub-distribution hazard ratio from this (misspecified) proportional sub-distribution hazards model can be interpreted as a time-averaged sub-distribution hazard ratio. Alternatively, if a proportional sub-distribution hazards analysis is desired after generating event time data using Beyersmann's method, the "reduction factor" approach suggested by Putter *et al.* (63) could be explored.

2.4. Summary

In this chapter, I have described the statistical methods used in my research, my approach to simulating event time data, the benefits of my simulation approach, and the challenges of simulating survival data. In Chapter 5, I use the methods described here to simulate and analyse competing risks data, in scenarios in which some event times are missing. In Chapter 6, I repeat this process using a MSM analysis. In Chapter 7, I apply the best MI methodology from my simulation studies in analysis of the NHS CBB dataset (using competing risks and MSM analysis models). In the next chapter (Chapter 3), I describe missing data methods and concepts.

CHAPTER 3. REVIEW OF METHODS FOR HANDLING MISSING DATA

3.1. Introduction

Missing covariate and outcome data is a common problem when conducting an analysis. Many methods have been developed to address the potential bias caused by missing data. Naïve methods include treating missing data as a separate category (if missing data are in an exposure or covariate) or replacing the missing value with a single value, such as the average of the observed data. However, these methods run a high risk of producing biased parameter estimates and underestimating variance (46). This chapter comprises a review of more rigorous methods for handling missing data.

The methods discussed in this chapter are summarised in Table 3.1, overleaf, including any restrictions on their use. This is followed by a general overview of missing data methods and concepts in Sections 3.2-3.4, with an emphasis on the most widely used approach for handling missing data: multiple imputation (MI). MI methods most pertinent to my research, that is, those for handling missing event times in survival analyses, are described in more detail in Section 3.5. In the NHS CBB dataset, missing event times are assumed to have occurred in a known, finite, time period (hence, can be considered interval-censored). Therefore, in Section 3.6, I give a brief overview of full maximum-likelihood (FML) methods for interval-censored event times.

Table 3.1. Methods for handling missing data

Missing data method	**Analysis model**	**Imputation model**	**Missing data mechanism**	**Restrictions**
Complete case analysis	Any	None	MCAR, MAR, MNAR	Generally requires that missingness does not depend on the analysis outcome
Weighting methods	Any	None	MCAR, MAR	Sensitive to the choice of weighting model; performs best with a single incomplete variable/monotone missing data, and large sample size
Multiple imputation	Any	Any	MCAR, MAR	Missing data mechanism is ignorable; imputation and analysis models are compatible and correctly specified
	Any	Joint model across all incomplete variables		The joint model must be specifiable: multivariate linear, log-linear, general location model
		Fully conditional specification (FCS): univariate model for each incomplete variable		None, given data MAR (conditional on the observed data)
	Non-linear relation-ship	Impute then transform and, relatedly, passive imputation		Analysis results tend to be biased towards the null
		Transform then impute (just another variable)		Imputed values may be inconsistent with one another; may perform badly when analysis is logistic regression and/or data are MAR
		Substantive model compatible (SMC)-FCS; stacked FCS MI		Computationally intensive; current software restricted to linear, logistic, Cox regression and multilevel analysis models
	Skewed variables	Impute on the original scale using a parametric imputation model		May lead to bias if non-linear association between incomplete covariate and outcome
		Impute on the original scale using predictive mean matching (PMM)		Type 1 PMM generally unbiased, even with non-linear associations; type 2 PMM leads to some bias
		Transform then impute		Analysis results tend to be biased towards the null
	Bounded variables	Truncate imputed values during or post-imputation	MCAR, MAR	Analysis results tend to be biased

Missing data method	Analysis model	Imputation model	Missing data mechanism	Restrictions
Multiple imputation contd.	Linear/ logistic regression	MI then deletion with same model for imputation and analysis	MCAR, MAR – incomplete outcome, covariates fully observed	Biased if auxiliary data available and outcome MAR
	Survival analysis	Should include the censoring indicator and a representation of the survivor function	MCAR, MAR	Less biased using cumulative baseline hazard than survival or log-survival time to represent the survivor function
		SMC-FCS for a Cox regression analysis; stacked FCS MI		Unclear if applicable to incomplete outcome as well as incomplete covariates
		Based on the survivor function – parametric/semi-parametric model or PMM		For imputation of missing/censored survival times; generally, parametric/semi-parametric methods developed to-date cannot be combined with imputation of incomplete covariates
Full maximum likelihood methods	Survival analysis	None	MCAR, MAR	For analysis of interval-censored survival times; generally cannot handle a mixture of exact and interval-censored times, or wide intervals relative to the change in hazard

MCAR, missing completely at random; MAR, missing at random; MNAR, missing not at random – see Section 3.2 for full definitions

3.2. Complete case analysis

The simplest solution to missing data is to exclude cases with missing values. This is commonly known as complete case analysis (CCA) or case deletion (46, 78). CCA can be inefficient (78), especially if data are missing across multiple variables. If the probability that data are missing is independent of the observed and missing data (missing completely at random, MCAR), CCA is equivalent to taking a random sample of the data and hence produces unbiased estimates. More generally, estimates from CCA are unbiased as long as missingness is not related to the analysis outcome, given the covariates in the analysis model (79). If this is true, CCA is valid even if the probability that data are missing depends on the observed data (missing at random, MAR) or depends on the values of the missing data themselves, given the observed data (missing not at random, MNAR). Bartlett *et al.* (80) describe certain circumstances in which CCA using a logistic regression model will allow unbiased estimation of the exposure odds ratio (even if missingness is related to the analysis outcome). Namely, if missingness is related only to the analysis outcome, or only to the analysis outcome and covariates. However, this does not apply if the binary outcome was generated from a continuous measure and missingness depends on the continuous measure. Bartlett *et al.* note that these results are also applicable to Cox regression analysis, providing follow-up is the same for all subjects and the event rate is low.

Galati and Seaton (81) argue that, regardless of the missingness mechanism, CCA is valid if cases are "available at random", such that the probability that all data are observed is constant across all covariate values. They give the example of participants in a trial who may be over-representative of the sickest patients, but where dropouts are also over-representative of the sickest patients. Hence the combination of selection bias and dropout might result in a roughly representative sample in the complete cases.

3.3. Weighting methods

Weighting adjustments can be used to correct suspected bias from CCA (82). This can be achieved by stratifying the complete case dataset into weighting classes, or more generally, by using inverse probability weighting (IPW), where weights are obtained from a model for the probability that data are complete (the weighting model). IPW relies on the assumption that the weighting model is correctly specified, and that the probability of a complete case is non-zero (83). However, in contrast to multiple imputation (see Section 3.4.2), it does not require compatibility between the weighting model and the analysis model. Since weighting methods use only complete cases, they can be inefficient (82). They can also be sensitive to the choice of weighting model and difficult to apply in practice if variables relating to missingness are themselves incomplete (84). Therefore, weighting methods are most useful when a single variable is incomplete or when data are monotone missing, and the sample size is large (82).

3.4. Multiple imputation

MI methods (and related approaches, such as Bayesian methods) are frequently used to handle missing data, with potential benefits in terms of both bias-correction and precision (85). MI involves a three-step process. First, missing data are replaced by plausible values using some specified mechanism. To account for the uncertainty about missing values, multiple copies of the dataset are created (86). White *et al.* (87) suggest as a "rule of thumb" that the number of copies of the dataset should at least equal the percentage of incomplete cases. Secondly, the desired analysis is performed separately for each imputed dataset. Finally, results from each separate analysis are pooled to obtain overall estimates of the mean and variance of analysis model parameters; Rubin's rules (88) are the standard method for this. Madley-Dowd *et al.* (89) found that, even with a large proportion of missing data, MI was beneficial in terms of reducing bias and improving efficiency compared with CCA. They note that the "fraction of

missing information", a parameter-specific measure of the information lost due to missingness, gives a better indication of the efficiency gains from using MI than the proportion of missing data.

Van Buuren (90) describes two broad categories of multiple imputation methods: (i) joint modelling (JM) and (ii) multiple imputation using fully conditional specification (FCS, also referred to as "chained equations") (91). In each method, missing values are replaced with realisations of the posterior distribution of a parametric model, known as the imputation model (92). JM involves the explicit specification of a joint parametric model across all incomplete variables. Common choices are the multivariate linear or log-linear distributions.

However, where an appropriate joint model cannot be easily specified, FCS offers a more flexible alternative. In this method, rather than drawing directly from a joint model, a separate model is specified for each incomplete variable, conditioning on all other observed and imputed data. The process uses an iterative algorithm (93, 94), described briefly as follows:

For each incomplete variable X_j, let $\mathbf{X}_{-j}$ represent the set of variables excluding X_j, and x_j^{obs} and x_j^{mis} represent the observed and missing data respectively for variable X_j. Then, for each X_j, specify an imputation model with probability distribution function $p(x_j \mid \boldsymbol{x}_{-j}, \theta_j)$, and prior distribution $p(\theta_j)$ for the (unknown) model parameter θ_j. Assume, without loss of generality, that variables $X_1,\ldots,X_R$ are incomplete and $X_{R+1},\ldots,X_J$ are fully observed ($R \leq J$).

(i) Each imputation consists of T iterations. In the first iteration, arbitrary starting values are assigned to all missing data.

(ii) Iterations $t = 2,\ldots T$, consist of the following draws:

$$\theta_1^{(t)} \sim p(\theta_1)\, p(x_1^{obs} \mid x_2^{(t-1)}, x_3^{(t-1)}, \ldots\ldots, x_R^{(t-1)}, x_{R+1}, \ldots\ldots, x_J, \theta_1)$$

$$x_1^{mis(t)} \sim p(x_1^{mis} \mid x_2^{(t-1)}, x_3^{(t-1)}, \ldots\ldots, x_R^{(t-1)}, x_{R+1}, \ldots\ldots, x_J, \theta_1^{(t)})$$

(continued overleaf)

$\theta_2^{(t)} \sim p(\theta_2)\, p(x_2^{obs} \mid x_1^{(t)}, x_3^{(t\text{-}1)}, \ldots\ldots, x_R^{(t\text{-}1)}, x_{R+1}, \ldots\ldots, x_J, \theta_2)$
$x_2^{mis(t)} \sim p(x_2^{mis} \mid x_1^{(t)}, x_3^{(t\text{-}1)}, \ldots\ldots, x_R^{(t\text{-}1)}, x_{R+1}, \ldots\ldots, x_J, \theta_2^{(t)})$
.
.
.
$\theta_R^{(t)} \sim p(\theta_R)\, p(x_R^{obs} \mid x_1^{(t)}, x_2^{(t)}, \ldots\ldots, x_{R\text{-}1}^{(t)}, x_{R+1}, \ldots\ldots, x_J, \theta_R)$
$x_R^{mis(t)} \sim p(x_R^{mis} \mid x_1^{(t)}, x_2^{(t)}, \ldots\ldots, x_{R\text{-}1}^{(t)}, x_{R+1}, \ldots\ldots, x_J, \theta_R^{(t)})$

(iii) Values from the T^{th} iteration are retained as the imputed dataset. The process is repeated M times to create M imputed datasets.

Due to its flexibility, the FCS method is particularly useful when incomplete variables are a mixture of continuous, binary and categorical variables or when constraints are imposed on some variables. In the special cases where the univariate imputation models are all linear regression models with no interaction terms, or all log-linear regression models with only two-way interactions, the FCS and JM approaches are equivalent (90). More generally, Seaman and Hughes (95) argue that for a correctly specified restricted general location model (a loglinear model for the categorical variables and a multivariate linear regression model for the continuous variables, with additional restrictions that any conditional distribution is a univariate regression with main effects only), posterior distributions under the JM and FCS will be asymptotically equivalent. They found that although the JM approach can be more efficient in this case, it is also less robust to misspecification. Valid inference using FCS is based on the assumption that a joint model can be defined. A criticism of FCS is that, other than in the few special cases previously described, this implicit assumption cannot be verified (and is unlikely to be true in applications to real data).

Rather than replacing missing data directly with draws from a parametric model, predictive mean matching (PMM) has been suggested as an alternative imputation method (96). A combination of PMM and imputation using direct draws from parametric models can be specified within the same FCS imputation scheme (with one or other chosen for each incomplete variable). In PMM, for

each subject *i* with a missing value for variable X, the following steps are performed:

(i) Calculate the distance between the predicted values of X for subject *i* and for all *h* subjects with an observed value for X (see below for definitions of the predictive distance).

(ii) Identify a donor pool of subjects (usually fixed in size) for which the predictive distance is minimised.

(iii) Randomly select a subject *d* from the donor pool.

(iv) Replace the missing value of subject *i* with the observed value of subject *d*.

Morris *et al.* (92) describe three types of PMM, which differ in how the predictive distance in step (i) is calculated. In type 1 PMM, the predictive distance is calculated as $\left|\boldsymbol{\beta}^{*}\boldsymbol{z_i} - \widehat{\boldsymbol{\beta}}\boldsymbol{z_h}\right|$, where $\widehat{\boldsymbol{\beta}}$ denote the estimates of regression coefficients $\boldsymbol{\beta}$ from a regression of X on predictors $\mathbf{Z}$, fitted to all *h* observed values of X; $\boldsymbol{\beta}^{*}$ denotes a random draw from the posterior distribution of $\boldsymbol{\beta}$ (or its Normal approximation) (87) and $\boldsymbol{z_i}$ and $\boldsymbol{z_h}$ denote the values of $\mathbf{Z}$ for subjects *i* and *h*, respectively. In type 0 the predictive distance is $\left|\widehat{\boldsymbol{\beta}}\boldsymbol{z_i} - \widehat{\boldsymbol{\beta}}\boldsymbol{z_h}\right|$ and in type 2 it is $|\boldsymbol{\beta}^{*}\boldsymbol{z_i} - \boldsymbol{\beta}^{*}\boldsymbol{z_h}|$. Several studies have found that type 2 PMM leads to under-coverage (92, 97, 98).

In PMM, imputed values are always selected from the pool of observed data, which may be an advantage when imputing discrete data or when the analysis model includes non-linear terms. A related method, "local residual draws" (LRD), proceeds as for PMM, but the imputed value is the sum of the residual for donor *d* (the difference between the observed and predicted value) and the predicted value for subject *i* (92). Both PMM and LRD may protect against misspecification of the imputation model. However, Morris *et al.* (92) suggest that both methods can perform badly when there is a strong association between the incomplete covariate and the outcome of interest. They suggest focus should be on improving the specification of the imputation model rather than using PMM or LRD as a way of correcting for a poorly specified model.

3.4.1. Ignorability

Most MI methods developed to date assume that the missing data mechanism (MDM) is ignorable, that is, that the parameters of the analysis model are distinct from the parameters of the MDM (79, 99). Generally, ignorability holds for data MCAR and MAR. If data are MNAR, the missingness mechanism is always non-ignorable. In contrast to CCA, for which non-ignorability can be accommodated under certain conditions, imputation with a non-ignorable MDM requires the correct specification of the missingness mechanism and this is usually unknown. Where a non-ignorable missingness structure is suspected, Little (79) suggests an approach called subsample ignorable likelihood (SIL). This combines CCA with the increased efficiency of MI. In SIL, a subsample of cases is identified in which all subjects have observed values of the MNAR covariates. For this subsample, missingness in other variables is assumed to be ignorable and hence, MI can be performed without explicitly modelling the missingness mechanism. Provided that the missingness mechanism for covariates MNAR is not related to the analysis outcome, Little shows that SIL is unbiased in scenarios where CCA or MI using all data may not be. However, as in MI, the validity of the method relies on specific assumptions about the MDMs, which are usually unverifiable.

White and Carlin suggest using subject-matter knowledge to decide whether the MDM is likely to be ignorable and, if in doubt, perform both CCA and MI (100). If inferences are different for each, they recommend careful investigation to determine why the differences occur. A flaw in this advice is that if data are MNAR and missingness depends on the analysis outcome, then both approaches will give biased results. In this case, a suitable method for handling data MNAR should be used instead (85). Hence, knowledge of the likely MDM is essential before deciding how to handle missing data.

3.4.2. Compatible imputation and analysis models

As previously described, MI requires the specification of both an imputation model and an analysis model. It is important that the imputation model is compatible (101, 102), or, relatedly, "congenial" (103), with the analysis model. Von Hippel (102) describes this as ensuring: "any relationship in the analysis model should also be part of the imputation model". Von Hippel notes that this justifies, for example, the inclusion of the analysis outcome variable in the imputation model, which can appear counter-intuitive at first sight. If the outcome variable is *not* included, covariate values will be imputed that have no association with the outcome, which will bias any resulting analysis towards the null. More formally, compatibility requires that a model exists which has conditional distributions equal to the imputation and analysis models (101, 104). Bartlett *et al.* (101) note that the weaker property of semi-compatibility can be useful. Semi-compatibility occurs when an imputation model can be made compatible by setting one or more imputation model parameters to zero in the analysis model. This allows for the inclusion of auxiliary variables in the imputation model, that is, variables predictive of missingness and incomplete data, that are not included in the analysis model.

Hughes *et al.* (105) found that Rubin's variance estimator (88) was sensitive to incompatibility and use of incorrect imputation models. They compared Rubin's estimator with the estimator suggested by Robins and Wang (106); full mechanism bootstrapping and a robust version of Rubin's variance estimator, which incorporates a sandwich estimator for within imputation variance (107). The estimator of Robins and Wang performed well for large samples but less so for small samples (since it relies on large sample properties); a further limitation is that it relies on specification of a joint imputation model so cannot be used with FCS. Full mechanism bootstrapping also performed well but was less efficient; a limitation of this estimator is that it requires explicit specification of the MDM when data are MAR. The sandwich estimator performed least well of

the three comparator estimators, although performance was comparable in the presence of severe non-normality. However, it has the advantage of ease of implementation, without the need for a joint model or knowledge of the MDM. In a subsequent study of the use of MI with bootstrapping, Bartlett and Hughes (108) found that it is valid to use bootstrapping after MI, provided the imputation and analysis models are compatible, but may lead to incorrect coverage if they are not. If uncongeniality is suspected, bootstrapping before MI is recommended.

Vansteelandt *et al.* (109) suggest that incompatibility and model misspecification can be avoided by using IPW methods. In particular, they argue that doubly robust IPW estimators, which have the advantage over standard IPW methods of incorporating information from partially observed cases, both minimise bias introduced by model misspecification, and maximise efficiency.

An additional compatibility issue arises in FCS MI: that of compatibility between the univariate imputation models and a single joint model. Liu *et al.* (110) show that achieving this form of compatibility is sufficient for convergence to the posterior distribution of a full Bayesian model, which in turn implies appropriate coverage when using FCS MI. Hughes *et al.* (94) develop this idea further and show that, for finite samples, in addition to compatibility, a "non-informative margins" condition is required. This means that, for each imputed variable, the joint distribution can be factorised into independent conditional distributions for the imputed variable and all other variables. They demonstrate that this condition holds for the multivariate linear and some forms of multinomial distribution, but not for the general location model. However, they suggest that for any models that fail to meet the non-informative margins condition, order effects (in which results depend on the order in which variables are imputed in the FCS algorithm) can occur. These would usually be small and negligible for large samples.

3.4.3. Non-linear terms and interactions

A common cause of incompatibility is the failure to account for non-linear or interaction terms in the imputation model when these are required in the analysis model. For example, a quadratic form of a covariate may be specified in the analysis model, but a linear form is used in the imputation model. A simple approach is to 'impute then transform', that is, to apply any required transformation after imputation, *e.g.* as in the previous example, by imputing a linear form of a covariate and then using the square of that value in the analysis model. A related approach is "passive imputation" (111), in which values of the transformed variable are calculated from the (imputed or observed) untransformed variable during the imputation step. Then both untransformed and transformed versions of the variable are used as predictors in imputation models for other incomplete variables.

Von Hippel (102) argues that both methods lead to biased estimates and advocates reversing the method to 'transform then impute' (coined by others as 'just another value', JAV). In JAV, untransformed and transformed versions of the variable are imputed independently (then, as for passive imputation, both are used as predictors in imputation models for other incomplete variables). The JAV approach may yield imputed values that are inconsistent with one another, *e.g.* it does not guarantee that imputed quadratic values are the square of imputed linear values. However, Von Hippel found JAV led to unbiased estimates when using a JM approach to impute squared and two-way interaction terms in linear and probit regression models.

Morris *et al.* (112) applied JAV to the imputation of ratios. They found that imputing the ratio independently of the two covariates that represented the numerator and denominator, led to unbiased estimates with correct coverage. In contrast, passive imputation led to biased results. In linear regression analysis, Seaman *et al.* (113) found that JAV gave consistent estimation only when data

were MCAR, although bias was generally small when data were MAR. However, they found that JAV led to severe bias for logistic regression analysis and advise the use of PMM instead in this case. Tilling *et al.* (114) found JAV led to bias in linear and logistic regression analysis models including two-way interaction terms. They advise including all interactions implied by the interaction specified in the analysis model *e.g.* if the analysis is a regression of Y on X, Z and XZ, the imputation model for Z should include X, Y and XY.

Bartlett *et al.* developed another approach to compatibility by modifying the FCS method, termed "substantive model compatible" FCS (SMC-FCS) (101). This version of FCS involves specifying the imputation model as the product of two densities (for the analysis model and incomplete covariates, respectively). Since this specification of the imputation model will not usually have a standard parametric form, Bartlett *et al.* use rejection sampling to impute suitable values when the analysis is a linear regression, logistic regression or Cox PH model. In simulation studies with squared and two-way interaction terms in linear and Cox regression models, they found passive imputation led to estimates that were biased towards the null, while SMC-FCS gave unbiased results. They found that JAV (102) and a polynomial approach suggested by van Buuren for quadratic terms (115, 116) also performed well, though with some bias in certain MAR scenarios. Goldstein *et al.* (117) use a Bayesian approach to handle incomplete covariates, when these include interactions and non-linear terms, in multi-level analysis models. Compatibility between imputation and analysis models is obtained by using a Metropolis-Hastings acceptance ratio. Diaz-Ordaz *et al.* (118) also consider a multi-level MI approach for the analysis of cluster randomised trial data. The clustered nature of the data is incorporated into the imputation model using a random-cluster effect.

Morris *et al.* (104) suggest a different approach when fractional polynomials (FP) are required for the analysis model and the exact form of FP is unknown at the imputation stage. Their method involves repeatedly sampling from the observed

data using an approximate Bayesian bootstrap, determining the best form of FP for each sample, drawing from the distribution for this FP, and then passively imputing the linear form of the missing covariate. Morris *et al.* (104) apply their method to FP of dimension one, but note that applications to FP of dimension two or more could be extremely computationally intensive. Although they discuss problems with using PMM, JAV and SMC-FCS for FP models of this type, results for these methods are not compared with their suggested method. Hence, it is not clear whether it performs better in practice.

Recently, Beesley and Taylor (119) have suggested an alternative method for ensuring compatibility between the imputation and analysis models when covariate data are missing. Their approach is based on imputation "stacking", in which all imputed datasets are combined to create one large dataset. The stacked approach is shown to be valid by Robins and Wang (106). Contrary to the standard approach, Beesley and Taylor (119) suggest imputation models for the incomplete covariates that do not use the outcome as a predictor. Instead, the analysis model is incorporated via weights. Analysis estimates are then obtained by fitting a weighted version of the analysis model to the stacked dataset. Beesley and Taylor argue that their approach is valid when covariates are MAR, even when missingness depends on the outcome (although they acknowledge that a small degree of bias may occur in this case).

3.4.4. Skewed and bounded variables

Several studies have focused on imputation of skewed or bounded variables *e.g.* survival times or grouped continuous variables such as income brackets. Royston describes an implementation of FCS MI for skewed or bounded variables using Stata software (120). An assumption of the method described is that the variable, if fully observed, would follow a normal distribution. A log transformation is suggested for survival times and other skewed and/or strictly positive variables, and a truncated normal is suggested for bounded variables. Royston points out that the normality assumption can lead to incompatibility

between the imputation and analysis model which may lead to results biased towards the null. Von Hippel (121) also found that attempts to normalise skewed variables via transformation led to biased results and remarks that: "The point of imputation is not that the imputed values should *look* like observed values. The point is that the imputed variable should *act* like the observed variable when used in analysis." He advises the use of untransformed variables in the imputation model when the purpose of the analysis is to estimate associations, but the use of transformed variables when estimating percentiles. He notes the best solution would be to use a flexible imputation model that can cope with non-normal distributions rather than forcing a skewed distribution to fit a linear regression model.

Kwon and Park (122) adapted the LRD method to accommodate boundaries associated with incomplete variables by adding the residual divided by the distance between the predicted value and its boundary, which they call a "proportioned residual." They found that their method was less biased than LRD and PMM methods and was robust given skewed distributions. In contrast Rodwell et al. (97) found that restricting the range of imputed values, either during the imputation step or post-imputation, led to substantial bias. They conclude that it is best to impute on the original scale without range restrictions.

Lee and Carlin (98) expanded on the work of von Hippel and Rodwell *et al.*, comparing various transformations of non-normal data to imputation on the original scale and PMM, in linear and logistic regression analyses. They found that imputation on the original scale and PMM resulted in broadly unbiased estimates when the relationship between the completely observed outcome and the incomplete covariate was linear. When the relationship was non-linear, only PMM gave unbiased results. Type 1 PMM performed better than type 2 PMM (type 2 PMM led to under-coverage in estimation of the mean, and bias when the outcome was a binary variable). Transforming the data led to bias in all scenarios.

3.4.5. Missing outcomes

Much of the literature on general MI methods assume that outcome data are all observed (90-92, 101, 123). Little (46) asserts that if covariates are complete and values of the outcome variable are MAR, then cases with missing outcome contribute no information to the analysis model. In a method described as "multiple imputation, then deletion" (MITD), Von Hippel (124) expands on another comment by Little, such that if covariate values are to be imputed, then cases with missing outcome "can provide a minor amount of information" by improving imputation of missing covariates in cases with observed outcomes. Von Hippel argues that once cases with missing outcomes have fulfilled their useful purpose, *i.e.* by improving the imputation model, they should be removed from the analysis model. This approach assumes that there are no auxiliary variables which would improve the prediction of the missing outcome values. Sullivan *et al.* (125) used a JM MI approach to assess MITD (124) in the presence of auxiliary variables. They found that MITD resulted in biased estimates when missingness in the outcome variable was associated with an auxiliary variable, with bias increasing with the proportion of missing data and the strength of the association between the auxiliary variable and the outcome (125). In this situation, Sullivan *et al.* suggest the best approach is to include imputed outcome data in the analysis model. They argue that to exclude these data would exclude the information provided by the auxiliary variable, and hence would violate the MAR assumption.

3.5. Use of multiple imputation in survival analysis

In survival analysis, it is well-established that both the censoring indicator and a representation of the survivor function should be included in the imputation model when imputing covariates (91, 101, 123). In a study of Cox proportional hazards models with right-censored survival times, White and Royston (123)

found that the inclusion of the cumulative baseline hazard in linear or logistic regression imputation models for a single incomplete covariate led to smaller bias than if the survival time or log-survival time was included. They found there was some bias towards the null in scenarios where there was a strong association between the incomplete covariate and the outcome of interest. The Nelson-Aalen estimator of the cumulative hazard was found to be an appropriate approximation of the cumulative baseline hazard.

Bartlett *et al.* (101) found further improvement through application of their SMC-FCS method, described previously. They found SMC-FCS led to broadly unbiased estimates with correct coverage while White and Royston's method led to some bias towards the null. Keogh and Morris (126) extended the approaches of both White and Royston and Bartlett *et al.* to accommodate time-dependent covariates. Their results agreed with those of Bartlett *et al.* (101). Bartlett and Taylor (127) also extended the SMC-FCS approach to competing risks by considering the cause-specific Cox model, but did not consider direct modelling of the cumulative incidence function *e.g.* by using the Fine and Gray model (62).

3.5.1. Missing outcomes in survival analysis

As discussed in Chapter 2, Section 2.2, it is routinely the case that some study subjects will not have experienced the event(s) of interest by the end of the observation period. The event times for these subjects are said to be censored (specifically, right-censored). Standard survival methods can accommodate right-censored event times, provided the censoring mechanism is independent of the survival time itself (possibly conditional on model covariates) (50-52). Typically, imputation of unobserved event times is performed to overcome two main departures from standard survival model assumptions: (i) dependent censoring and (ii) interval-censoring. These types of censoring, and methods for handling them, are described in subsequent sections.

(i) Dependent censoring

Dependent, or informative, censoring means that the censoring mechanism is related to the survival time itself. It occurs, for example, when the sickest patients drop out of a clinical trial and are lost to follow-up. Jackson *et al.* (128) used MI to explore the effect of dependent censoring in a Cox regression analysis. They imputed the censored failure times (by treating them as data MAR) by drawing directly from the survivor function for a Cox model fitted to all observed failure times. They found their method performed well with small bias and adequate coverage. Since the baseline hazard function in a Cox model does not have a parametric form, they used bootstrap sampling to create multiple versions of the baseline hazard. They note that using a parametric specification for the baseline hazard would remove the need to use bootstrapping and would allow a conventional MI approach (in which random draws are made from the posterior distribution of the baseline hazard function parameters).

Hsu *et al.* (129) used a method analogous to PMM to handle dependent censoring. They included auxiliary variables associated with censoring in the imputation scheme to make the MAR assumption plausible. Here the predictive distance used in the PMM procedure was based on a composite risk-score rather than a linear regression model. They performed a simulation study comparing this method to an IPW method (130). They found that MI was more efficient and IPW was less biased (as is generally the conclusion when MI and IPW methods are compared). Xiang *et al.* (131) obtained unbiased results using a similar approach, but with a restricted mean model to define the set of donors. As in the LRD method, the imputed value is the sum of the residual from the restricted mean model for the donor and the predicted value for the missing failure time.

(ii) Interval censoring

As described in Chapter 1, Section 1.3.2, interval-censoring means that the event time is not observed exactly but is known to lie within a particular time interval.

If the censored intervals are relatively narrow, Sun (47) suggests substituting a single point from the interval (usually the mid-point if there is no prior information about the part of the interval where the event is more likely to occur).

As a more rigorous alternative, MI can be used to handle interval-censoring (by treating the interval-censored times as missing data). In several studies, described below, the imputation model is based on the survivor function. Hsu *et al.* (132) extended their PMM-like method for imputation of right-censored observations (129), to interval-censored data, by conditioning on the censored interval boundaries. Pan (133) compared two MI methods involving iterative draws from the survivor function for a Cox model, conditional on the censored interval boundaries: the Poor Man's Data Augmentation (PMDA) and Asymptotic Normal Data Augmentation (ANDA) (134). They recommend using ANDA because they found PMDA underestimated SE in some scenarios. Chen and Sun (135) adapted Pan's method for an additive hazards analysis model. Delord and Genin (136) adapted Pan's method for a competing risks analysis, using both the non-parametric estimator of the cumulative incidence function and the Fine and Gray sub-distribution hazards regression model (62).

IPW methods have also been used to address interval-censoring. Kim *et al.* (137) proposed an additive-multiplicative hazards regression model (combining proportional hazards and additive hazards models) for semi-competing risks data, accommodating interval-censoring by the application of IPW over the censoring interval. Hyun *et al.* (138) used IPW techniques to allow for interval-censored times in prevalence-incidence mixture models, also accounting for competing risks by modelling the sub-distribution hazard function.

3.6. Full maximum likelihood methods for handling interval-censored outcomes

As an alternative to MI, FML methods for handling interval-censored outcomes have been developed. Early methods were based on the "simplified" likelihood function, which assumes non-informative censoring (139). The simplified likelihood function takes the following form: for a cumulative distribution function F(t), the contribution to the simplified likelihood function by subject i, whose event time is known to lie in the interval (L_i, R_i], is proportional to {F(R_i) – F(L_i)}. Hudgens *et al.* (140) used this approach to derive the NPMLE of the cumulative incidence function for competing risks data.

Parametric models for interval-censored competing risks data have also been considered. Such models can be particularly useful for small sample sizes and wide censoring intervals (47), and can be more efficient (141) than a non- or semi-parametric approach. Hudgens *et al.* (141) used a Gompertz distribution for the cumulative incidence function for each event type in a competing risks framework. They found that their estimators had smaller bias and variance than the equivalent NPMLE. Mitra *et al.* (142) extended the approach of Hudgens *et al.*, allowing for missing event types, considering both non-proportional and proportional sub-distribution hazard models.

An alternative FML approach uses the penalised likelihood and/or smoothing functions to increase computational efficiency, and to avoid the need for strong parametric assumptions. Li (143) developed a penalised likelihood approach for a cause-specific PH model. They also developed a spline-based estimator for the log baseline cumulative sub-distribution hazard (144) in a Fine and Gray (62) regression model. Bakoyannis *et al.* (64) extended Li's method to estimate the cumulative incidence function in a competing risks framework, allowing different models for each event type. They argue that their approach yields more

semi-parametrically efficient regression model estimates than the approach of Delord and Genin (136).

FML methods for handling interval-censored outcomes in a MSM analysis have also been developed. For example, Frydman and Szarek (145) proposed an NPMLE of the distribution of time to first event in an illness-death Markov model, with interval-censored intermediate event times. Foucher *et al.* (146) used a similar approach for semi-Markov models with parametric (Weibull) transition intensity models. Beesley and Taylor (147) used an expectation-maximisation (EM) algorithm to fit a multistate cure model, incorporating an imputation-type method to handle missing values of cure status and time of disease recurrence. They note that their imputation approach is "improper" because a single estimate of the predictive distribution parameter is used throughout the imputation step (rather than drawing a new parameter value in each iteration as in MI).

Time-homogeneous (*i.e.* assuming that the transition intensity is constant over time) Markov models have also been used to handle interval-censored event times in MSMs (148-151). Jackson (148) notes that this implies that the censoring intervals are ignorable, which would be valid if, for example, the censoring intervals were determined by clinic visits at fixed time-points. Healy and Degruttola (149) considered approaches to fitting time-homogeneous Markov models when the time origin and the states were known with error. Jackson extended their approach to time-inhomogeneous models, by assuming the transition intensity was constant over each censoring interval, referred to as "piecewise" constant hazards (148).

Assuming piecewise constant hazards, Joly e*t al.* (150) and Machado and van den Hout (151) used penalised maximum likelihood estimation with spline functions for the baseline intensities. Machado *et al.* (152) describe an improvement of the

smoothing parameter estimation in the method of Machado and van den Hout. They compared their new method with that of Joly *et al.* and Jackson, finding that the latter performed less well than the other methods. They note that this is probably because Jackson's method only allows a limited choice of models to be fitted, whereas the other methods are more flexible. They found that none of the methods performed well when censoring intervals were wide, relative to the change in the hazards.

3.7. Discussion

This chapter has provided an overview of missing data methods. The focus of this review has been on methods most relevant to my research, that is, methods for handling missing covariate and outcome data in competing risks and MSM analyses.

In the NHS CBB dataset, the missingness of event times depends on the type of event experienced (as well as patient, donor and transplant characteristics) (see Chapter 4, Section 4.4). In this case, CCA will be biased because missingness of event times depends on the analysis outcome.

In addition to CCA, available methods can be broadly categorised as (i) applying weighting methods, (ii) applying MI strategies or (iii) taking a FML approach. Weighting methods are most useful when a single variable is incomplete or when data are monotone missing, and the sample size is large (82). In the NHS CBB dataset, both event times and covariate data are missing, and only 432 transplants are available for analysis (see Chapter 4, Sections 4.3-4.4). Therefore, weighting methods are not considered in my analysis.

In the NHS CBB dataset, event times for the events of interest are frequently not reported, although the times of competing events generally are (see Chapter 4, Section 4.4). The interval boundaries for event times are wide and generally the

same for all patients: for acute GvHD, missing event times are assumed to occur in the acute period (between day 0 and day 100, or the time of the patient's death, if this occurs before day 100) and for chronic GvHD, missing event times are assumed to occur in the chronic period (beyond day 100). The upper boundary for missing event times for chronic GvHD and relapse will be the time of the patient's death or last observation time. Hence, the missing event times in the NHS CBB dataset deviate from the assumptions of the FML methods developed so far, in two ways: (i) event times are a mixture of exact and interval-censored times and (ii) interval boundaries are wide relative to the observed event times.

Therefore, in my research, I focus on MI strategies for handling missing event times, using the FCS approach. The appeal of FCS MI is its flexibility: the FCS method can accommodate a mixture of continuous, binary, and categorical incomplete variables, non-linear associations, and any constraints imposed on incomplete variables. A mixture of observed and missing data can be accommodated, and additional data that are predictive of the missing times, but not required for the analysis model, can be used during the imputation step to inform the imputed times. After imputation, any desired complete data method may be used. To assess whether MI is an improvement over FML, I compare MI strategies with a suitable FML method in a competing risks analysis (see Chapter 5, Section 5.2.7).

FCS MI is valid if the analysis and imputation models are correctly specified, and each variable being imputed is not MNAR. Based on research to date, it is clear how to impute missing covariate data when these are not MNAR: use FCS MI to impute variables using an appropriate parametric model by, for example, imputing continuous variables using linear regression, binary variables using logistic regression and categorical variables using multinomial regression.

Optimal methods for imputing missing event times are less clear. It is not evident that the general recommendations for imputation of missing outcomes

(described in Section 3.4.5) apply to imputation of missing event times in a survival analysis model, when both non-linear associations and censoring need to be accounted for. In this thesis, I will address key questions about imputation of event times. These questions are described in more detail below.

The first question is whether it is necessary to constrain the imputed event times to lie within the censoring interval boundaries. Additionally, event times tend to be right-skewed: it is not clear to what extent skewness should be accounted for in the imputation model. Another challenge is that of compatibility between the imputation and analysis models. It is not clear whether it is important that the imputation scheme incorporates the cumulative incidence function in a competing risks framework, or the order of the event times in a MSM. In such models, estimands of interest depend on all the cause-specific hazards (see Chapter 2, Section 2.2.2), so the challenge of compatibility between the imputation and analysis models increases dramatically. The method of Delord and Genin (136) is currently the only compatible MI method available for non-parametric estimation of the cumulative incidence function in the presence of competing risks. Finally, to date, the use of MI methods for handling interval-censored event times in a MSM analysis has not been assessed. Therefore, in this thesis, I will compare the performance of MI methods in a MSM analysis.

CHAPTER 4. OVERVIEW OF THE NHS CORD BLOOD BANK DATASET

4.1. Introduction

This chapter provides an overview of patient, donor and transplant baseline characteristics and the types of events experienced by patients in the NHS CBB dataset, including the percentage of missing data for each variable. Associations between missingness of event times and all other variables are explored, and possible missingness mechanisms for event times are described.

4.2. Description of the NHS Cord Blood Bank dataset

Between 1996 and 2015, 432 first (for the patient) transplants were performed using CB donated to the NHS CBB. An additional 50 transplants were excluded. These were of different types from the majority of transplants, but were only present in small numbers. The reasons for exclusion were as follows:

- The transplant was the second or later transplant for a patient, where the first transplant was a CB transplant (N = 12)
- The transplant was the second or later transplant for a patient, where the first transplant was not a CB transplant (N = 12)
- The donor was related to the patient (N = 12)
- The patient received a non-standard source of HSCs *e.g.* CB plus PB (N = 14)

4.3. Baseline patient, donor and transplant characteristics

Baseline patient, donor and transplant characteristics of the 432 transplants in the NHS CBB dataset are summarised in Table 4.1 and briefly described overleaf.

Table 4.1. Baseline patient, donor and transplant characteristics of the NHS CBB dataset.

Characteristic	**Level**	**N (%)**
Year of transplant	1996-2000	30 (7)
	2001-2005	66 (15)
	2006-2010	172 (40)
	2011-2015	164 (38)
Country of transplant	France	52 (12)
	UK	138 (32)
	Other European country	83 (19)
	USA	117 (27)
	Other non-European country	41 (9)
	Not reported	1(<1)
Patient age at transplant (years)	<16	195 (45)
	16-39	84 (19)
	40-59	103 (24)
	60+	50 (12)
Disease type at transplant	Acute leukaemia	217 (50)
	Other blood cancer[1]	126 (29)
	Non-malignant disorder[2]	89 (21)
Disease status at transplant	Partial/complete remission	158 (37)
	Relapse	23 (5)
	Other[3]	123 (28)
	Not reported	128 (30)
Conditioning regimen	Intensive	209 (48)
	Reduced intensity	206 (48)
	Not reported	17 (4)
Number of CB units received	Single cord	232 (54)
	Double cord	200 (46)
GvHD prophylaxis	Yes	10 (2)
	No	0 (0)
	Not reported	422 (98)

Table 4.1 contd.

Characteristic	Level	N (%)
Donor-recipient CMV match	Negative to positive	114 (26)
	Positive to negative	81 (19)
	Negative to negative	70 (16)
	Based on one donor[4]	40 (9)
	Positive to positive	40 (9)
	Based on one donor[4]	35 (8)
	Not reported	52 (12)
Donor-recipient sex match	Male to female	89 (21)
	Female to male	136 (31)
	Male to male	68 (16)
	Based on one donor[4]	55 (13)
	Female to female	51 (12)
	Based on one donor[4]	30 (7)
	Not reported	3 (1)
Number of donor-recipient HLA mismatches[5]	Well-matched: 0 or 1	144 (33)
	Based on one donor[4]	75 (17)
	Poorly-matched: 2 or more	161 (37)
	Not reported	52 (12)
TNC dose at infusion ($\times 10^7$/kg)	Low: <3.0	54 (13)
	Based on one donor[4]	38 (9)
	Medium: 3.0-5.0	61 (14)
	Based on one donor[4]	69 (16)
	High: > 5.0	74 (17)
	Based on one donor[4]	31 (7)
	Not reported	105 (24)

[1] Other blood cancer includes lymphoproliferative and plasma cell disorders, myelodysplastic syndromes and myeloproliferative disorders.
[2] Non-malignant disorder includes histiocytic disorder, solid tumour, bone marrow failure syndrome, haemoglobinopathy, primary immune deficiency and inborn error of metabolism.
[3] Other disease status includes acute, chronic and accelerated phase, refractory disease, transformed to acute leukaemia, blastic crisis, MDS, MDP and non-malignant disorders.
[4] Based on data for one donor of a double cord transplant, where data for the second donor are not available. For TNC dose, the dose reported here is double the dose reported for one donor.
[5] HLA A and B loci at antigenic level and DR-B1 at allelic level

Most transplants were performed in the last 10 years of the study period (336, 78% during 2006-2015 vs. 96, 22% during 1996-2005). This reflects increasing adoption of CB transplant as a standard treatment (14), as well as increasing numbers of transplant centres providing information to the Eurocord Registry (36). Approximately half (45%) of all transplant recipients were paediatric

patients (aged < 16 years). The mean age of paediatric patients was 5 years and of adult patients was 45 years.

HSC transplantation is used to treat a variety of disease types (14) and this is reflected in the NHS CBB cohort. Most patients (343, 79%) were treated for a blood cancer, with acute leukaemia the most common disease type in both adults and paediatric patients (133, 56% vs. 84, 43%). Non-malignant disorders, such as primary immune deficiency and inborn error of metabolism, are usually diagnosed and treated in childhood (153) and accounted for nearly half (80, 41%) of all paediatric transplants in the NHS CBB cohort. Only 9 (4%) adult patients were treated for a non-malignant disorder. Where reported, most patients (158 of 215 patients with status reported, 73%) were in partial or complete disease remission at transplant. Where reported, intensive and reduced intensity conditioning were each used in about half the transplants in the study cohort. GvHD prophylaxis was only reported for 10 (2%) patients, and all these patients received GvHD prophylaxis. Therefore, using MI to impute the missing values of GvHD prophylaxis status is not appropriate because the observed data are not good predictors of the missing values (assuming that not all patients received GvHD prophylaxis) (85). There was no additional information available to determine whether GvHD prophylaxis was or was not received by the remaining 422 patients. Therefore, this variable was treated as completely unobserved, and was excluded from subsequent analyses.

Of all 432 transplants, 200 (46%) were double cord transplants. As described in Chapter 1, Section 1.2.7, a double cord transplant is where the patient receives CB from two different donors during their transplant. Double cord transplants are usually given when a single cord would provide insufficient HSCs for the patient's body weight (14). Most adult patients in the NHS CBB cohort received a double cord transplant (188, 79% of adults vs. 12, 6% of paediatric patients). In most double cord transplants (180 out of 200, 90%), one of the two donors did not donate via the NHS CBB, and hence, none of this donor's data were available.

Hereafter, data based on one of the two donors of a double cord transplants will be referred to as 'partially observed' donor data, to distinguish it from completely missing data (*i.e.* values for which no information was available). Strategies for classifying characteristics that depend on both donors' data in double cord transplants, namely, sex and CMV match between donors and recipient, number of HLA mismatches between donors and recipient, and TNC dose at infusion, are described below.

4.3.1. Strategies for handling double cord transplant donor data

Donor-recipient sex and CMV status match

In double cord transplants, a mismatch in the donor-recipient sex or CMV status was defined as at least one donor of the opposite sex or CMV status to the patient. No distinction was made between a mismatch involving one or both donors and it was assumed that the effect of a mismatch would be the same regardless of the number of mismatched donors. Hence, if the known donor was of the opposite sex or CMV status to the recipient, a mismatch could be identified from partially observed donor data. However, in many transplants (85, 20% for sex match and 75, 17% for CMV match), a sex or CMV mismatch could not be determined from partially observed donor data because the known donor was the same sex or CMV status as the patient. For these cases, the match based on the known donor is reported; such cases are identified in separate rows in Table 4.1.

Number of donor-recipient HLA mismatches

For transplants with partially observed donor data, in which the patient was a poor HLA match for the known donor (2 or more HLA mismatches), knowledge of the second donor's HLA types could only increase the number of mismatches. Hence, poorly-matched transplants could be identified from partially observed donor data. However, in cases in which the known donor was well-matched (0

or 1 mismatches), the overall number of HLA mismatches could not be fully determined. For these cases, the number of mismatches based on the known donor is reported; such cases are identified in a separate row in Table 4.1.

TNC dose at infusion

For transplants with partially observed donor data, in which the TNC dose at infusion for the known donor was reported (N=138), the mean dose was 1.99×10^7/kg. This was less than half the TNC dose reported for single cord transplants or double cord transplants where both TNC doses were reported (mean TNC dose: 5.98×10^7/kg, N=179 for single cord transplants and 4.46×10^7/kg, N=10 for double cord transplants). For the 10 double cord transplants with both doses reported, the mean ratio of the two TNC doses was 1.2. Therefore, for transplants with partially observed donor data, the reported TNC dose was doubled to give an estimate of the total TNC dose. These cases are also identified in separate rows in Table 4.1.

4.3.2. Missing patient, donor and transplant data

Some baseline variables had completely missing data (*i.e.* not even partial donor information was available, see Table 4.1). These variables were (with percentage missing) disease status (30%), conditioning regimen (4%), sex match (1%), CMV match (12%), number of HLA mismatches (12%), TNC dose at infusion (24%), and country of transplant (<1%). Only 197 transplants had complete baseline data; this reduced to 132 transplants if partially observed donor data were treated as missing data.

My research is focused on missing event times, not missing covariate data; in my simulation studies, I assume that all covariate data are observed. Therefore, in my exploratory analyses of missing event times in the NHS CBB dataset, I use simple methods for handling missing baseline data. I treat values inferred from partially observed donor data as the true values. For each variable with completely missing values (as defined above), missing values were included in a

separate "missing" category. My approach has the advantage of being straight-forward to apply and uses the available data for all 432 transplants. However, it has been shown that use of "missing" categories can lead to biased analysis results (154). Therefore, in final analyses of the NHS CBB dataset (see Chapter 7, Section 7.4.2-7.4.3), missing baseline data, including values based on partially observed donor data, were imputed using FCS MI.

4.4. Summary of post-transplant events and their missingness

The median post-transplant follow-up of patients in the NHS CBB dataset was 3 years (Kaplan-Meier estimate, censoring follow-up time at death). Only seven patients had less than 100 days follow-up (with a range of 61-99 days follow-up). At least one post-transplant event was reported for each patient.

Table 4.2 shows the number (percentage) of patients in the NHS CBB dataset who experienced each type of post-transplant event. This table also shows the number (percentage) of events with a missing event time.

Table 4.2. Events experienced, and number (percentage) missing an exact event time.

Event	**N (%)**	**N (%) missing exact time of onset**
Acute GvHD	241 (56)	57 (24)
Chronic GvHD	82 (19)	29 (35)
Relapse	89 (21)	22 (25)
Myeloid engraftment	373 (86)	3 (1)
Graft failure prior to engraftment	37 (9)	4 (11)
Death	196 (45)	0 (0)

Cases of acute GvHD, chronic GvHD and relapse were reported without an exact event time in at least 24% of cases. Of the 22 cases of relapse without an exact event time, relapse was inferred from the cause of death in 20 cases. In the other two cases, relapse was reported without an associated event time, but the patient was alive at last follow-up. All times of death were reported.

When the number and type of missing event time per patient were examined (Tables 4.3a and b), it was found that, of all 432 patients, 66 (15%) patients were missing one event time, 23 (5%) were missing two event times (of which 18 were missing times of both acute and chronic GvHD), and one patient was missing three event times (times of acute GvHD, chronic GvHD and relapse).

Table 4.3a. Events missing exact event times, for patients with only one missing event time.

Event type	**Acute GvHD**	**Chronic GvHD**	**Relapse**	**Graft failure**	**Myeloid engraftment**	**Death**
Number of patients	34	10	16	3	3	0

Table 4.3b. Events missing exact event times, for patients with two missing event times.

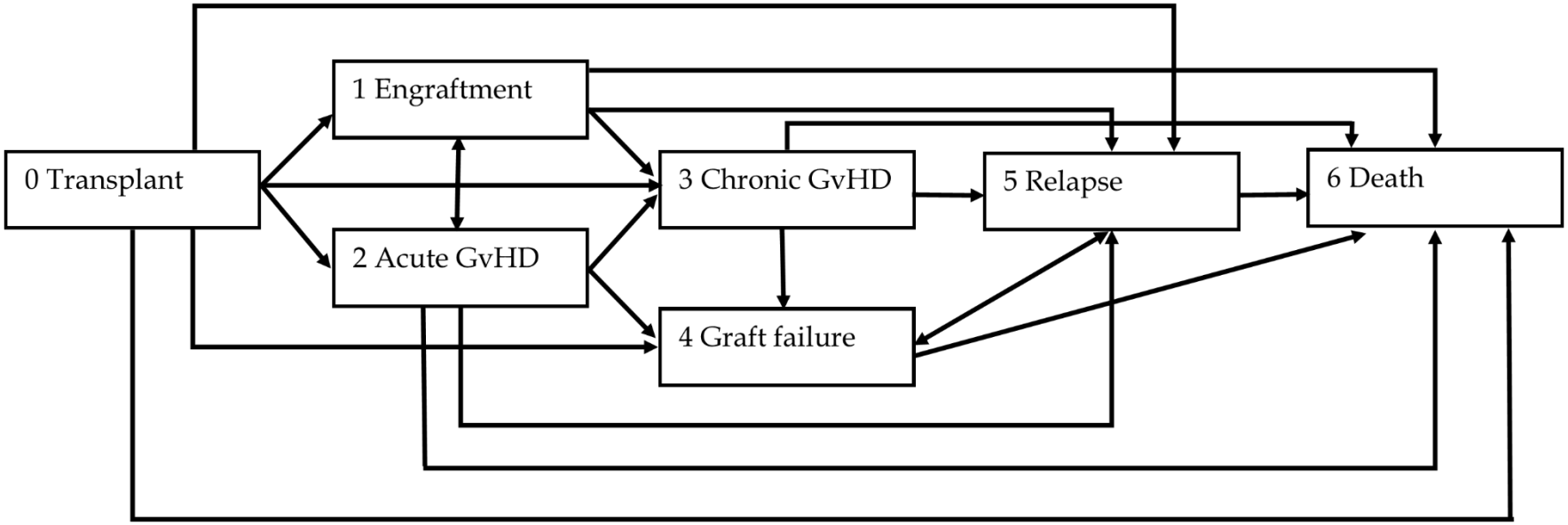
First missing event time	**Second missing event time**					
	Acute GvHD	Chronic GvHD	Relapse	Graft failure	Myeloid engraftment	Death
Acute GvHD		18	4	0	0	0
Chronic GvHD			0	0	0	0
Relapse				1	0	0
Graft failure					0	0
My. Engraftment						0
Death						

Numbers represent the number of patients with each combination of missing events, regardless of event time order.

4.4.1. Combinations of events experienced

All combinations of events experienced by patients in the NHS CBB dataset are shown as a MSM in Figure 4.1 overleaf. Note that, for patients with some missing event times (N=90), the observed event orders are consistent with the event orders of patients with completely observed event times (N=342).

Figure 4.1. Multi-state model for post-transplant events experienced by patients in the NHS CBB dataset.



There are 22 separate transitions between states in the MSM, some of which were experienced only by a small number of patients *e.g.* from graft failure to relapse or to chronic GvHD from states other than acute GvHD. Broadly, event combinations experienced by patients in the NHS CBB dataset can be categorised into five sub-groups, depending on whether acute GvHD, chronic GvHD, and/or relapse were experienced (Table 4.4). The full list of events experienced per patient is included in the Appendices (Tables A.1a and b).

Table 4.4. Combinations of events experienced by patients.

Events experienced (regardless of event order)	**N**
Acute GvHD & chronic GvHD*	68
Acute GvHD without chronic GvHD*	173
Chronic GvHD without acute GvHD*	14
Relapse without GvHD	44
Neither GvHD nor relapse	133
ALL PATIENTS	**432**

In each sub-group, patients may also experience myeloid engraftment, graft failure and/or death.
*Patients in these sub-groups may also experience relapse.

For all but three of the 342 patients with completely observed event times, relapse was followed by only graft failure or death. There were three clinically anomalous cases, in which relapse was reported to have occurred between transplant and transplant-related events:

(i) For one patient, relapse was reported to have occurred shortly after (at day 30 post-transplant) myeloid engraftment and acute GvHD; the patient was subsequently reported to have experienced chronic GvHD and was still alive more than five years post-transplant. In this case, early relapse did not appear clinically plausible given the other events experienced. In subsequent analyses, it was assumed that relapse had not occurred for this patient.

(ii) For two patients, relapse was reported to have occurred at the same time as myeloid engraftment and acute GvHD, or shortly before myeloid engraftment. In both cases, the patient died within 100 days of transplant, so early relapse appeared clinically plausible. Therefore, in both cases, a small amount of time (0.1 days) was added to the time of relapse to ensure

consistency with the event order observed for all other patients.

4.4.2. Missingness of event times by event combination

To explore whether missingness of event times depended on the combination of events experienced, patients in the NHS CBB dataset were categorised into the five sub-groups described in Table 4.4 above. This meant that patients with different types of missing event times tended to be in separate groups. A summary of missing event times for these patient sub-groups is shown in Table 4.5. Note that, in survival analysis, analysing patients according to events experienced after time zero can lead to bias (155). For example, patients need to be alive and in follow-up for long enough after transplant (at least 100 days) to experience chronic GvHD. Hence, experiencing chronic GvHD could appear to be protective if patients are classified as such at baseline because, by definition, people with chronic GvHD cannot die within the first 100 days. Appropriate survival analysis approaches, in this case, are to treat chronic GvHD as a time-varying covariate in a model for time to death or to use a MSM approach (see for example, the approach by Glimelius *et al.* (156)). However, in exploration of missingness mechanisms, and application of MI (where the purpose is prediction of missing values), it is valid to use all observed data, including values measured at a later time-point (157).

Table 4.5. Summary of missing event times, by patient sub-group.

Events experienced	**N**	**No. of patients (no. of missing times for each event type)**					
		AGvHD	**CGvHD**	**Relapse**	**Myeloid engraft-ment**	**Graft failure**	**Death**
AGvHD & CGvHD*	68	68 (22)	68 (24)	11 (1)	68 (0)	0 (0)	17 (0)
AGvHD w/o CGvHD*	173	173 (35)	n/a	33 (7)	165 (2)	7 (1)	79 (0)
CGvHD w/o AGvHD*	14	n/a	14 (5)	1 (0)	13 (0)	1 (0)	3 (0)
Relapse w/o GvHD	44	n/a	n/a	44 (14)	31 (0)	10 (1)	37 (0)
Neither GvHD nor relapse	133	n/a	n/a	n/a	96 (1)	19 (2)	60 (0)
ALL PATIENTS	**432**	**241**	**82**	**89**	**373**	**37**	**196**

AGvHD = acute GvHD; CGvHD = chronic GvHD; w/o = without.
In each sub-group, patients may also experience myeloid engraftment, graft failure and/or death.
*Patients in these sub-groups may also experience relapse.

Table 4.5 shows that most missing event times occur in three sub-groups of patients, namely:

(a) Patients who experience acute and chronic GvHD. Of the 68 patients, 8 (12%) were missing the time of either acute or chronic GvHD, and 19 (28%) were missing both GvHD times.

(b) Patients who experience acute without chronic GvHD. Of the 173 patients, 35 (20%) were missing the time of acute GvHD.

(c) Patients who experience relapse without GvHD. Of the 44 patients, 14 (32%) were missing time of relapse.

Numbers of patients missing both GvHD and relapse times were very small: one case among patients who experienced both acute and chronic GvHD; seven cases among patients who experienced only acute GvHD; no cases among patients who experienced only chronic GvHD.

4.5. Associations of individual-level characteristics with missingness of event times

4.5.1. Methods for describing associations with missingness of event times

The association between individual-level characteristics and missingness of event times was investigated, to explore whether an assumption of event times MCAR or MAR (conditional on the observed data) was plausible (and hence, whether MI was a suitable method for handling missing event times). Separate analyses were performed for the three sub-groups of patients with the most missing event times, described above, namely: patients who experienced both acute and chronic GvHD, patients who experienced acute without chronic GvHD, and patients who experienced relapse without GvHD.

For patients who experienced GvHD, only associations with missingness of GvHD times, not relapse times, were investigated. This was due to the small numbers of missing relapse times in these patient sub-groups. Again, due to small numbers of missing event times, the three sub-groups were not split into further sub-categories of event combinations in this analysis, and the analysis was not performed for any other patient sub-groups.

All baseline variables were included in this analysis. The post-transplant variables considered were grade of acute GvHD (in analyses of patients who experienced acute GvHD with and without chronic GvHD); grade of chronic GvHD (in the analysis of patients who experienced acute and chronic GvHD); incidence of myeloid engraftment within 100 days post-transplant (with competing events of graft failure or death prior to engraftment) and overall survival within two years post-transplant.

Statistical methods

Firstly, individual-level characteristics were summarised as number, percentage (or median, inter-quartile range (IQR) for engraftment and overall survival), for cases with and without exact event times. Distributions of these characteristics were compared univariably (for cases with and without exact event times) using Fisher's exact test (158) (baseline variables), Gray's test (61) (incidence of myeloid engraftment) and the log-rank test (54) (overall survival). The incidence of graft failure and death prior to engraftment was not compared for those with observed and missing event times due to the small numbers of these event types.

In the univariable comparisons, for each baseline variable in turn, all values were included in the summary but missing cases for that variable were excluded when calculating the test statistic. This approach allowed identification of variables whose observed data was associated with missingness of event times. A small number of missing event times for myeloid engraftment, or the competing event of graft failure prior to engraftment, were excluded from analysis (there were no

missing death times). The cumulative incidence of myeloid engraftment, graft failure and death prior to engraftment, and overall survival function were also presented graphically, to illustrate any differences in the distribution of these events for patients with and without exact event times for GvHD and relapse.

Additionally, to inform my later simulation study designs (see Chapters 5 and 6), multivariable logistic regression was performed for each patient sub-group (with a binary indicator of missingness of the event time as the outcome). Inclusion of variables in the regression models was decided using forward selection based on the likelihood ratio test (159), selecting from the set of variables for which there was some evidence of an association with missingness of event times in the univariable comparisons (*i.e.* those with a p-value < 0.1).

4.5.2. Results: univariable associations with missingness of event times

Table 4.6 summarises patient characteristics, comparing those with and without event times, for the following sub-groups:
(a) Comparing those with observed times of both acute and chronic GvHD, and those missing either or both times (among patients who experienced both events)
(b) Comparing those with an observed time of acute GvHD, and those missing a time of acute GvHD (among patients experiencing acute without chronic GvHD)
(c) Comparing those with an observed time of relapse, and those missing a time of relapse (among patients experiencing relapse without GvHD).

In the descriptions below, these patient sub-groups are referred to as sub-groups (a) to (c), respectively. For brevity, only variables for which there was some evidence of an association with missingness of event times (*i.e.* with a p-value < 0.1) are shown in Table 4.6.

Table 4.6. Variables associated with missingness of event times.

Factor	Level	Acute and chronic GvHD			Acute without chronic GvHD			Relapse without GvHD		
		Missing GvHD time(s) (N=27)	Both GvHD times reported (N=41)	P	Missing acute GvHD time (N=35)	Acute GvHD time reported (N=138)	P	Missing relapse time (N=14)	Relapse time reported (N=30)	P
Country of transplant	France	0 (0)	10 (24)	<0.001	0 (0)	24 (17)	<0.001	0 (0)	3 (10)	0.001
	UK	3 (11)	14 (34)		12 (34)	53 (38)		1 (7)	12 (40)	
	Other Eur. country	4 (5)	11 (27)		3 (9)	31 (23)		1 (7)	8 (27)	
	USA	17 (63)	2 (5)		17 (49)	15 (11)		11 (79)	4 (13)	
	Other non-Eur. country	3 (11)	4 (10)		3 (9)	15 (11)		1 (7)	3 (10)	
Year of transplant	1996-2000	0 (0)	4 (10)	0.239	0 (0)	12 (9)	<0.001	1 (7)	2 (7)	0.102
	2001-2005	3 (11)	4 (10)		0 (0)	31 (23)		0 (0)	6 (20)	
	2006-2010	13 (48)	23 (56)		14 (40)	43 (31)		4 (29)	13 (43)	
	2011-2015	11 (41)	10 (24)		21 (60)	52 (38)		9 (64)	9 (30)	
Number of CB units received	Single cord	12 (44)	19 (46)	>0.999	14 (40)	85 (62)	0.034	3 (21)	17 (57)	0.050
	Double cord	15 (56)	22 (54)		21 (60)	53 (38)		11 (79)	13 (43)	
Donor-recipient CMV match	Negative to positive	10 (37)	4 (10)	0.014	10 (29)	42 (30)	0.587	1 (7)	10 (33)	0.244
	Positive to negative	3 (11)	15 (37)		9 (26)	25 (18)		3 (21)	4 (13)	
	Negative to negative	5 (19)	13 (32)		8 (23)	38 (28)		2 (14)	6 (20)	
	Positive to positive	7 (26)	6 (15)		2 (6)	17 (12)		4 (29)	7 (23)	
	Not reported	2 (7)	3 (7)		6 (17)	16 (12)		4 (29)	3 (10)	
Myeloid engraftment	*Median, IQR (days)*	21 (17-30)	21 (16-27)	0.910	17 (11-27)	24 (18-31)	<0.001	30 (23-51)	26 (15-na[1])	0.998

[1] Upper quartile not reached

Values are N, column % unless indicated otherwise.

Associations with baseline characteristics

The distribution of country of transplant ($p \leq 0.001$ in each case) was different for patients with missing vs. observed event times for all sub-groups (a) to (c). Most missing event times occurred for transplants in the USA, accounting for 17/27 (63%) cases with missing GvHD times in sub-group (a), 17/35 (49%) cases with missing acute GvHD times in sub-group (b), and 11/14 (79%) cases with missing relapse times in sub-group (c).

The distribution of year of transplant was different for patients with missing vs. observed times of acute GvHD in sub-group (b) ($p < 0.001$). In the early years of the study period (1996-2005), there were no missing acute GvHD times. Similarly, there were few missing event times in 1996-2005 in sub-groups (a) and (c). However, there were smaller differences in the distribution of year of transplant for patients with missing vs. observed event times in these sub-groups ($p > 0.10$ in each case).

In sub-groups (b) and (c), patients with missing event times were more likely to receive double cord transplants than those with observed times (60% and 79% received double cord transplants when event times were missing; 38% and 43% received double cord transplants when event times were observed; p=0.03 and 0.05, for sub-groups (b) and (c), respectively). In sub-group (a), there was little difference in the proportion of double cord transplants (56% vs. 54% double cord transplants, respectively, for those with missing and observed event times; $p > 0.9$).

In sub-group (a), there were differences in the distribution of donor-recipient CMV match for those with missing and observed times GvHD times (63% vs. 25% CMV+ patients, respectively, for those with missing and observed event times; p=0.01). There were smaller differences in the distribution of donor-recipient CMV match for patients with missing and observed event times in the

other two patient sub-groups (for sub-groups (b) and (c), respectively, 35% and 36% of patients were CMV+ when event times were missing; 42% and 56% patients were CMV+ when event times were observed; p=0.6 and 0.2).

Associations with post-transplant events

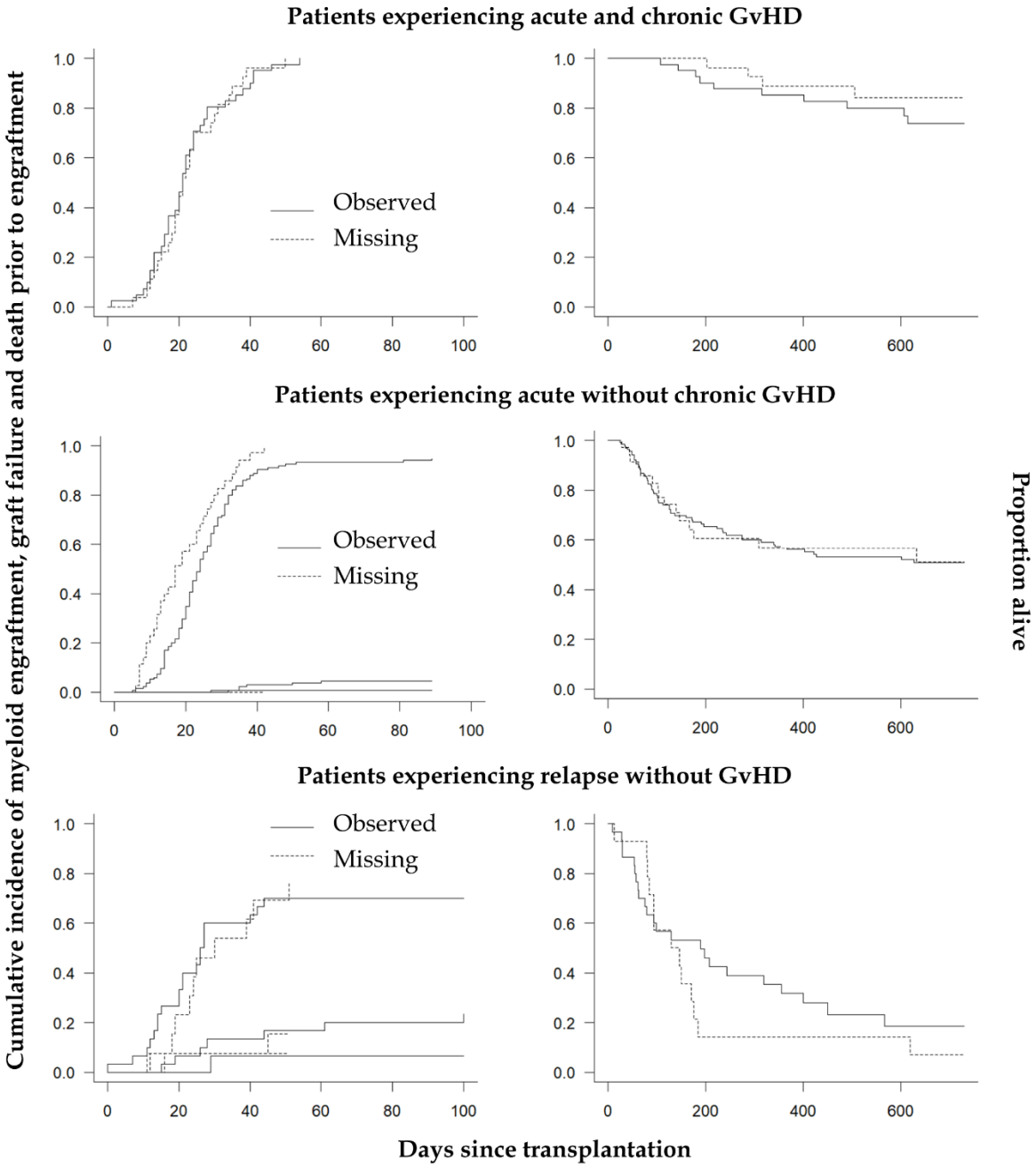
In sub-group (b), the distribution of times of myeloid engraftment was different for patients with missing vs. observed acute GvHD times (p<0.001, Figure 4.2). The median time of engraftment was 17 days (IQR: 11-27 days) and 24 days (IQR: 18-31 days), respectively. Four patients with missing myeloid engraftment or graft failure times were excluded from the analysis of myeloid engraftment.

Comparing patients with missing and observed event times in sub-group (a), there was little difference in the distribution of time to myeloid engraftment (p>0.9, median 21 days, IQR: 17-30 days vs. median 21 days, IQR: 16-27 days). Similarly, for those with missing and observed event times in sub-group (c), there was little difference in the distribution of time to myeloid engraftment (p>0.9, median 30 days, IQR: 23-51 days vs. median 26 days, lower quartile=15 days, upper quartile not achieved).

In addition, comparing patients with missing and observed event times, there was no evidence of a difference in the distribution of survival times for any of the patient sub-groups (p=0.4, survival at 2 years: 84% vs. 74% for sub-group (a); p>0.9, survival at 2 years: 51% vs. 51% for sub-group (b); and p=0.3, survival at 2 years: 7% vs. 19% for sub-group (c), Figure 4.2). In Figure 4.2, there appear to be some small differences in the survival curves for patients in sub-groups (a) and (c). These are based on very small numbers of patients, so it is difficult to draw conclusions about any observed differences.

Figure 4.2. Comparison of patients with observed (solid line) and missing (dotted line) event times.

LH plot: Cumulative incidence of myeloid engraftment and competing events.*
RH plot: Overall survival.



* In sub-group (a) (top LH plot), all patients experienced engraftment.
In sub-groups (b) and (c) (middle and bottom LH plot, respectively), upper pair of lines represent engraftment, middle pair represent graft failure prior to engraftment, and bottom pair represent death prior to engraftment.

4.5.3. Results: multivariable logistic regression of covariates and outcomes associated with missingness of event times

Table 4.7 summarises the results of a multivariable logistic regression analysis of missingness of event times for patients in sub-groups (a), (b) and (c).

Table 4.7. Variables associated with missingness of event times in a multivariable logistic regression analysis.

Covariate (baseline)	Total (N)	Missing event time (N)	Odds ratio (95% CI)	P
Acute and chronic GvHD	68	27		
Country of transplant (not USA)	49	10	1.00 (-)	<0.001
USA	19	17	33.15 (6.55 – 167.76)	
Donor-recipient CMV+ match* (-/-)	18	5	1.00 (-)	0.014
-/+	14	10	6.50 (1.38 – 30.68)	
+/-	18	3	0.52 (0.10 – 2.61)	
+/+	13	7	3.03 (0.68 – 13.61)	
Not reported	5	2	1.73 (0.22 – 13.67)	
Acute without chronic GvHD	**163**	**35**		
Country of transplant (not USA)	131	18	1.00 (-)	<0.001
USA	32	17	7.11 (3.03 – 16.71)	
Time to myeloid engraftment in weeks*	163	35	0.69 (0.50 – 0.94)	0.011
Number of CB units received* (single cord)	94	14	1.00 (-)	0.025
Double cord	69	21	2.42 (1.11 – 5.27)	
Relapse without GvHD	**44**	**14**		
Country of transplant (not USA)	29	3	1.00 (-)	<0.001
USA	15	11	23.83 (4.56 – 124.67)	
Number of CB units received* (single cord)	20	3	1.00 (-)	0.025
Double cord	24	11	4.79 (1.11 – 20.78)	

* In a model excluding country of transplant but including any other listed covariates.

In multivariable analyses, country of transplant was the strongest predictor of missingness of event times in all three patient sub-groups (Table 4.7, $p<0.001$ in each case). Country of transplant was included as a binary variable in regression models (USA or not), due to the small number of missing event times for some

countries. Transplantation in the USA was associated with increased odds of missingness of event times, compared with transplantation elsewhere, but 95% confidence intervals (CI) were wide (odds ratio (OR) 33.15, 95% CI: 6.55 – 167.76 for patients in sub-group (a); OR 7.11, 95% CI 3.03-16.71 for patients in sub-group (b), and OR 23.83, 95% CI 4.56-124.67 for patients in sub-group (c), respectively).

In models including country of transplant, there were no other covariates with strong associations with missingness of event times. However, in models excluding country of transplant, results were as follows. For patients in sub-group (a), the odds that GvHD times were missing were greater for transplants with a CMV- donor(s) and CMV+ recipient compared with transplants in which both donor(s) and recipient were CMV- (OR 6.50, 95% CI 1.38 – 30.68).

Similarly, without adjusting for country of transplant in the analysis of patients in sub-group (b), time to myeloid engraftment (p=0.01) and number of CB units received (p=0.03) were associated with missingness of acute GvHD times. The odds that the acute GvHD time was missing was reduced by 30% for each additional week until myeloid engraftment (OR 0.69, 95% CI 0.50 – 0.94). The odds that the acute GvHD time was missing was also greater for double cord recipients compared with single cord recipients (OR 2.42, 95% CI 1.11 – 5.27). For simplicity, eight patients who did not achieve myeloid engraftment and an additional two patients missing time to engraftment were excluded from the analysis of patients in sub-group (b). Note that in the final analysis of the NHS CBB dataset (Chapter 7), all missing times of engraftment are imputed.

Finally, without adjusting for country of transplant in the analysis of patients in sub-group (c), the number of CB units received (p=0.03) was associated with missingness of relapse times. The odds that the relapse time was missing was nearly five times greater for double cord recipients compared with single cord recipients (OR 4.79, 95% CI 1.11 – 20.78).

To check whether patterns of missing acute GvHD times were different when considering all patients who experienced acute GvHD (not just those without chronic GvHD), the multivariable logistic regression analysis (of variables associated with missingness of acute GvHD time) was repeated for all 251 patients who experienced acute GvHD. The results were very similar: transplants performed in the USA (OR 13.00, 95% CI 6.23 – 27.14) and, without adjusting for country of transplant, double cord compared with single cord transplants (OR 2.56, 95% CI 1.38 – 4.77) were associated with increased odds of missing acute GvHD times. Longer times to myeloid engraftment (OR 0.97, 95% CI 0.94 – 1.01) were associated with reduced odds of missing acute GvHD times.

4.6. Possible missingness mechanisms for event times in the NHS CBB dataset

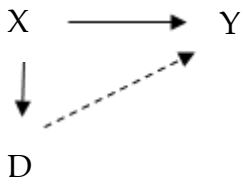
The analyses above suggest that missingness of event times is associated with the event type, the country of transplant, and other observed data. Hence, it is unlikely that event times are MCAR.

Previous clinical studies have found an association between overall survival and the timing of acute GvHD, chronic GvHD and relapse (160-162). For each type of event, earlier onset (*i.e.* smaller event times) was associated with increased mortality. In addition, Lee *et al.* (160) found that early onset acute GvHD was associated with more severe grades of acute GvHD than late onset acute GvHD. In the NHS CBB dataset, I found little difference in overall survival for those with missing and observed times of acute GvHD, chronic GvHD and relapse (see Figure 4.2, described previously). I also found little difference in grade of acute GvHD (comparing patients with missing and observed times of acute GvHD, 73% vs. 67% of patients experienced grade 2-4 acute GvHD, respectively; p=0.8).

The results above suggest that a MNAR mechanism is unlikely. My argument for drawing this conclusion is as follows:

Assume that there is an association between the event time X (for example, the time of acute GvHD) and another variable Y (for example, the time of death), as per the clinical literature. This association is depicted in Figure 4.3, below, by the arrow between X and Y. Also assume that the acute GvHD time, X, is MNAR. In Figure 4.3, this is depicted by the arrow between X and D (a binary variable indicating whether the time of acute GvHD was observed or missing). In this case, a relationship will be induced between the missingness indicator, D, and the time of death, Y, through a backdoor path. In other words, X will behave like a confounder for the relationship between D and Y.

Figure 4.3. Diagram showing relationships between the incompletely observed variable X (time of acute GvHD), the completely observed variable Y (time of death), and D (a binary variable indicating whether the time of acute GvHD was observed or missing).



As described previously, in my analysis of the NHS CBB dataset, I found no evidence of an association between time of death, Y, and the missingness indicator, D. As per Figure 4.3, this implies one of the following:

(i) The time of acute GvHD, X, is not related to the missingness indicator, D, *i.e.* X is not MNAR.

(ii) There is a relationship between X and D, but it is exactly balanced out by the relationship between Y and D (or unmeasured confounders).

(iii) There is no association between X and Y.

Since reason (i) appears the most plausible, it is likely that the time of acute GvHD, X, is not MNAR.

Using a similar argument, the small differences in overall survival (for those with missing and observed times of chronic GvHD and relapse), and in grade of acute GvHD (for those with missing and observed times of acute GvHD), suggest that missingness of event times depends on observed covariates, and not the event times themselves. Therefore, it is more plausible that event times in the NHS CBB dataset are MAR (conditional on the observed data) than MNAR.

4.7. Discussion

In this chapter, I have described patient, donor and transplant baseline characteristics and the types of events experienced by patients in the NHS CBB dataset, including the percentage of missing data for each variable. In addition, I have explored associations between missingness of event times and other variables. Finally, I have described possible missingness mechanisms for event times.

Usually, following allogeneic HSC transplant, patients are not discharged from hospital until there is evidence of myeloid engraftment and acceptable platelet and red cell counts (163). The results in this chapter suggest that events which occur while the patient is still in hospital (myeloid engraftment, graft failure) are well-reported. However, GvHD and relapse can occur once the patient has been discharged from hospital. These events are less well-reported in the NHS CBB dataset. These events may occur outside a clinical setting, with initial identification of symptoms by the patient. The time of onset of a particular condition may not be recorded at a subsequent GP appointment or hospital clinic. In addition, the aetiology of certain symptoms such as skin rashes can be ambiguous, which could also make the time of onset difficult to determine.

There are substantially more missing event times for transplants in the USA than in other countries. The Eurocord Registry, which provided the post-transplant events data for the NHS CBB dataset, cannot contact USA transplant centres

directly. Data retrieval for USA transplant centres relies on affiliation of these centres with a third party (the Center for International Blood and Marrow Transplant Research) (36), which may limit data retrieval. In addition, there are more missing event times for transplants in more recent years than in early years of the study period. This suggests that as the number of transplant centres providing data to the Eurocord Registry has increased over the years, it has become more difficult to obtain complete data.

Since event types, as well as individual-level characteristics, are associated with missingness of event times, CCA will result in biased estimates (see Chapter 3, Section 3.2). These associations suggest that event times are not MCAR. Although it is not possible to rule out a MNAR mechanism, associations which would have suggested a MNAR mechanism (see Section 4.6) are not apparent in the NHS CBB dataset. This suggests that a MAR mechanism (conditional on the observed data) is plausible. Hence, it may be appropriate to use MI methods that assume MAR to handle missing event times. In addition, only 116 transplants in the NHS CBB dataset have complete covariate and outcome data (see Chapter 7, Section 7.6). Therefore, using MI will greatly increase the precision of estimates, compared with CCA.

As discussed in Chapter 3, the optimal MI method for imputing missing event times is not clear, especially, as in my analyses, when there are competing risks or the analysis is a MSM. Key questions are whether it is necessary to constrain the imputed event times to lie within specific boundaries; to what extent skewness should be accounted for in the imputation model; and whether it is important that the imputation scheme incorporates the cumulative incidence function in order to achieve unbiased estimates in a competing risks framework, or the order of the event times in a MSM. In Chapters 5 and 6, I will explore these questions using simulation studies.

CHAPTER 5. IDENTIFYING THE BEST METHOD FOR HANDLING MISSING EVENT TIMES IN A COMPETING RISKS ANALYSIS

5.1. Introduction

In Chapter 4, the types of events experienced by patients in the NHS CBB cohort, and the percentage of events missing exact event times were described. I found that patients experienced various combinations of myeloid engraftment, graft failure, acute GvHD, chronic GvHD, relapse and death. The percentage of missing event times varied by event type; the times of relapse, acute and chronic GvHD were missing in at least 24% of cases, but only small numbers of engraftment and graft failure times were missing, and all death dates were reported.

In Chapter 3, following a review of missing data methods, I concluded that the optimal methods for handling missing event times have not yet been identified. I described the key questions about handling missing event times that are still to be addressed. In this chapter, I evaluate MI strategies for imputing event times in a competing risks analysis, investigating the extent to which interval boundaries, features of the data distribution and of the analysis model should be accounted for in the imputation model. The imputation methods that I consider are: PMM (predictive mean matching); log-linear regression with post-imputation back-transformation; linear regression with and without restrictions on the imputed values, and Delord and Genin's method (136) based on sampling

from the cumulative incidence function. I also evaluate the FML method proposed by Bakoyannis *et al.* (64). Note that I explain how these methods address my research questions in Section 5.2.7.

5.2. Description of simulation studies to assess methods for handling missing event times in a competing risks analysis

5.2.1. Aim of the simulation studies

The aim of the simulation studies was to assess bias and precision when estimating the cumulative incidence function in a competing risks analysis (see Section 5.2.4), comparing the methods described above for handling missing event times. An additional aim was to assess the sensitivity of MI methods to large proportions of missing data and/or misspecification of the imputation model.

5.2.2. Design of the simulation studies

Three simulation studies were performed. The design elements of the three simulation studies are summarised in Table 5.1 overleaf, and described in detail in subsequent sections.

Table 5.1. Design elements of the three simulation studies

Study design element	Study 1	Study 2	Study 3
Data-generating mechanism	Sampling from real data	Sampling from parametric distributions	Sampling from real data
Estimand(s) of interest – cumulative incidence at a single time point	Cumulative incidence of acute GvHD at largest event time	Cumulative incidence of acute GvHD at 100 days post-transplant	Cumulative incidence of acute GvHD at 100 days, chronic GvHD at one year, and relapse at one year post-transplant
Estimand(s) of interest – percentile of event times	Lower quartile time (time by which 25% of patients experienced acute GvHD)	Median time of acute GvHD	Lower quartile time of acute GvHD; 10^{th} percentile time of chronic GvHD and relapse
Number of simulated datasets	1000	1000	1000
Missing data mechanisms	Acute GvHD times MCAR, MAR, MNAR	Acute GvHD times MCAR, MAR, MNAR	Acute GvHD, chronic GvHD, and relapse times MAR, MNAR
Methods for handling missing event times	CCA, various MI methods, and a likelihood-based method	CCA, various MI methods, and a likelihood-based method	Best-performing method from Studies 1 and 2
Performance measures	Standardised bias and model-based SE	Standardised bias and model-based SE	Standardised bias and model-based SE

5.2.3. Choice of data-generating mechanism

Two different sampling methods were used for generating data, to assess whether results were sensitive to the choice of DGM. The two sampling methods were: (i) sampling from the real NHS CBB data, and (ii) sampling from parametric distributions. In the first two simulation studies, in which event times were missing for a single event type, sampling method (i) was used to generate data in the first simulation study and method (ii) was used in the second simulation study. In the third simulation study, in which event times were missing for multiple event types, only method (i) was used. Note that a fourth simulation study, in which event times were missing for multiple event types and sampling method (ii) was used, is described in the next chapter.

DGM 1

In the first DGM, simulated data were generated by sampling with replacement from a full data version of the NHS CBB dataset (FULL-CBB). Censored patients *i.e.* those who had not experienced any of acute GvHD, graft failure or death by last follow-up, were included in the FULL-CBB dataset, because standard techniques for calculating the cumulative incidence can accommodate censored events (164).

DGM1: Description of the creation of the first full data version of the NHS Cord Blood Bank dataset

The FULL-CBB dataset was created to be complete with regards to time of acute GvHD and the competing events of graft failure or death prior to acute GvHD, as well as time of myeloid engraftment (because this variable was a strong predictor of missingness of acute GvHD). Of all 432 transplants in the NHS CBB dataset, cases missing these event times were replaced by sampling from cases with observed event times, as described in Table 5.2. Stratified sampling was used to replace cases with missing acute GvHD times. The stratification variable was the

number of CB units received because this was the strongest predictor of missingness of acute GvHD times, apart from country of transplant (which could not be used in the sampling strategy because there were fewer transplants with observed acute GvHD times than missing in the USA). Stratified sampling was not used to replace cases with missing times of graft failure or myeloid engraftment because there were only a small number of cases with missing times for these event types.

Table 5.2. Strategy to replace incomplete cases when creating the FULL-CBB dataset.

Event	**Cases with missing event time (N)**	**Cases with observed event time (N)**	**Stratification variable***
Acute GvHD	57	184	Number of CB units received (63% double cord)
Graft failure prior to acute GvHD	3	27	None**
Myeloid engraftment	4	425	None**

* Using random sampling without replacement.
** Due to small number of cases with missing event time.

DGM 2

In the second DGM, simulated data were generated using a method similar to that described by Grand *et al.* (165). Event times were generated by direct draws from a parametric function, using inverse transform sampling (166). A separate function was defined for each event type j, where $j = 1, 2, 3$ denotes the events of acute GvHD, graft failure prior to acute GvHD and death prior to acute GvHD, respectively. It was assumed that all patients would experience one of the three event types.

Inverse transform sampling was performed as follows:

a) For each subject i, the event type j was determined with probability 0.65, 0.25 and 0.1, respectively, for $j = 1, 2, 3$.
b) A u_i drawn from a uniform (0,1) distribution was used to determine each event time t_i such that:
   $u_i = p_j(t_i)$

where $p_j(t)$ is the probability density function for the j^{th} event type. A log-normal distribution $p_j(t) = LN(\mu_j, \sigma_j^2)$ was used for each event type, with $\mu_1 = log(26)$, $\sigma_1 = log(2)$; $\mu_2 = log(43)$, $\sigma_2 = log(2)$; $\mu_3 = log(77)$, $\sigma_3 = log(4)$.

The choice of distribution and parameter values were guided by the distribution of event times, measured in days, for the NHS CBB dataset. Each event time was rounded up to the nearest whole number to increase computational speed and to reflect the real data. Administrative censoring at one year post-transplant was applied for all patients, to reflect usual practice in transplant registry studies *i.e.* follow-up until a fixed time-point of clinical interest. One auxiliary variable was included for each subject, that is, whether the subject received a double cord transplant (rather than a single cord), by sampling from a Bernoulli distribution with probability 0.45. This variable was chosen because, in the real NHS CBB dataset, the number of CB units received is predictive of the acute GvHD times themselves, as well as the missingness of acute GvHD times.

DGM 3

In the third DGM, simulated data were generated using the same method as described for DGM 1, by sampling with replacement from a full data version of the NHS CBB dataset (FULL-CBB2).

DGM3: Description of the creation of the second full data version of the NHS Cord Blood Bank dataset

The FULL-CBB2 dataset was created to be complete with regards to all event times. Of all 432 patients in the NHS CBB dataset, 90 with any missing event times were replaced by sampling from patients who had experienced similar combinations of events and had all event times observed, as described in Table 5.3. As before, country of transplant and donor-recipient CMV status match were not used as stratification variables, even though they were the strongest predictors of missingness of event times, because there were fewer patients with observed times than missing in some categories of these variables.

Table 5.3. Strategy to replace incomplete cases when creating the FULL-CBB2 dataset.

Event(s) experienced	**Cases with missing event time(s) (N)**	**Cases with observed event time(s) (N)**	**Stratification variable***
Acute and chronic GvHD	27	41	None**
Acute without chronic GvHD	41	132	Number of CB units received (56% double cord)
Chronic without acute GvHD	5	9	None**
Relapse without GvHD	14	30	Number of CB units received (79% double cord)
Graft failure without GvHD nor relapse	2	17	None**
Myeloid engraftment without GvHD nor relapse	1	95	None**

* Using random sampling without replacement.
** Due to small number of cases with missing event time.

5.2.4. Estimands of interest

In the first and second studies, in which event times were missing for a single event type, acute GvHD was the event of interest. In the third study where event times were missing for multiple event types, acute GvHD, chronic GvHD and relapse were the events of interest. For each event of interest, the estimands of interest were:

(i) **Cumulative incidence at specific time points**

The specific time points chosen for each event of interest in each study were guided by the DGMs (see Table 5.1). In the first simulation study, in which data were generated by sampling from the real NHS CBB data, the specific time point chosen for acute GvHD was the largest event time, *i.e.* the estimand was the percentage of patients who experienced acute GvHD at some point post-transplant. Acute GvHD only occurs during approximately the first 100 days post-transplant, according to the standard clinical definition of acute GvHD (17) (observed times were within 119 days in the NHS CBB dataset). This clinical criterion imposes a limit on

the maximum event time and it seemed reasonable not to impose a further restriction by estimating the cumulative incidence at a particular time-point. However, there is potential for imputed event times to be far larger than 100 days, making this estimand sensitive to outlying values. Therefore, in the second and third simulation studies, the cumulative incidence was estimated only within the real clinical period in which acute GvHD could occur, that is, the cumulative incidence was estimated at 100 days post-transplant. Chronic GvHD and relapse can occur at any point post-transplant. Hence, the cumulative incidence was estimated at a time-point typically of clinical interest, namely, at one year post-transplant.

I expected little bias in estimates of the cumulative incidence of acute GvHD because event types were not imputed (*i.e.* although event times were sometimes missing, the associated event types were always observed), and the cumulative incidence was estimated at the point by which most, if not all, events should have occurred. However, it is less clear that the cumulative incidence of chronic GvHD and relapse would be unbiased.

(ii) **Percentiles of event times**

As above, the percentiles of event times chosen for each event of interest in each study were guided by the DGMs (see Table 5.1). Percentiles were chosen so that they would occur before the time point of the associated cumulative incidence of interest. For example, when the cumulative incidence of acute GvHD at 100 days post-transplant was the estimand of interest, the percentile of interest occurred within 100 days. I included percentiles of event times as a measure of the shape of the cumulative incidence function. I wanted to assess whether incorrect imputation affected the shape of the cumulative incidence function more than the estimate at any particular time-point.

In clinical studies, the median event time is more commonly reported than other percentiles. However, in the first and third simulation studies (in which data were generated by sampling from the real NHS CBB data), the lower quartile time of acute GvHD was used as the estimand of interest. This was because acute GvHD was reported for just over 50% of patients in the original NHS CBB cohort, so the median was unlikely to be reached in all the simulated datasets. However, in the second study, the DGM was designed so that the median could be estimated in every simulated dataset. For chronic GvHD and relapse, due to the low incidence of these events, the 10th percentile time was chosen as the estimand of interest. In each study, the design of the DGM ensured the 10th percentile was reached within one year of transplant.

Statistical methods

The cumulative incidence function was estimated using the non-parametric Aalen-Johansen estimator (167, 168) (see Chapter 2, Section 2.2); its standard error (SE) was estimated using the Greenwood-style estimator described by Marubini and Valsecchi (169), as this has been shown to be more accurate than other proposed estimators (170, 171). An estimator of the SE of a percentile was derived using the delta method (see Section 5.4.9). Analysis was performed using the 'mstate' R package (172).

5.2.5. Number of simulation datasets

In each study, one thousand simulated datasets were created using the DGMs described above. In Studies 1 and 3, the size of each dataset was 432, which was the size of the real NHS CBB dataset. In Study 2, the size of each dataset was 500; this was chosen to approximately match the size of the real dataset. Based on the mean estimates and empirical SE for the 1000 datasets, with complete data (*i.e.* with no missing event times, see Section 5.4), and using the formula suggested by Burton *et al.* (173) with a type 1 error of 5%, 1000 datasets allows estimation of the

cumulative incidence and percentiles of event times with an accuracy of 0.1% and 0.1 days, respectively.

5.2.6. Methods for simulating missing data

In Studies 1 and 2, a MCAR missing data mechanism (MDM) was considered by setting a random 10%, 30% or 50% of all event times to missing, regardless of event type. Next, MAR and MNAR MDMs were considered in all three studies. In each MAR or MNAR scenario, missingness depended on event type, such that only times of the event of interest (acute GvHD, chronic GvHD or relapse) were missing, but not times of the competing events (graft failure or death prior to the event of interest). Although event times MAR (conditional on the observed data) is the most likely MDM for the real NHS CBB cohort (see Chapter 4, Section 4.6), MNAR MDMs were also considered to assess the impact on bias and precision if MI was used in this scenario. The MAR and MNAR scenarios are described below. Scenarios with a different design for each simulation study are identified separately.

In each scenario, j = 1, 2, 3 denotes the event experienced (j = 1: acute GvHD, j = 2: chronic GvHD, j = 3: relapse). For each subject i, π_{ij} denotes the probability that event times for the j^{th} event are missing; x_{1i} = 1 for a double cord transplant and 0 otherwise; x_{2i} is the number of weeks from transplant to myeloid engraftment; t_{ij} is the event time for subject i for the j^{th} event and $t_{j(p\%)}$ is the p^{th} percentile of event times for the j^{th} event, ordered from smallest to largest. Note that times of graft failure and death are always observed and hence, for brevity, probability statements are not included below for these event types. The missingness scenarios are:

a) MAR: The probability that event times are missing depends only on the event type, such that 10%, 30% or 50% of event times are missing for acute GvHD, and graft failure and death times are fully observed. This scenario was considered in Studies 1 and 2 only:
   π_{i1} = 0.1, 0.3 or 0.5

b) MAR: The probability that event times are missing depends on the event type and covariates, with 10%, 30% or 50% of event times missing for acute GvHD, and graft failure and death times are fully observed. Note that Study 2 included only one covariate, x_{1i} (the number of CB units received), in the DGM:

$\pi_{i1} = c \times \text{logit}^{-1}(-0.914 + 0.910x_{1i} - 0.220x_{2i})$ in Studies 1 and 3

$\pi_{i1} = c \times \text{logit}^{-1}(-0.9 + 0.9x_{1i})$ in Study 2

In each study, the constant c was chosen so that approximately 10%, 30% or 50% of all event times were missing.

In Study 3, 10%, 30% or 50% of event times were also missing for chronic GvHD and relapse. Missingness definitions were the same for acute and chronic GvHD. This reflected the fact that GvHD times tended to be both reported or both missing in the real NHS CBB dataset (although in the simulation study, as in the real data, it was possible for the acute GvHD time to be missing and the chronic GvHD time to be observed, and vice versa):

$\pi_{i2} = c \times \text{logit}^{-1}(-0.914 + 0.910x_{1i} - 0.220x_{2i})$

$\pi_{i3} = c \times \text{logit}^{-1}(-1.735 + 1.472\, x_{1i})$

As before, for each event type (j=1,2,3), the constant c was chosen so that approximately 10%, 30% or 50% of all event times were missing. The coefficients in the above formulae were the parameter estimates from logistic regression models applied to the original NHS CBB dataset (see Chapter 4, Section 4.5.3).

c) MNAR: The shortest 10%, 30% or 50% of all event times are missing for acute GvHD, and graft failure and death times are fully observed:

$$\pi_{i1} = \left\{ \begin{array}{l} 1 \text{ if } t_{i1} < t_{1(p\%)} \\ 0 \text{ otherwise} \end{array} \right\} \quad \text{in Studies 1, 2 and 3}$$

where p = 10, 30 or 50.

Additionally, in Study 3, the shortest 10%, 30% or 50% of all event times are missing for chronic GvHD and relapse, and graft failure and death times are fully observed:

$$\pi_{ij} = \begin{cases} 1 \text{ if } t_{ij} < t_{j(p\%)} \\ 0 \text{ otherwise} \end{cases}$$

where j = 2, 3 and p = 10, 30 or 50

d) MNAR: The longest 10%, 30% or 50% of all event times are missing for acute GvHD, and graft failure and death times are fully observed. This scenario was considered in Studies 1 and 2 only:

$$\pi_{i1} = \begin{cases} 1 \text{ if } t_{i1} > t_{1(p\%)} \\ 0 \text{ otherwise} \end{cases}$$

where p = 90, 70 or 50.

The MNAR mechanisms described above could occur in practice if patients with early-onset acute GvHD tended to have milder or ambiguous symptoms. This could make identification of the exact date of onset more difficult and would result in shorter acute GvHD times being more likely to be missing. Alternatively, if acute GvHD occurred after the patient was discharged from hospital, diagnosis may only occur at infrequent clinic visits. This scenario would result in longer acute GvHD times being more likely to be missing.

5.2.7. Methods for handling missing event times

Studies 1 and 2

In Studies 1 and 2, the following methods for handling missing event times were considered in all MCAR, MAR and MNAR scenarios:

(i) Linear imputation model with no restrictions on the imputed values (NORM), as per Rodwell's advice (97) to impute skewed continuous variables on the original scale with no range-restrictions and no post-

imputation rounding (see Chapter 3, Section 3.4.4). This is implemented in the 'mice' R package using the method='norm' option.

(ii) Linear imputation model, as per (i), without auxiliary variables (*i.e.* any variable other than event type) in the imputation model (NORMNOAUX). The aim of this scenario was to assess whether bias and precision were worse if variables predictive of missing event times, and/or the probability that they are missing, are excluded from imputation models.

(iii) Type 1 PMM imputation model with no restrictions on the imputed values (PMM), as per Lee and Carlin's (98) finding that it is preferable to use type 1 PMM than to transform non-normal data (see Chapter 3, Section 3.4.4). Type 1 PMM is implemented in the 'mice' R package using the method='pmm' option and the default of five donors. This number of donors has been shown to provide adequate coverage and efficiency (92).

(iv) Type 1 PMM, as per (iii), without auxiliary variables in the imputation model (PMMNOAUX). This method was considered for the same reason as NORMNOAUX.

(v) Log-linear imputation model with post-imputation back-transformation (LOGNORM). I considered this method to assess whether imputation of percentiles was improved when the normal distribution assumption was made more plausible, as per von Hippel's advice (121) (see Chapter 3, Section 3.4.4). The natural logarithm of acute GvHD time was imputed and the exponential of the imputed time was used in the analysis model.

(vi) Linear regression with restrictions on the imputed values (RESNORM). I considered this method to test Rodwell's comment that restricting the imputed values can increase rather than decrease bias (see Chapter 3, Section 3.4.4). Within each imputation step, for each missing time value, a

value was drawn and compared with pre-specified boundaries. If the drawn value was outside the boundaries, then a new value was drawn. This process was repeated until all imputed values were within the boundaries or until the process had been carried out 200 times. The number of repeats was chosen to be large to minimise the number of imputed values outside the boundaries.

In both studies, a minimum boundary of 0 was used. In Study 1, a maximum boundary of 119 days or the time of the patient's death (if this was observed within 119 days) was used. A value of 119 days was used as the maximum because this was the largest observed time of acute GvHD in the NHS CBB dataset. In the NHS CBB dataset, of the 251 patients who experienced acute GvHD, 45 died within 119 days of transplant. Hence, for most patients, the maximum boundary was 119 days.

In Study 2, a maximum boundary of 100 days was used, for consistency with the clinical criteria. Times of death after acute GvHD were not generated in Study 2 (only the time of the first event experienced was generated), hence the boundary of 100 days applied for all patients in Study 2. As a sensitivity analysis, in Study 2, for the first MAR scenario, RESNORM was also implemented with the boundary comparison performed up to 500 times.

(vii) The MI method proposed by Delord and Genin (136) (MICI). In this method, missing event times are sampled from the observed event times based on the current estimate of the cumulative incidence function, conditional on user-specified boundaries. The sampling probability is determined by the current estimate of the cumulative incidence function. The boundaries are as specified for RESNORM.

(viii) The semi-parametric maximum likelihood approach of Bakoyannis *et al.* (64) for interval-censored event times (INTCCR), estimating the baseline cumulative incidence function from a proportional sub-distribution hazards (Fine and Gray (3)) model fitted without covariates. Interval boundaries are as specified for RESNORM. This approach was used to compare a FML method with MI methods. I did not use a strictly non-parametric FML method because Sun (47) advises that such methods should be avoided when censoring intervals are wide. In addition, I did not consider a parametric FML method because this would not allow a direct comparison with the non-parametric analysis that was performed when MI methods were used. To the best of my knowledge, INTCCR is the only semi-parametric FML method for handling interval-censored competing risks data that can be applied using standard software. It was of interest to assess the authors' suggestion that their method allows better estimation of SE than MICI (174).

In Study 1, two additional methods for handling missing event times were considered in MAR and MNAR scenarios, namely:

(ix) Single imputation by replacement with the median of the observed acute GvHD times (MED).

(x) Single imputation by replacement with the mid-point of the specified boundaries (MID).

Naïve methods such as these have been used in HSC studies (45). These methods were included to demonstrate their inferior performance compared with MI or FML approaches.

I compare results of the above methods with those of a complete case analysis (CCA).

Study 3

In Study 3, only the best method identified from Studies 1 and 2 was applied, to confirm that similar results were obtained when event times were missing for multiple event types.

Variables used in imputation models

In all three studies, for MI methods (i), (iii), (v) and (vi), imputation models included auxiliary variables, as well as the type of event experienced (which was required for compatibility with the analysis model). Note that in MI methods (ii) and (iv), (NORMNOAUX and PMMNOAUX), auxiliary variables were deliberately excluded. It is not possible to include auxiliary variables in Delord and Genin's method (MICI) (175). In Studies 1 and 3, the auxiliary variables were those predictive of event times (based on the clinical literature, Chapter 1, Table 1.1) or predictive of event times and their missingness (based on analysis of the NHS CBB dataset, Chapter 4, Section 4.5). These were: all baseline patient, donor and transplant variables; the time from transplant to myeloid engraftment or competing event (graft failure or death prior to engraftment); indication of myeloid engraftment (engraftment vs. graft failure or death prior to engraftment); post-transplant survival time; indication of alive or dead at the survival time. In Study 2, the auxiliary variable was the number of CB units received.

Imputation method implementation

MI methods (i) to (vi) were implemented using the 'mice' R package (19). To assess the standard performance of this package, the default of five imputations was used in all studies. This number of imputations generally gives estimates with adequate efficiency (87). However, White *et al.* (87) argue that, to ensure adequate reproducibility (small Monte Carlo error) as well as efficiency, the number of imputations should equal at least the percentage of incomplete cases. Therefore, as a sensitivity analysis, in Study 2, for the first MAR scenario, methods were also implemented using 50 imputations (because the largest

percentage of missing event times was 50%). In Studies 1 and 2, only one variable (the time of acute GvHD) was incomplete. Therefore, no iteration was performed *i.e.* the number of cycles per imputation was one. In Study 3, the default of five iterations was used. Van Buuren found that convergence is generally achieved within this small number of iterations; notably, the number of iterations required to achieve convergence in MI is much less than in other applications of Markov chain Monte Carlo methods (90).

MICI was implemented using an adaptation of the 'MIICD' R package (20). I adapted the code used in the published package in order to output the set of imputed cumulative incidence functions. This enabled me to correct a mistake in the calculation of the SE of the cumulative incidence in the published code (I have alerted the package authors), and to calculate the SE of the percentile times (see Appendices, Section A.6).

INTCCR was implemented using the 'intccr' R package. This package cannot handle exactly observed times and therefore boundaries were specified for all acute GvHD times. For the exactly observed times in the simulated datasets, the left boundary was set to be the observed time minus one day and the right boundary to be the observed time. Graft failure and death prior to acute GvHD were combined into one event type, because the package only allows for two event types. Non-parametric bootstrap sampling is used in the package to estimate the SE of regression parameter estimates. I adapted the published code for calculating SE so that I could apply the same method when calculating the SE for each of my estimands (see Appendices, Section A.6).

All other methods were implemented using my own R code (see Appendices, Section A.6). R version 3.5.2 and 'mice' version 3.3.0 was used for all results.

5.2.8. Performance measures

In all three studies, for each estimand β (following the notation of Burton et al. (173)), performance measures of interest were:

(i) Standardised bias, defined as $\left(\frac{\bar{\hat{\beta}}-\beta}{SE(\hat{\beta})}\right)$, where $\bar{\hat{\beta}}=\sum_{k=1}^{1000}\hat{\beta}_k/1000$, $\hat{\beta}_k$ is the estimate for the k^{th} simulated dataset, $SE(\hat{\beta})$ is the standard deviation of the $\hat{\beta}_k$ and β is the true value. I used standardised bias, rather than absolute bias, as a performance measure, following the advice of Burton *et al.* (173). They note that standardised bias can be more informative than bias or percentage bias because it accounts for the uncertainty of the estimate. In addition, standardised bias allows comparison between estimands that vary in scale. Hereafter, for brevity, the term "bias" will refer to the standardised bias. Any reference to absolute bias will be identified as such.

(ii) Average model-based SE *i.e.* $\sqrt{\sum_{k=1}^{1000}\widehat{SE}^2(\hat{\beta}_k)/1000}$ where $\widehat{SE}(\hat{\beta}_k)$ is the estimated standard error (SE) within each simulated dataset.

In Studies 1 and 3, the true values of each estimand were defined as the values calculated using the full data versions of the NHS CBB dataset (FULL-CBB and FULL-CBB2, respectively). In Study 2, since acute GvHD times were generated from a specific parametric distribution, I calculated the true value of each estimand analytically using standard results (176).

5.3. Description of the estimator of the standard error of a percentile of event times

In the simulation studies, I derived an estimator of the model-based SE of a percentile of event times, by adapting Collett's estimator of SE for the median all-cause survival time (53). I replaced the survivor function in Collett's definition with the cumulative incidence function. This estimator is based on the delta method. It is defined as follows, for the p^{th} percentile of event times for the j^{th} event type (for brevity, this estimand is referred to as $\hat{t}(p)$ and the subscript j is not shown in the definitions below):

$$\widehat{SE}[\hat{t}(p)] = \frac{1}{\hat{f}\{\hat{t}(p)\}} \widehat{SE}[\hat{F}\{\hat{t}(p)\}]$$

The SE of the cumulative incidence function at $\hat{t}(p)$, $\widehat{SE}[\hat{F}\{\hat{t}(p)\}]$, was calculated using the Greenwood-style estimator (167, 169) mentioned previously.

I estimated the probability density function for the cumulative incidence at $\hat{t}(p)$, $\hat{f}\{\hat{t}(p)\}$, by calculating a local gradient at $\hat{t}(p)$, such that:

$$\hat{f}\{\hat{t}(p)\} = \frac{\hat{F}\{\hat{u}(p)\} - \hat{F}\{\hat{l}(p)\}}{\hat{u}(p) - \hat{l}(p)}$$

where $\hat{u}(p) = \min\left\{t_i \middle| \hat{F}(t_i) \geq \frac{p}{100} + \epsilon\right\}$ and $\hat{l}(p) = \max\left\{t_i \middle| \hat{F}(t_i) \leq \frac{p}{100} - \epsilon\right\}$,

for all event times i for the j^{th} event type and small ϵ.

Collett warns that his estimator of SE is only approximately correct, but that superior methods are much more computationally difficult (53). Brookmeyer and Crowley (177) also commented that it can be difficult to obtain an accurate estimate of the probability density function of a percentile.

5.3.1. Methods: study to assess the estimator of the standard error of a percentile of event times

I performed an additional simulation study to assess my estimator of SE of a percentile of event times, using 1000 simulated datasets. My estimand of interest was the lower quartile time of acute GvHD. Various elements of the estimator which may affect its performance were considered. These were combined factorially in the simulation studies:

(i) The size of ϵ. The effect of the size of ϵ on the SE estimate was explored by comparing performance for six different values of ϵ (0.005, 0.01, 0.03, 0.05, 0.1, 0.15).

(ii) The estimator of the probability density function. A simple gradient function is used by Collett to estimate the probability density function $\hat{f}\{\hat{t}(25)\}$. Here, as an alternative, the gradient coefficient from a univariable linear regression of the observed incidence over time was also considered.

(iii) The sample size. To investigate whether sample size affected bias and coverage of the estimator of SE, analysis was performed using a simulated dataset of the same size as the real NHS CBB dataset (DGM1, N=432), and a large dataset (DGM2, N=10,000). For DGM1, data were generated using the same method as in Study 1 (see Section 5.2.3). For DGM2, data were generated using the same method as in Study 2 (see Section 5.2.3).

(iv) The units used to measure event times. In DGM1, since data were generated by sampling from the real NHS CBB dataset, event times were measured in whole days. In DGM2, I was able to vary the units used to measure event times: I used whole days and 0.1 day increments.

Performance measures of interest were the average model-based SE and coverage. Coverage was defined as the proportion of 95% CI including the true value of the lower quartile, where the 95% CI was defined for the k^{th} simulated dataset as $\hat{t}_k(25) \pm 1.96\ \widehat{SE}_k[\hat{t}_k(25)]$ (53). For reference, the absolute bias of the estimated lower quartile time of acute GvHD, the empirical SE, and the Monte

Carlo 95% CI of absolute bias (178) were also calculated. In addition, for DGM2, since acute GvHD times were generated from a specific parametric distribution, I calculated the theoretical lower quartile time of acute GvHD and its SE analytically using standard results (176).

5.3.2. Results: study to assess the estimator of the standard error of a percentile of event times

Results of the study to assess my estimator of SE are shown in Table 5.4.

Table 5.4. Performance of the estimator of SE for the lower quartile time of acute GvHD.

Sample size	**Estimator of the probability density function**				
	ϵ	**Simple gradient**		**Linear regression**	
		ModSE	**Coverage**	**ModSE**	**Coverage**
432 (time measured in whole days)	**Absolute bias in lower quartile = -0.38; EmpSE = 2.04; TheorSE n/a**				
	0.005	1.85	0.78	2.12	0.85
	0.01	1.89	0.84	2.09	0.87
	0.03	1.85	0.88	2.01	0.90
	0.05	1.79	0.89	1.95	0.91
	0.1	1.73	0.92	1.82	0.92
	0.15	1.81	0.92	1.81	0.92
10,000 (time measured in whole days)	**Absolute bias in lower quartile = -0.08; EmpSE = 0.29; TheorSE=0.24**				
	0.005	0.26	0.91	0.26	0.91
	0.01	0.26	0.91	0.26	0.91
	0.03	0.26	0.91	0.26	0.91
	0.05	0.27	0.91	0.27	0.91
	0.1	0.27	0.91	0.27	0.91
	0.15	0.28	0.91	0.28	0.91
10,000 (time measured in 0.1 days)	**Absolute bias in lower quartile = -0.01; EmpSE = 0.26; TheorSE=0.24**				
	0.005	0.26	0.95	0.26	0.95
	0.01	0.26	0.95	0.26	0.95
	0.03	0.26	0.95	0.26	0.95
	0.05	0.26	0.95	0.26	0.95
	0.1	0.27	0.95	0.27	0.95
	0.15	0.28	0.96	0.28	0.96

ModSE, average model-based SE; EmpSE, empirical SE; TheorSE, theoretical SE
Monte Carlo CI for absolute bias is (-0.50, -0.26) for N=432; (-0.10, -0.06) for N=10,000, time in whole days; (-0.03, 0.01) for N=10,000, time in 0.1 days

Table 5.4 shows that the estimated lower quartile time of acute GvHD was slightly biased for both sample sizes when time was measured in whole days. It was unbiased (*i.e.* the Monte Carlo interval for absolute bias contained 0) when the sample size was 10,000 and time was measured in 0.1 days. In the former scenarios, the negative bias in the estimated lower quartile time explains the

under-estimation of SE using my estimator and the under-coverage (75) (when using a simple gradient to calculate the probability density function). In the latter scenario, in which the estimated lower quartile time was unbiased, estimates of SE using my estimator were close to the empirical SE and theoretical SE. In addition, coverage was equal to the nominal value when ϵ was 0.05 or smaller. For very small ϵ in combination with a dataset of size 432, larger estimates of SE were obtained if linear regression was used, compared with using a simple gradient to calculate the probability density function. Otherwise, results were very similar whether using a simple gradient or linear regression.

In summary, when using this estimator of SE, these results suggest that ϵ between 0.01 and 0.05 is the best choice to minimise bias and optimise coverage. This is particularly the case when rounding of event times cannot be avoided. Therefore, in the first and third simulation studies, ϵ of 0.03 was used. In the second study, ϵ of 0.01 was used. For simplicity, a simple gradient function was used for probability density function estimation.

5.4. Results: simulation study to assess methods for handling missing event times in a competing risks analysis

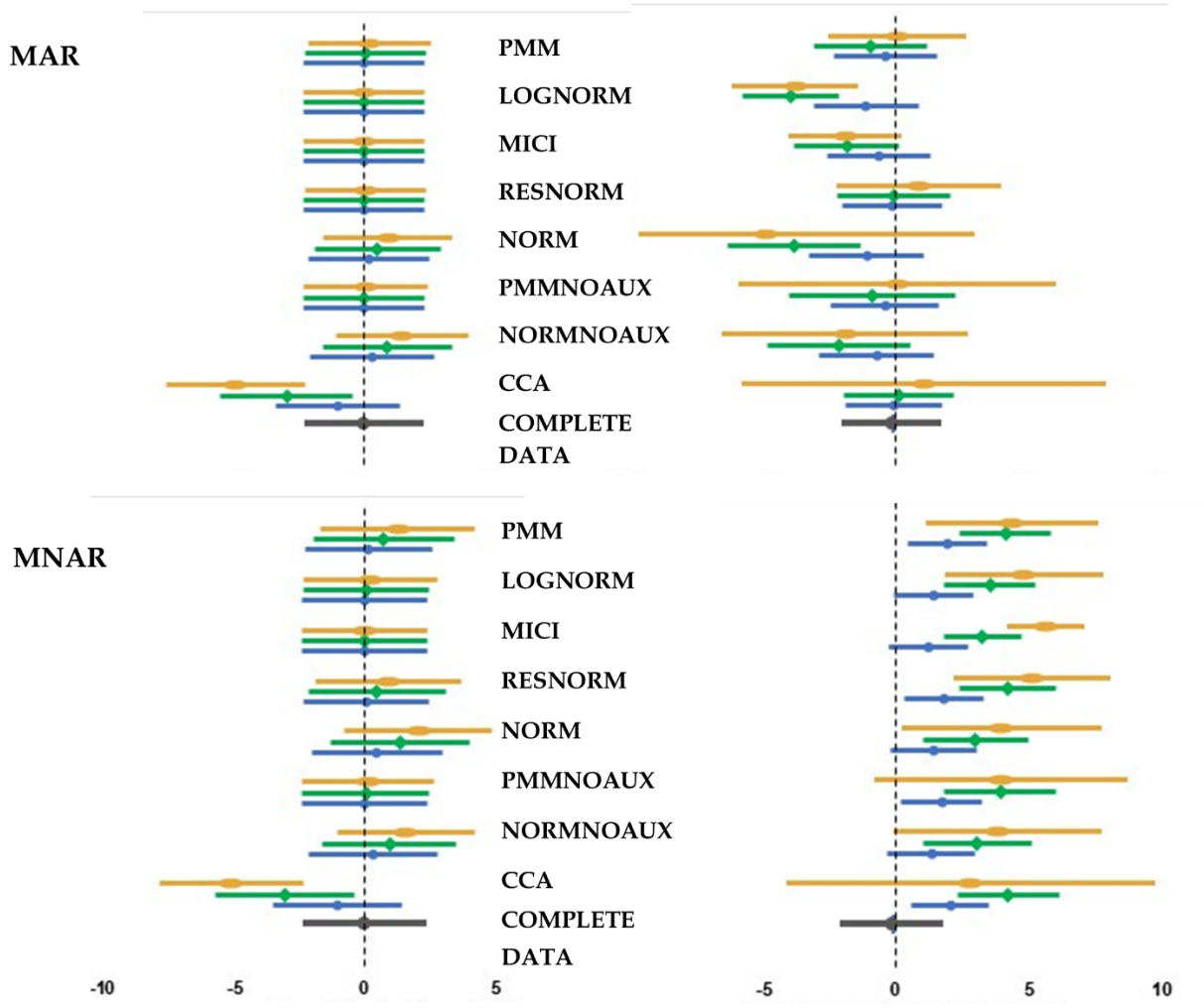
5.4.1. Results from Studies 1 and 2: event times missing for a single event type

The true values of the estimands in Study 1 (in which simulated data were generated by sampling from the real NHS CBB dataset) were 55.58% for the cumulative incidence at the largest event time and 26.00 days for the lower quartile time of acute GvHD. In Study 2 (in which simulated data were generated from parametric distributions), the true values were 63.11% for the cumulative incidence at 100 days and 44.00 days for the median time of acute GvHD. In both studies, estimates of cumulative incidence based on complete simulated data (*i.e.* without any missing event times) were unbiased (given Monte Carlo 95% CI for absolute bias of (-0.15, 0.15) and (-0.10, 0.18) for Studies 1 and 2, respectively). In both studies, estimates of percentile times based on complete simulated data were slightly biased (given Monte Carlo 95% CI for absolute bias of (-0.44, -0.18) and (0.50, 0.96) for Studies 1 and 2, respectively). This bias may be due to measuring event times in whole days, as per the study results described in Section 5.3.2.

Figures 5.1 and 5.2 show the (standardised) bias $\left(\frac{\bar{\hat{\beta}}-\beta}{SE(\hat{\beta})}\right)$ and average model-based SE of the estimands of interest in each study, for MAR and MNAR MDMs, for various MI methods, and percentages of missing data. In these figures, methods for handling missing data are ranked in preferential order (most preferred at the top of each plot), based on bias and precision. For reference, the complete data estimates are also shown.

Figure 5.1. Study 1: Standardised bias and average model-based SE for cumulative incidence of acute GvHD at the largest event time, and lower quartile time of acute GvHD.

10% (blue circle), 30% (green diamond) or 50% (yellow oval) of event times are set to missing using the following missing data mechanisms: MAR (depending on event type and covariates) and MNAR (shortest times missing).



PMM, MI by Type 1 predictive mean matching with no restrictions on the imputed values

LOGNORM, MI by log-normal imputation with post-imputation back-transformation

MICI, Delord and Genin's MI method

RESNORM, MI by normal regression with restrictions on the imputed values

NORM, MI by normal regression with no restrictions on the imputed values

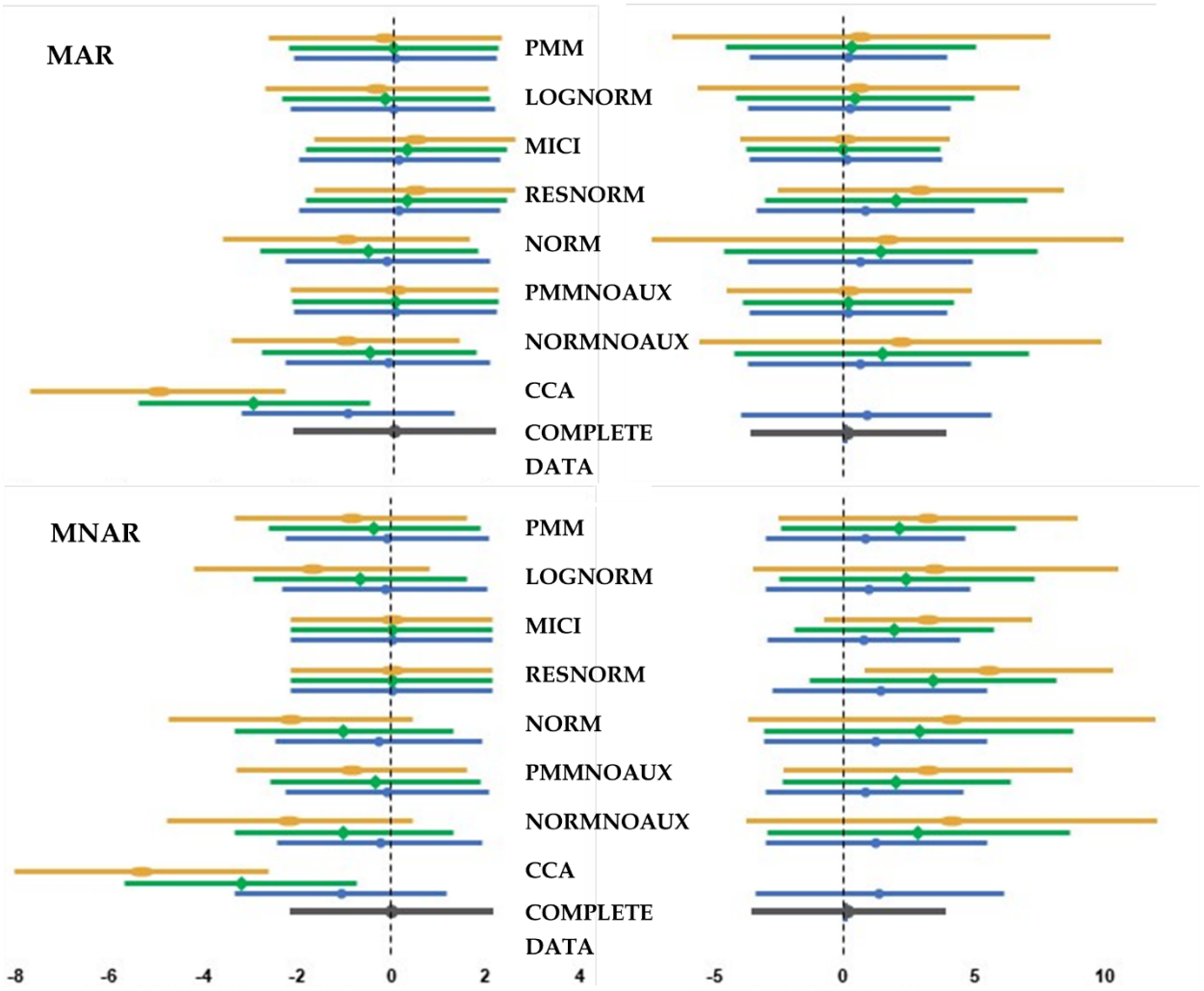
PMMNOAUX, as for PMM excluding the auxiliary variables from the imputation model

NORMNOAUX, as for NORM excluding the auxiliary variables from the imputation model

CCA, complete case analysis

Figure 5.2. Study 2: Standardised bias and average model-based SE for cumulative incidence of acute GvHD at 100 days and median time of acute GvHD.

10% (blue circle), 30% (green diamond) or 50% (yellow oval) of event times are set to missing using the following missing data mechanisms: MAR (depending on event type and covariate) and MNAR (shortest times missing)



*In complete case analyses, less than 50% patients experienced aGvHD so the median time to aGvHD could not be estimated.

PMM, MI by Type 1 predictive mean matching with no restrictions on the imputed values

LOGNORM, MI by log-normal imputation with post-imputation back-transformation

MICI, Delord and Genin's MI method

RESNORM, MI by normal regression with restrictions on the imputed values

NORM, MI by normal regression with no restrictions on the imputed values

PMMNOAUX, as for PMM excluding the auxiliary variable from the imputation model

NORMNOAUX, as for NORM excluding the auxiliary variable from the imputation model

CCA, complete case analysis

Results were generally similar for MCAR and both MAR MDMs considered; likewise, results were similar for both MNAR MDMs considered. Therefore, only results for one MAR scenario (the most likely MDM for the NHS CBB dataset, *i.e.* where the probability that event times are missing depends on the event type and covariates) and one MNAR scenario (shortest times missing) are illustrated in Figures 5.1 and 5.2.

MED and MID methods were included in Study 1 for completeness but are not illustrated in Figure 5.1. Results using the semi-parametric maximum likelihood approach (INTCCR) are also not illustrated in Figures 5.1 and 5.2. This was because results for INTCCR were very similar to those for MICI, and because MI, not FML methods, are the focus of my studies. Results for these three methods, and other key features of the results from each study, are summarised below. Tables of all results are included in the Appendices (Tables A.2 and A.3).

Study 1 results

In Study 1, when event times were MCAR (*i.e.* when there was an equal probability of any acute GvHD, graft failure or death time being missing), CCA estimates of the cumulative incidence were unbiased as expected. As per the complete data results, CCA estimates of the lower quartile time of acute GvHD had a small amount of bias (given 10%, 30% and 50% missing times, bias for the lower quartile was: -0.14, -0.13, -0.16, respectively). Of the MI methods, only PMMNOAUX and LOGNORM estimates of the cumulative incidence were unbiased. For these methods, estimates of the SE of the cumulative incidence were smaller than CCA estimates of SE. PMMNOAUX and LOGNORM estimates of the lower quartile time of acute GvHD were also least biased (given 10%, 30% and 50% missing times, bias < 0.8 in each case) of all MI and FML methods. For these methods, estimates of the SE of the lower quartile time were comparable to CCA estimates (except given 50% missing event times, in which case PMMNOAUX estimates of SE were larger than for the other two methods). NORM, MICI and INTCCR (the FML method) estimates of the cumulative

incidence were most biased. Of all methods considered, RESNORM estimates of the lower quartile time of acute GvHD were most biased. Reasons for this are considered in Section 5.5, following these results.

When event times were MAR or MNAR, CCA estimates of the cumulative incidence greatly underestimated the true cumulative incidence (Figure 5.1). This was because times for outcomes other than acute GvHD (death or graft failure prior to acute GvHD) were fully complete and only cases of acute GvHD were under-represented. When data were MAR and missingness depended on event type and covariates (Figure 5.1), bias was small for all methods (including INTCCR, the FML method) except CCA and methods based on unrestricted linear imputation (NORM and NORMNOAUX). Methods based on an unrestricted linear imputation model (NORM and NORMNOAUX) over-estimated the cumulative incidence. Reasons for this are considered in Section 5.5, following these results. Restricting the range of imputed values (RESNORM) reduced the bias of the cumulative incidence estimates compared to NORM and NORMNOAUX.

When event times were MNAR (shortest times missing, Figure 5.1), all methods except LOGNORM, MICI and PMMNOAUX over-estimated the cumulative incidence, with performance worsening as the percentage of missing data increased. However, results for LOGNORM, MICI, PMMNOAUX and INTCCR were unbiased, even with a large percentage of missing data. As expected, SE for most methods increased with the volume of missing data, reflecting the additional uncertainty due to imputation. However, SE for MICI, LOGNORM and RESNORM (in the MAR scenario) was the same as the complete data SE, regardless of the percentage of missing data (see Section 5.5 for an explanation of these results).

In general, bias was larger for estimates of the lower quartile than for cumulative incidence (Figure 5.1). In the MAR scenario, most methods, except CCA and

RESNORM, under-estimated the lower quartile. RESNORM and PMM were the least biased MI methods. When event times were MNAR (shortest times missing), all methods over-estimated the median (bias > 1 for all methods), even when only 10% of acute GvHD times were missing.

In all scenarios, including auxiliary data in the PMM and NORM imputation models did not noticeably reduce bias, although it did generally reduce SE compared with PMMNOAUX and NORMNOAUX, respectively.

Although MED and MID estimates of the cumulative incidence were unbiased, estimates of the lower quartile time were biased. In addition, as expected, SE of the cumulative incidence was under-estimated for both methods (SE was the same as the complete data SE, regardless of the percentage of missing data). SE of the lower quartile time was also greatly under-estimated (and decreased as the percentage of missing data increased) using the MED method. Conversely, SE of the lower quartile time was the same or larger using the MID method, compared with other methods.

Study 2 results

In all missing data scenarios, results from Study 2 were consistent with those from Study 1. As per Study 1, when event times were MCAR, CCA estimates of cumulative incidence were unbiased as expected. As per the complete data results, CCA estimates of the median time of acute GvHD had a small amount of bias when event times were MCAR (given 10%, 30% and 50% missing times, bias for the median was: 0.20, 0.21, 0.19, respectively). Of the MI methods, only PMM and PMMNOAUX estimates of the cumulative incidence were unbiased. PMM and PMMNOAUX estimates of the median time of acute GvHD were also least biased (given 10%, 30% and 50% missing times, bias < 0.4 in each case) of all MI and FML methods. In addition, PMM and PMMNOAUX estimates of the SE were smaller than CCA estimates of SE, for both estimands. NORM, MICI and INTCCR (the FML method) estimates of the cumulative incidence were most

biased. Of all methods considered, RESNORM and NORM estimates of the median time of acute GvHD were most biased (see Section 5.5 for an explanation of these results).

As per Study 1, CCA estimates of the cumulative incidence greatly underestimated the true cumulative incidence, for event times both MAR and MNAR. When event times were MAR and missingness depended on event type and covariates (Figure 5.2), bias was small for all methods (including INTCCR, the FML method) except CCA and methods based on unrestricted linear imputation (NORM and NORMNOAUX). NORM and NORMNOAUX under-estimated the cumulative incidence at 100 days because the imputed times tended to be larger than with other imputation methods. PMM and PMMNOAUX performed particularly well, with negligible bias, even when 50% of event times were missing. Restricting the range of imputed values (RESNORM) reduced the bias of the cumulative incidence estimates compared to NORM and NORMNOAUX. As expected, SE for most methods increased with the volume of missing data, reflecting the additional uncertainty due to imputation. However, SE was under-estimated for MICI, RESNORM and, to some extent, INTCCR; SE was the same or smaller than the complete data SE in each scenario and was similar regardless of the volume of missing data. Reasons for this are considered in Section 5.5, following these results.

An additional analysis (to those considered in Study 1) explored the boundary comparison performed as part of the RESNORM imputation method. When the boundary comparison was performed in RESNORM a maximum of 200 times, a small number of imputed values were not within the boundaries. There was no change in the results when the boundary comparison was performed up to 500 times, but this did ensure that all imputed values were within the boundaries. When event times were MNAR (shortest times missing), all methods under-estimated the cumulative incidence except MICI, RESNORM and INTCCR methods. However, when event times were MNAR (longest times missing), the

cumulative incidence was over-estimated for all methods, as expected (see Appendices, Table A.3), including MICI, RESNORM and INTCCR methods (standardised bias was approximately 0.8 for each of these methods, for each percentage missing data). This suggests that the lack of bias for these methods when shortest times were missing was specific to that particular MNAR scenario (see Section 5.5 for an explanation of these results).

In CCA (for MAR and MNAR scenarios with 30% or 50% missing event times), less than 50% of patients (of those with completely observed event times) experienced acute GvHD. Therefore, the median time to acute GvHD could not be estimated using CCA in these scenarios. When data were MAR, bias was smaller for the median than was observed for the lower quartile in Study 1; MICI, PMM and INTCCR methods gave least biased estimates. As per Study 1, the size of the standardised bias was larger for estimates of the median than for cumulative incidence. NORM and NORMNOAUX over-estimated the median and restricting the range of imputed values did not improve median estimation. When data were MNAR, all methods over-estimated the median, with bias greater than 0.5 for all methods, even when only 10% of acute GvHD times were missing. For both estimation of the cumulative incidence and median, including the auxiliary variable (number of CB units) in the PMM and NORM imputation models did not reduce the bias and resulted in larger SE. This was the case even when the probability that acute GvHD times were missing depended on the auxiliary variable. An additional analysis in this study explored whether results were sensitive to the number of imputations performed; the same pattern of results was seen when the number of imputations was increased to 50 (see Appendices, Table A.3).

5.4.2. Results from simulation study 3: event times missing for multiple event types

Of all the MI methods considered in Studies 1 and 2, type 1 PMM performed best, with small bias and smaller SE than CCA, when imputing missing times under the MAR assumption. Therefore, Type 1 PMM was applied in Study 3, when event times were missing for multiple event types. Table 5.5 shows the standardised bias and average model-based SE of estimates of cumulative incidence and percentiles of event times, for MAR and MNAR MDMs. Results are consistent with those for the previous simulation studies (in which only acute GvHD times were imputed). Using type 1 PMM, estimates of cumulative incidence and percentiles of event times for acute GvHD, chronic GvHD and relapse have small bias when event times are MAR. In the MNAR scenario, bias was greater for estimates of the percentiles of event times than for cumulative incidence.

Table 5.5. Standardised bias (StdBias) and average model-based SE (ModSE) for event times (a) MAR (dependent on event type, the number of CB units received and time to myeloid engraftment), (b) shortest times MNAR, after MI using type 1 PMM.

Estimand	**Cumulative incidence (%)[1]**						**Percentile of event times (days)[2]**					
Event type (true result)	**Acute GvHD (55.56)**		**Chronic GvHD (19.38)**		**Relapse (16.89)**		**Acute GvHD (26.00)**		**Chronic GvHD (173.00)**		**Relapse (132.00)**	
Missing data mechanism	**Std Bias**	**Mod SE**	**Std Bias**	**Mod SE**	**Std Bias**	**Mod SE**	**Std Bias**	**Mod SE**	**Std Bias**	**Mod SE**	**Std Bias**	**Mod SE**
Complete data	-0.04	2.39	0.08	2.02	0.14	1.87	-0.26	1.67	0.12	17.46	0.11	36.82
MAR (event type + covars)	-0.36	2.47	0.20	2.05	0.21	1.91	-0.54	1.98	-0.46	22.73	-0.20	37.10
MNAR (shortest times missing)	-1.31	2.54	-0.57	2.02	-0.24	1.91	3.33	1.85	1.42	20.99	1.22	39.03

[1] At 100 days for acute GvHD and one year for chronic GvHD and relapse

[2] Lower quartile for acute GvHD and 10th percentile for chronic GvHD and relapse

Monte Carlo SE for bias was <0.1 for all estimates of cumulative incidence and lower quartile time of acute GvHD, and <1.2 for all estimates of the 10th percentile time of chronic GvHD and relapse.

5.5. Discussion

In this chapter, simulation studies to assess methods for handling missing event times were described, when estimating cumulative incidence and percentiles of event times in a competing risks framework. The aim of the studies was to evaluate the extent to which data boundaries, the data distribution and the analysis model should be represented in the imputation model when imputing event times. In addition, MI strategies were compared with a suitable FML method. Methods were compared for different DGMs, different missing data scenarios, various percentages of missing data, and with event times missing for a single or multiple event types.

Estimates based on complete simulated data (*i.e.* without any missing event times) were slightly biased for estimates of the percentiles of event times. The distribution of percentile estimates (based on the simulated event times) was slightly skewed (the median estimates of percentiles of event times using complete data exactly matched the true value in each study). This may be because event times were rounded to the nearest day (see study results in Section 5.3.2).

In the NHS CBB dataset, the missingness of event times depends on the type of event experienced. Hence, CCA will be biased because missingness of event times depends on the analysis outcome. However, even if event times were MCAR, or MAR but missingness did not depend on the analysis outcome, given the covariates in the analysis model (in which case, CCA estimates would be unbiased), my results suggest that MI can be used to improve precision.

My simulation study results suggest that estimates of cumulative incidence and percentiles of event times are sensitive to imputation model misspecification. Sampling from a set of observed times without reliance on a specific parametric

distribution (type 1 PMM and, in the studies with event times missing for a single event type, the Delord and Genin method, MICI), resulted in the least biased estimates, on average, when event times were MAR. In PMM, missing values are replaced by sampling at random from a donor pool of patients (with observed values) who are 'similar' to the subject with missing data. In MICI, sampling is from the set of event times that lie within specified boundaries, where the sampling probability is determined by the current estimate of the cumulative incidence function. However, SE was under-estimated using MICI. MICI also has the drawback of being a univariate imputation method *i.e.* other variables with missing data cannot be imputed in the same imputation model (175).

Restricting the range of imputed values generally reduced the bias for estimates of cumulative incidence, though not for estimates of percentiles of event times. MICI and the restricted linear imputation model (RESNORM) both truncate values during the imputation step and this seems to lead to under-estimates of the SE for both methods. Results from MNAR studies suggest that these methods tend to result in smaller imputed times, perhaps because larger times, which may be outside the boundaries, are discarded. For this reason, I do not recommend the use of restricted range methods. My study results do not completely agree with von Hippel's advice to transform variables with a skewed distribution when estimating percentiles (here, via log-transformation); although estimates were improved in comparison with an untransformed linear regression imputation model, log-transformation resulted in some bias.

Contrary to other publications in this field (97, 121), in my simulation studies, imputing on the original scale led to bias. However, this may be due to imputation model misspecification, more specifically, due to the constant variance assumption of the linear regression imputation model (87). In my DGMs, guided by the distribution of event times for each event type in the NHS CBB cohort, the standard deviation of simulated times of acute GvHD and graft

failure times was approximately 30 days in all three studies. However, the standard deviation of simulated times of death was approximately 300 days in studies 1 and 3, and 85 days in study 2. Including all event types in the imputation models resulted in a model with estimated variance of times of acute GvHD that was greater than desired. This explains the tendency of the linear imputation method towards larger times compared with other imputation methods. Correcting this by limiting the linear imputation model to the subset of patients who experienced acute GvHD, resulted in estimates with negligible bias in MAR scenarios even when 50% data were missing (Appendices, Table A.3, NORMSUBGP method).

The difference in the distribution of simulated death times (compared with the distributions of acute GvHD and graft failure times) may also explain the poor performance of the RESNORM, MICI and INTCCR methods when event times were MCAR. In the MCAR scenarios, event times were incompletely observed for all event types (*i.e.* there was an equal probability of any acute GvHD, graft failure or death time being missing). For the RESNORM, MICI and INTCCR methods, I used the same interval boundaries, regardless of event type (in Study 1, minimum = 0 and maximum = min(119 days, patient's death time); in Study 2, minimum = 0 and maximum = 100 days). However, in each study, the specified maximum boundary under-estimated the true maximum death time (based on the distribution of simulated death times). Hence, the number of deaths during the time period of interest (approximately the first 100 days post-transplant) may have been over-estimated. Since the cumulative incidence of acute GvHD depends on the failure rates of all event types (see Chapter 2, Section 2.2.2), this could lead to biased estimates. My results suggest that MI and FML methods that require the specification of interval boundaries are sensitive to the choice of boundaries.

Given these findings, I recommend exploring the distribution of event times for different sub-groups of patients prior to imputation. In theory, for any sub-

group for which the distribution of event times is substantially different, a separate imputation model could be fit. However, this may not be practical in studies with large numbers of variables included in the imputation models or many incomplete variables to be imputed. Overall, I recommend PMM, because it is more robust to model misspecification.

In Study 2, due to the DGM design, there was no association between the auxiliary variable and event times. This explains why inclusion of the auxiliary variable in the imputation model did not reduce bias and increased SE, even when missingness of the event time depended on the auxiliary variable (157). All variables, including outcome variables, included in the analysis model must be included in the imputation model. Only auxiliary variables that are predictive of the incomplete variable, or predictive of the incomplete variable and its missingness, should be included in the imputation model. Variables that are only predictive of missingness will not reduce bias and could increase the SE. However, (since lack of association is difficult to verify in practice) I recommend including all variables that are thought to be predictive of the incomplete variable in the imputation model (based on a detailed exploration of the potential missingness mechanism for each incomplete variable in the analysis model, as well as subject-matter knowledge).

When data were MNAR, imputation resulted in biased estimates even when only a small percentage of data were missing. Generally, FCS MI methods that assume MAR are not recommended when data are MNAR, and my results support this.

Performance of the FML method considered (INTCCR) was similar that for restricted range MI methods (RESNORM and MICI), and worse than the best performing MI method (PMM). There was no evidence of any advantage in using a FML method rather than FCS MI to handle missing event times. FCS MI offers many advantages, notably the ability to accommodate a mixture of

continuous, binary, and categorical incomplete variables; the option to include auxiliary data during imputation; flexibility when choosing the analysis model; and accommodation of a mixture of exactly observed and missing times. There is only one standard software application of FML methods that allows calculation of the non-parametric estimate of the cumulative incidence function. This is the package 'MLEcens'(179) available in R. However, it only allows for two competing events and does not provide standard error (SE) estimation. In contrast, MI can be easily implemented in many statistical packages including SAS (SAS Institute Inc., Cary, NC, USA), Stata (StataCorp, College Station, TX, USA) and R *software.*

In the simulation studies described here, data were generated by sampling either from real data, or from parametric distributions (the advantages and disadvantages of each type of DGM were summarised in Chapter 2, Table 2.1). In my simulation studies, results using each type of DGM were similar. This suggests that the performance of the different methods was not related to the DGM used. I recommend using DGMs based on both real data and simple parametric distributions, to provide reassurance that method performance is not sensitive to the choice of DGM.

It is particularly important to handle missing times of acute GvHD, chronic GvHD and relapse appropriately because these events are commonly-reported outcomes in HSC transplant studies. Little and Rubin (82) note that CCA will always be biased if missingness depends on the analysis outcome. My study results support this statement. However, CCA and naïve imputation methods are frequently used to handle missing times in HSC studies (45, 48). I recommend using PMM to impute missing event times when the analysis includes estimation of the cumulative incidence and percentiles of event times, and event times are assumed to be MAR (conditional on the observed data). It is straight-forward to implement PMM in many software packages although it should be noted that

SAS software only implements type 2 PMM (92) which has inferior performance to type 1 PMM (97, 98).

The simulation studies described in this chapter considered a simple analysis model (non-parametric estimation of the cumulative incidence function). In the next chapter, I investigate whether results also apply to a MSM including an intermediate event and covariates.

CHAPTER 6. IDENTIFYING THE BEST MULTIPLE IMPUTATION METHOD FOR HANDLING MISSING EVENT TIMES IN A MULTI-STATE MODEL

6.1. Introduction

In Chapter 5, I explored methods for handling missing event times in a competing risks analysis using simulation studies. For the MI methods I applied, I found that sampling from a set of observed times, without reliance on a specific parametric distribution, resulted in the least biased estimates when event times were MAR. Restricting the range of imputed values generally reduced the bias for estimates of cumulative incidence but led to under-estimation of SE. Overall, I found the best method was type 1 PMM, because of its flexibility and its robustness to model misspecification.

In this chapter, using an extensive simulation study, I consider missing event times in a multi-state setting. I compare the best MI methods (identified using a competing risks analysis model, see Chapter 5), when the analysis model is a MSM. In this chapter, I only consider MI strategies. In my competing risks analysis, I found no evidence of an advantage in using FML methods rather than MI to handle missing event times. Furthermore, in a review of FML methods for handling interval-censored event times in MSMs, Machado *et al.* (152) found that none of the available methods performed well when censoring intervals were wide, relative to the change in the hazards. In my dataset, intervals are wide

relative to the observed event times. Hence, the FML methods developed to date are unlikely to perform well when applied to my data. In contrast, MI offers many advantages, notably the option to include auxiliary data during imputation, flexibility when choosing the analysis model and accommodation of a mixture of exactly observed and missing times.

6.2. Description of the simulation study comparing imputation methods for handling missing event times in a multi-state model

6.2.1. Aim of the simulation study

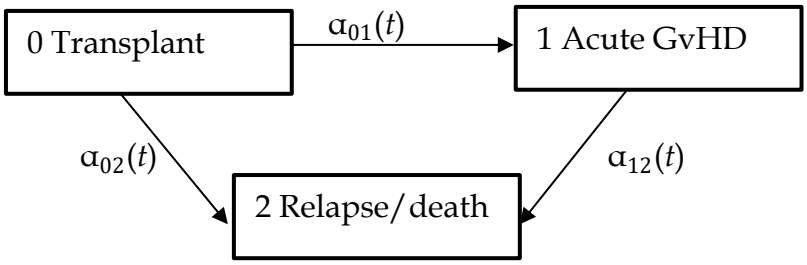
The aim of the study was to evaluate the bias and precision of estimates from an MSM in different missing data scenarios, comparing the best previously identified MI methods for handling missing event times.

6.2.2. Multi-state model structure

The MSM used in this study is based on the MSM suggested by Keiding *et al.* to describe the events experienced after HSC transplant (180). My version reduces the MSM of Keiding *et al.* to the simplest three-state MSM, that is, a uni-directional "illness-death" model (73), as described in Chapter 2, Section 2.2.3.

Here, transplant is the initial state for all patients, acute GvHD the single intermediate state and relapse/death the single absorbing state (Figure 6.1).

Figure 6.1. The uni-directional illness-death model for events following HSC transplantation. $\alpha_{hj}(t)$ represents the transition intensity when moving from state h to state j



Consistent with the NHS CBB data, I make the following assumptions:

- The time of transplant is known, *i.e.* the time origin is observed for all patients.
- Subsequent events may be unobserved, *i.e.* there may be right-censoring, but I assume that the censoring distribution is independent of the transition time and state occupied and hence, standard methods can be applied (see Chapter 2, Section 2.2).
- For patients who experience both acute GvHD and relapse or death, I assume that acute GvHD always occurs before relapse or death.

The events included in the MSM differ from those considered in the competing risks model for time to acute GvHD, described in Chapter 5, in two ways:

(i) Relapse was not included in the competing risks model because it is not considered a competing risk for acute GvHD, according to the clinical literature (45, 59). However, events other than death are unlikely after relapse (for patients in the NHS CBB dataset, engraftment and GvHD never occurred after relapse, see Chapter 4, Section 4.4.1). In addition, disease-free survival is an outcome of clinical interest in many HSC transplant studies (45, 181, 182). Relapse is also of interest from a missing data perspective, because, like acute GvHD times, a large proportion of relapse times are missing. Hence, in my MSM analysis, I treat the

occurrence of either relapse or death as the absorbing state (see Chapter 2, Section 2.2.3).

(ii) Graft failure prior to acute GvHD was included in the competing risks model but is not included in the MSM. In the NHS CBB dataset, graft failure occurs before or after acute GvHD and/or relapse. In the MSM used here, for simplicity, I only consider one intermediate state (acute GvHD). Also, I assume that relapse is as an absorbing state (*i.e.* that a transition from relapse is not possible).

Again, for simplicity, the Markov property was assumed, hence time t is measured from entry into the initial state for all transitions (see Chapter 2, Section 2.2.3). I also assumed a PH structure for each transition intensity (see Chapter 2, Section 2.2.1). This means that the transition intensity, $\alpha_{hj}(t)$, at time t since transplant, when moving from state h to state j, is defined for the i^{th} subject with time-fixed covariates z_i as follows:

$$\alpha_{hj}(t) = \alpha^0_{hj}(t) \exp(\boldsymbol{\beta}'_{hj} z_i) \text{ for all } i, h, j$$

where $\alpha^0_{hj}(t)$ represents the baseline intensity at time t.

6.2.3. Data-generating mechanism

Data were generated using the method for simulating competing risks data described by Beyersmann *et al.* (77). MSM data can be generated by applying this method to a sequence of competing risks experiments. The method consists of four steps, which are performed for each set of r transitions representing a competing risks experiment ($r \geq 1$, with $r = 1$ representing the simplest model with only one absorbing state). These steps are briefly summarised below; further details on the application of these steps in my simulation study are described in Section 6.2.8:

1. The transition intensity function $\alpha_{hj}(t)$ is defined for each transition as a function of time t.

2. Event times are generated from the all-cause intensity function (the sum of all transition intensities), $\sum_r \alpha_{hj}(t)$.
3. A binomial experiment (or multinomial if $r > 2$) is performed to determine the transition associated with each event time, with probability $\frac{\alpha_{hj}(t)}{\sum_r \alpha_{hj}(t)}$ of each of the r transitions occurring.
4. Censoring times are generated independently, and as desired.

6.2.4. Description of complete case analysis of the NHS CBB cohort

A CCA of the NHS CBB cohort was used to inform the choice of (a) covariates and covariate parameters and (b) baseline intensity functions, to be used in the generation of simulated data. Here, 'complete case' means that patients in the NHS CBB cohort were included in the analysis if times of all the events of interest were observed, regardless of whether covariate data were observed or not.

Covariates and covariate parameters

A separate Cox model (51) was fitted for each transition intensity using the 'survival' R package (183). All clinically relevant patient, donor and transplant characteristics at time of transplant were included as covariates in each transition intensity model, namely: number of CB units transplanted (single or double); patient age (in years, assuming a linear association); disease type (acute leukaemia, other blood cancer, non-malignant disorder); disease status at time of transplant (in remission, in relapse, other); conditioning regimen (intensive or not); sex and CMV match between donor(s) and recipient; number of HLA mismatches between donor(s) and recipient (well-matched: 0 or 1 mismatches; or not: 2 or more mismatches); TNC dose at infusion ($\times 10^7$/kg patient weight: low, <3.0; medium, 3.0-5.0; high, >5.0). Patients in relapse at time of transplant were assumed to be relapse-free immediately post-transplant *i.e.* the transition from

state 0 to state 2 was assumed to occur at some time $t > 0$. The covariates most strongly associated with each outcome were decided using forward selection, based on the likelihood ratio test (159).

Baseline intensity functions

In practice, the Cox model allows covariate effects to be estimated without specification of the baseline intensity function, $\alpha_{hj}^{0}(t)$. However, when simulating data using the method described by Beyersmann *et al.*, it is necessary to define $\alpha_{hj}^{0}(t)$ explicitly. In this study, a parametric, Weibull distribution (184) was used to generate the baseline intensity function for each transition intensity model. The intensity function for a Weibull distribution $\alpha(t;k,\lambda)$, with shape parameter k and scale parameter λ, is defined as:

$$\alpha(t;k,\lambda) = \left(\frac{k}{\lambda}\right)\left(\frac{t}{\lambda}\right)^{k-1}$$

The Weibull distribution was chosen for two reasons:

(i) It is compatible with the PH assumption.

(ii) Its flexibility: depending on the values of the shape and scale parameters, transition intensities that are increasing, decreasing or constant over time can be fitted using a Weibull model. Thus, it is possible to choose a Weibull distribution that is consistent with the real NHS CBB data.

After identifying the covariates to be used in the simulation study (based on the CCA described above), the Weibull parameters to be used in the simulation study were determined as follows. A Weibull distribution was fitted to event times, measured in days, for each transition, for the set of patients with reference values of the chosen covariates, using the 'fitdistrplus' R package (185). This approach assumes a direct relationship between event times and the transition intensity and ignores the competing risks and conditional time elements (see Section 6.2.2) of the MSM used in my study. Therefore, an alternative method was also considered. In this method, a Weibull PH transition intensity model was fitted for each transition, with the chosen covariates in each model.

6.2.5. Results of complete case analysis of the NHS CBB cohort

There were 358 patients in the NHS CBB dataset with both acute GvHD and relapse/death times observed (Table 6.1).

Table 6.1. Number of patients experiencing acute GvHD and/or relapse/death, by type of missing event time.

Type of missing event time	**Patients (N)**	**Type of event experienced**		
		Acute GvHD	**Relapse/death without acute GvHD**	**Relapse/death after acute GvHD**
None	358	181	93	85
Acute GvHD	52	52	0	12
Relapse	17	3	14	3
Acute GvHD and relapse	5	5	0	5
All patients	**432**	**241**	**107**	**105**

Covariates and covariate parameters

A detailed interpretation of covariate associations with transition intensities will be provided as part of the full analysis of the NHS CBB dataset in Chapter 7. Here, to inform covariate selection in the simulation study, the covariates most strongly associated (*i.e.* with a p-value < 0.1) with each transition intensity are summarised in Table 6.2 overleaf.

Table 6.2. Predictors in the transition intensity models for the complete case NHS CBB dataset.

Covariate (reference value)	Total (N)	Events (N)	Hazard ratio (95% CI)		P*
Transplant to acute GvHD	358	181			
Disease status at transplant (partial/complete remission)	141	86	1.00	(-)	0.016
Relapse	23	9	0.67	(0.34-1.34)	
Other	110	55	0.80	(0.57-1.12)	
Not reported	84	31	0.53	(0.35-0.80)	
Transplant to relapse/death without acute GvHD	358	93			
Disease status at transplant (partial/complete remission)	141	31	1.00	(-)	0.004
Relapse	23	12	3.08	(1.57-6.04)	
Other	110	21	0.72	(0.41-1.25)	
Not reported	84	29	1.08	(0.65-1.80)	
Donor-recipient CMV status match (-/-)	95	21	1.00	(-)	0.035
-/+	95	29	1.40	(0.80-2.45)	
+/-	66	9	0.54	(0.25-1.18)	
+/+	62	20	1.14	(0.98-4.78)	
Not reported	40	14	1.82	(0.61-2.11)	
Acute GvHD to relapse/death	181	85			
Disease status at transplant (partial/complete remission)	86	44	1.00	(-)	0.005
Relapse	9	8	2.24	(0.92-5.44)	
Other	55	26	1.34	(0.79-2.26)	
Not reported	31	7	0.42	(0.18-0.98)	
Number of donor-recipient HLA mismatches (well-matched: 0/1)	64	21	1.00	(-)	0.027
Poorly-matched: 2 or more	95	52	1.88	(0.99-3.58)	
Not reported	22	12	2.99	(1.32-6.77)	
TNC dose at infusion ($\times 10^7$/kg) (Low: < 3.0)	40	25	1.00	(-)	0.028
Medium: 3.0-5.0	54	32	1.01	(0.59-1.72)	
High: > 5.0	41	16	0.83	(0.42-1.65)	
Not reported	46	12	0.40	(0.20-0.82)	
Recipient age (per 10-year increase)	181	85	1.17	(1.01-1.34)	0.035
Number of CB units transplanted (single cord)	108	51	1.00	(-)	0.095
Double cord	73	34	0.57	(0.30-1.09)	

*P-values are for models including all other covariates shown for each transition.

For all transitions, disease status at transplant was most strongly associated with the hazard rate. After adjusting for disease status at transplant, there was no

evidence of an association between any other covariate and the hazard of acute GvHD. However, in addition to disease status at transplant, donor-recipient CMV status match was associated with the hazard of relapse/death without acute GvHD. The number of donor-recipient HLA mismatches, recipient age, dose at infusion and number of CB units transplanted were associated with the hazard of relapse/death after acute GvHD.

In the simulation study, for simplicity, only two of the covariates identified above were included in the transition intensity models. These were both included as binary variables (disease status at transplant was simplified to 'in relapse at time of transplant or not'; assuming patients are not in relapse immediately post-transplant). The covariate in models of transition from transplant to acute GvHD, and from transplant to relapse/death, was disease status at transplant. The covariates in the model from acute GvHD to relapse/death were disease status at transplant and the number of CB units received. These covariates were chosen because they were associated with the times to events (disease status at transplant) or were associated with the times to events and the probability that event times were missing (number of CB units received) (see Chapter 4, Section 4.5). The covariates were included in both DGMs and the analysis model. No unmeasured variables were considered in the simulation study, so that any bias in model estimates could be attributed to the MI method.

Baseline intensity functions

The baseline intensity functions used in the simulation study are defined below, for each transition:

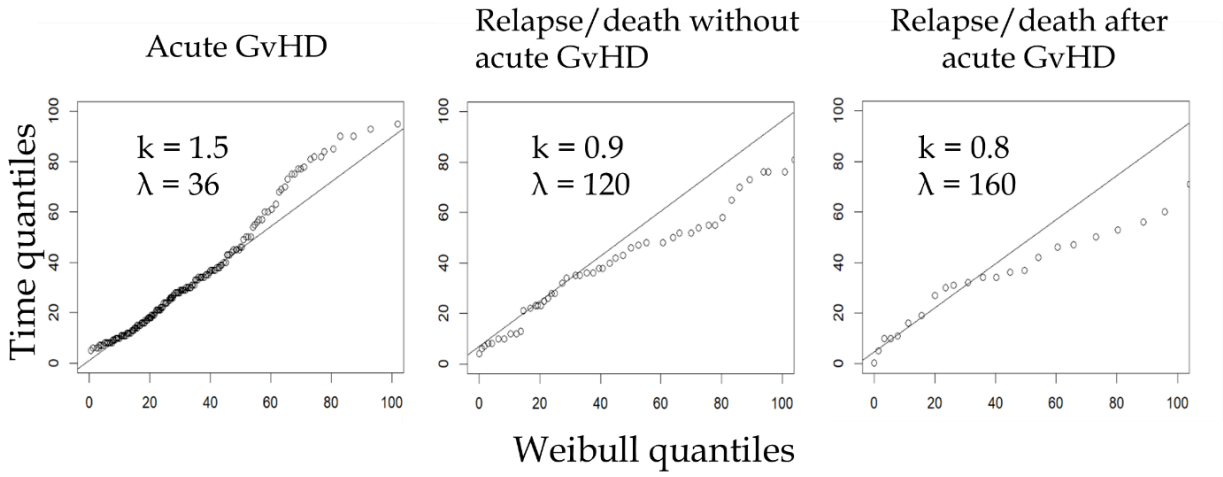
$$\alpha_{01}^{0}(t) = \left(\frac{1.5}{36}\right)\left(\frac{t}{36}\right)^{0.5} \qquad \alpha_{02}^{0}(t) = \left(\frac{0.9}{120}\right)\left(\frac{t}{120}\right)^{-0.1} \qquad \alpha_{12}^{0}(t) = \left(\frac{0.8}{160}\right)\left(\frac{t}{160}\right)^{-0.2}$$

The Weibull parameters in the models above were from Weibull distributions fitted to event times for each transition, for patients in the complete case NHS CBB cohort. The Q-Q plots in Figure 6.2 compare the distribution of event times

generated using the chosen Weibull models with the distribution of event times for patients in the complete case NHS CBB cohort. Since the Weibull models represent baseline intensity functions, only NHS CBB patients who had reference (baseline) values of the proposed covariates were included in the plots. That is, patients not in relapse at time of transplant and, for the transition from acute GvHD to relapse/death, patients not in relapse at time of transplant who received a single cord transplant. As shown in Figure 6.2, the chosen Weibull distribution is a good fit for the real data for times below the median point of each Weibull distribution. However, there is some lack of fit above the median for each transition, particularly for the two distributions of relapse/death times, which may be due to the small numbers of observed times above the median.

Figure 6.2. Q-Q plots of event time distributions, comparing NHS CBB data with theoretical Weibull distributions.

Weibull shape parameter (k) and scale parameter (λ) shown in each case.



The alternative method considered for choosing Weibull parameter estimates (*i.e.* fitting Weibull PH transition intensity models for each transition), resulted in very unrealistic estimates, particularly for scale parameters. Weibull parameter estimates were as follows.

- Transplant to acute GvHD: $k = 0.4$ and $\lambda = 760$ (equivalent to a distribution with median 290 days, IQR 1500 days).
- Transplant to relapse/death: $k = 0.5$ and $\lambda = 2600$ (equivalent to a distribution with median 1300 days, IQR 4700 days).

- Acute GvHD to relapse/death, $k = 0.06$ and $\lambda = 0.00004$ (equivalent to a distribution with median 0 days, IQR 0.02 days).

Again, this poor fit may be due to small numbers of patients/events in the complete case NHS CBB cohort.

6.2.6. Data-generating mechanism: application of Beyersmann's method

The MSM used in the simulation study consisted of a sequence of two competing risks experiments. Transitions from transplant (α_{01} and α_{02}) comprised the first competing risks experiment and the transition from acute GvHD to relapse/death (α_{12}) comprised the second experiment (in this case, there was only a single absorbing state). Informed by the results of the complete case NHS CBB analysis, the application of Beyersmann's method used in this study is described below. For each patient i (note that, for clarity, the subscript i is suppressed in the definitions overleaf), values of the covariates used in the DGM were generated by sampling from Bernoulli distributions with probability 0.2 of relapse at time of transplant, and probability 0.45 of a double cord transplant. Values for the two covariates were generated independently *i.e.* assuming no correlation between the two covariates. R code to generate data as per my DGM is included in the Appendices (Section A.6).

First experiment

1. The transition intensity functions $\alpha_{0j}(t)$ for the two transitions from transplant were defined as:

   $\alpha_{01}(t) = \left(\frac{1.5}{36}\right)\left(\frac{t}{36}\right)^{0.5}\exp\{-0.8x_1\}$ and $\alpha_{02}(t) = \left(\frac{0.9}{120}\right)\left(\frac{t}{120}\right)^{-0.1}\exp\{1.2x_1\}$

   where t is the time in days since transplant and $x_1 = 1$ for a patient in relapse at time of transplant and 0 otherwise.

2. Event times were generated from the all-cause intensity function, in this case, $\alpha_{01}(t) + \alpha_{02}(t)$. Event times can be generated easily via cumulative hazard

inversion if the intensity function has a standard distribution (184). However, here the all-cause intensity function is the sum of two Weibull distributions, which cannot be inverted using analytical methods. Crowther and Lambert (186) adapted the cumulative hazard inversion method to accommodate such functions. In this study, their methods were applied using the "simsurv" R package (187) to generate event times for the first competing risks experiment. To distinguish between event times from the first and second experiment, the simulated event time for the first experiment will hereafter be referred to as t_{0j}, where j denotes the state entered from transplant.

3. A binomial experiment was performed to determine the state j entered from transplant state 0 associated with each simulated event time t_{0j}, with the probability of entering the acute GvHD state (versus relapse/death) equal to $\frac{\alpha_{01}(t_{0j})}{\alpha_{01}(t_{0j}) + \alpha_{02}(t_{0j})}$.

4. The censoring mechanism was applied after event times had been generated from both experiments (because the second experiment used the uncensored times from the first experiment), so the censoring mechanism is not described here.

Second experiment

1. The transition intensity function $\alpha_{12}(t)$ was defined as:

   $$\alpha_{12}(t) = \left(\frac{0.8}{160}\right)\left(\frac{t}{160}\right)^{-0.2}\exp\{1.2x_1 - x_2\}$$

   where t and x_1 are defined as before and $x_2 = 1$ for a double cord transplant and 0 otherwise.

2. The second experiment consisted of just one possible transition; hence the all-cause intensity was $\alpha_{12}(t)$. The time of relapse/death, t_{12}, was required to be greater than the simulated time of acute GvHD, t_{01}. Therefore, relapse/death

times were generated from the conditional survival function, using cumulative hazard inversion. Cumulative hazard inversion proceeds as follows:

(a) The conditional survival function is defined as $S(t_{12} \mid t_{01}) = \frac{S(t_{12})}{S(t_{01})}$

where $S(t) = \exp\left[-\left(\frac{t}{160}\right)^{0.8} \exp\{1.2x_1 - x_2\}\right]$ with x_1 and x_2 defined as before, based on the standard result for a Weibull distribution (184).

(b) A value of u is drawn from a Uniform(0,1) distribution.

(c) Using the result that the conditional survival function is uniformly distributed across the range (0,1) gives $u = \frac{S(t_{12})}{S(t_{01})}$; substituting the expression for $S(t)$ and rearranging allows direct calculation of each t_{12} as:

$$t_{12} = 160\left\{\frac{-\log\left[u \exp\left[-\left(\frac{t_{01}}{160}\right)^{0.8} \exp\{1.2x_1 - x_2\}\right]\right]}{\exp\{1.2x_1 - x_2\}}\right\}^{1/0.8}$$

3. There was no need to determine the transition associated with each survival time in the second experiment because the only possible transition from acute GvHD was to relapse/death.

4. In this study, censoring was at random. Censoring times between one and five years post-transplant were generated for each subject by drawing from a Uniform(365, 1826) distribution.

Of the data simulated in this study, only times of relapse/death after acute GvHD were censored. This was because all simulated times of acute GvHD, and relapse/death without acute GvHD, were within one year of transplant. The simulated distribution of acute GvHD times (median 21 days, IQR 26 days) is consistent with both the standard clinical definition of acute GvHD (acute GvHD

occurs within the first 100 days post-transplant), and the distribution of observed times of acute GvHD in the complete case NHS CBB dataset (median 27 days, IQR 25 days).

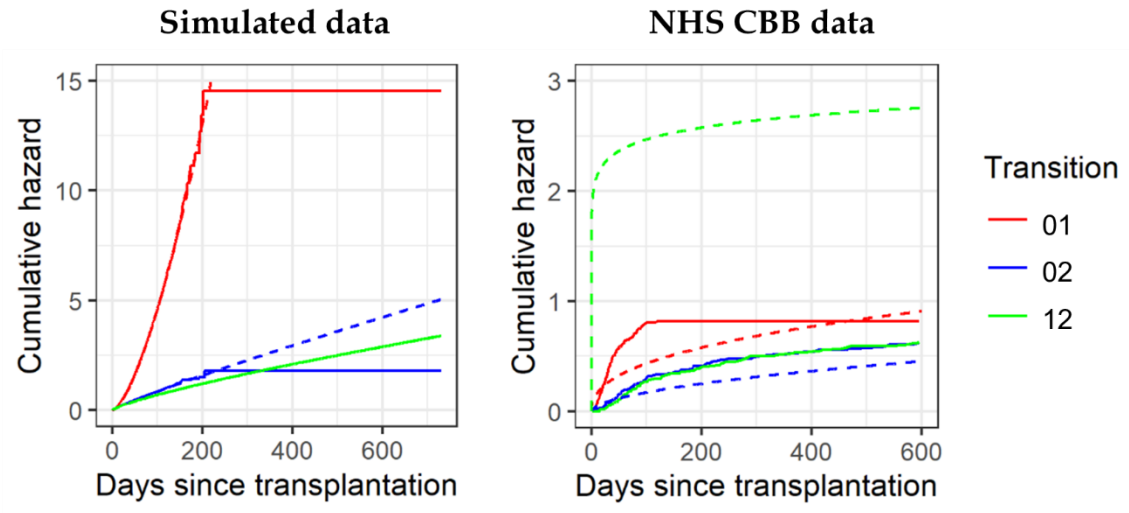
Comparing patients who experienced acute GvHD prior to relapse/death with those who did not (in all 1000 simulated datasets), the distribution of simulated (uncensored) relapse/death times were very different (after acute GvHD, median 200 days, IQR 386 days; without acute GvHD, median 14 days, IQR 22 days). A direct comparison with the complete case NHS CBB dataset is not possible, because many relapse/death times (in the real dataset) were censored. However, comparing patients who experienced acute GvHD prior to relapse/death with those who did not (in the NHS CBB dataset), the distribution of real (censored) relapse/death times appear more similar (after acute GvHD, median 77 days, IQR 205 days; without acute GvHD, median 56 days, IQR 119 days).

However, some aspects of the simulated and real relapse/death times are comparable. In both cases, relapse/death times are longer for patients experiencing acute GvHD than for those who experience relapse/death without acute GvHD. Furthermore, for patients in the complete case NHS CBB dataset who experienced relapse/death without acute GvHD, 86/93, 92%, were within one year of transplant.

In addition, Cox model estimates of the cumulative baseline hazard function (CBH) for each transition (shown as solid lines in the plots in Figure 6.3, overleaf) have a similar shape for both simulated and real data (shown on the left- and right-hand plots, respectively), at least within the range of transition times. However, as mentioned previously, the distribution of transition times differs between the simulated and real data. This can be seen in Figure 6.3 by comparing the time-point at which there is no further change in the CBH for each transition (the CBH becomes horizontal) *e.g.* for the transition from transplant to

acute GvHD, this occurs at ~ 200 days for the simulated data vs. 100 days for the real data.

Figure 6.3. Estimated cumulative baseline hazard function for each transition for the simulated and complete case NHS CBB datasets, fitting a Cox model (solid lines) and a Weibull model (dashed lines).



State indicators: 0 = transplanted; 1 = acute GvHD; 2 = relapse/death
Note that for the simulated data(left-hand plot), for the transition from acute GvHD to relapse/death, the solid line exactly overlays the dotted line.

In contrast to the Cox model estimates, there are substantial differences in the fit of Weibull model estimates of the CBHs between the simulated and real data. Taking the Cox CBHs as the 'true' CBHs (since these are estimated non-parametrically from the data), Figure 6.3 shows that the Weibull CBHs are a very good fit for the simulated data within the range of the simulated transition times. In addition, as expected, the parameter estimates from Weibull models fitted to the simulated data are the same as the Weibull parameters used to generate the simulated data.

However, it is clear that the Weibull CBHs are not a good fit for the real data. In particular, the Weibull CBHs appear to have a good fit to the real data only in certain time periods. For transitions from transplant, the Weibull CBHs are a good fit to the true CBHs only in the very early period post-transplant (all observed transition times are below the median point of the Weibull CBHs). For

the transition from acute GvHD to relapse/death, the shape of the Weibull CBH fits the shape of the true CBH only at later times post-transplant (all observed transition times are above the median of the Weibull CBH). The Weibull model greatly over-estimates the CBH in the first few days post-transplant for this transition. The lack of fit in the real data may be due to the small number of transitions relative to the range of observed transition times *e.g.* only 46 patients with reference values of covariates experienced relapse/death after acute GvHD, and the range of transition times was 14 - 1711 days post-transplant.

In summary, the difference between the simulated and real distributions of relapse/death times may be for the reasons discussed previously. Namely, the lack of fit between the Weibull models and the real data, and the assumption of a direct relationship between event times and the transition intensity functions.

6.2.7. Methods for simulating missing data

MDMs were chosen so that approximately 30% of event times were missing. This reflects the percentage of missing times in the real NHS CBB cohort. Firstly, a MCAR MDM was considered by setting a random 30% of event times to missing, regardless of event type. Next, MAR and MNAR MDMs were considered in three scenarios: (i) only times of acute GvHD missing, (ii) only times of relapse/death missing, (iii) times of both acute GvHD and relapse/death missing. Although event times MAR is the most likely MDM for the NHS CBB cohort (see Chapter 4, Section 4.5), MNAR MDMs were also considered, to assess the impact on bias and precision if MI was used in this scenario.

As discussed above, the DGM used in this study means that the distribution of simulated times to relapse/death differs substantially between patients who experience acute GvHD prior to relapse/death, and those who do not. Therefore, in scenario (ii), two different MAR MDMs were considered. Firstly,

dependent on acute GvHD, such that relapse/death times were either missing for patients who experienced acute GvHD, or for patients who did not experience acute GvHD (but not both in the same MDM). Secondly, independent of acute GvHD, such that relapse/death times were missing, regardless of whether acute GvHD was experienced or not. This was done to assess whether the heterogeneity of the two distributions of times to relapse/death would affect the performance of the chosen MI methods.

In scenario (iii), missingness of relapse/death times was not dependent on whether acute GvHD was experienced or not. In this scenario, different combinations of acute GvHD and relapse/death event times MAR and MNAR were considered. This was to assess whether the type of missingness was more important for transitions to an absorbing state than to an intermediate state. In addition, two different MNAR MDMs were considered: smallest event times missing, and largest event times missing. This was to assess how sensitive the results were to the choice of MNAR mechanism.

Note that, missingness of an event time for one event type may lead to missing times for more than one transition, because the event time may be associated with multiple transitions (Table 6.3).

Table 6.3. Missing transition times for each type of missing event time

Type of missing event time	**Transition intensity model (transition time missing or not)**		
	Transplant to AGvHD	**Transplant to R/death w/o AGvHD**	**AGvHD to R/death**
AGvHD	Missing	Missing*	Missing
R/death w/o AGvHD	Missing*	Missing	n/a
R/death after AGvHD	Not missing	Not missing*	Missing

AGvHD, acute GvHD; R/death, relapse/death; w/o, without
* Censored at time of competing event

Complete data MSM methods require the time of entry into both states *h* and *j* to be known for the transition from state *h* to state *j*. It would be possible to construct MDMs such that missingness was transition-specific. However, this

would not be realistic because it would mean that, for example, the time of acute GvHD was missing for one transition but observed for another. Therefore, in all MDMs, missingness is specified for event times, rather than transition times.

MAR and MNAR MDMs for each of the three scenarios discussed previously are described in detail overleaf. In each scenario, j = 1, 2, 3 denotes the event experienced (j = 1: acute GvHD, j = 2: relapse/death without acute GvHD, j = 3: relapse/death after acute GvHD). For clarity in explaining the missingness structure, the events of relapse/death, without or after acute GvHD, are identified as two separate event types. For each subject i, π_{ij} denotes the probability that event times for the j^{th} event are missing; t_{ij} is the event time for subject i for the j^{th} event and x_{2i} = 1 for a double CB transplant and 0 otherwise. Also, $t_{j(p\%)}$ is the p^{th} percentile of event times for the j^{th} event, ordered from smallest to largest and $t_{RD(p\%)}$ is the p^{th} percentile of all times of relapse/death (regardless of whether acute GvHD was experienced or not), ordered from smallest to largest.

(i) Missing acute GvHD times

a) MAR: The probability that acute GvHD times are missing depends on the number of CB units transplanted; relapse/death times are fully observed:

$\pi_{i1} = 0.2\ (1 + x_{2i})$ and $\pi_{i2} = \pi_{i3} = 0$

b) MNAR: The shortest 30% of all acute GvHD times are missing; relapse/death times are fully observed:

$$\pi_{i1} = \begin{cases} 1 \text{ if } t_{i1} < t_{1(30\%)} \\ 0 \quad \text{otherwise} \end{cases} \quad \text{and} \quad \pi_{i2} = \pi_{i3} = 0$$

(ii) Missing relapse/death times

a) MAR: The probability that relapse/death times are missing depends on the number of CB units transplanted and whether acute GvHD was experienced or not; acute GvHD times are fully observed:

$\pi_{i1} = \pi_{i3} = 0$ and $\pi_{i2} = 0.5\ (1 - 0.8\ x_{2i})$

b) MAR: The probability that relapse/death times are missing depends on the number of CB units transplanted, regardless of whether acute GvHD was experienced; acute GvHD times are fully observed:

$\pi_{i1} = 0$ and $\pi_{i2} = \pi_{i3} = 0.5\ (1 - 0.8\ x_{2i})$

c) MNAR: The shortest 30% of all relapse/death times are missing, regardless of whether acute GvHD was experienced; acute GvHD times are fully observed:

$\pi_{i1} = 0$ and $\pi_{iI} = \begin{cases} 1 \text{ if } t_{iI} < t_{RD(30\%)} \\ 0 \text{ otherwise} \end{cases}$ $l = 2, 3$

(iii) Missing acute GvHD and relapse/death times

a) Both MAR: The probability that acute GvHD and relapse/death times are missing depends on the number of CB units transplanted:

$\pi_{i1} = 0.1\ (1 + 4\ x_{2i})$ and $\pi_{i2} = \pi_{i3} = 0.5\ (1 - 0.8\ x_{2i})$

π_{ij} are defined such that acute GvHD times are more likely to be missing for patients receiving a double cord rather than a single cord transplant; the situation is reversed for relapse/death times. This design means that it is unlikely that both acute GvHD and relapse/death times are missing for any given subject, reflecting the pattern of missingness observed in the NHS CBB dataset (see Chapter 4, Section 4.5).

b) Acute GvHD times MNAR and relapse/death times MAR: The shortest 30% of all acute GvHD times are missing; the probability that relapse/death times are missing depends on the number of CB units transplanted:

$\pi_{i1} = \begin{cases} 1 \text{ if } t_{i1} < t_{1(30\%)} \\ 0 \text{ otherwise} \end{cases}$ and $\pi_{i2} = \pi_{i3} = 0.5\ (1 - 0.8\ x_{2i})$

c) Acute GvHD times MNAR and relapse/death times MAR: The largest 30% of all acute GvHD times are missing; the probability that relapse/death times are missing depends on the number of CB units transplanted:

$\pi_{i1} = \begin{cases} 1 \text{ if } t_{i1} > t_{1(70\%)} \\ 0 \text{ otherwise} \end{cases}$ and $\pi_{i2} = \pi_{i3} = 0.5\ (1 - 0.8\ x_{2i})$

d) Acute GvHD times MAR and relapse/death times MNAR: The probability that acute GvHD imes are missing depends on the number of CB units transplanted; the shortest 30% of all relapse/death times are missing:

$\pi_{i1} = 0.1\ (1 + 4\ x_{2i})$ and $\pi_{iI} = \begin{cases} 1 \text{ if } t_{iI} < t_{RD(30\%)} \\ 0 \text{ otherwise} \end{cases}$ $l = 2, 3$

e) Acute GvHD times MAR and relapse/death times MNAR: The probability that acute GvHD times are missing depends on the number of CB units transplanted; the largest 30% of all relapse/death times are missing:

$\pi_{i1} = 0.1\ (1 + 4\ x_{2i})$ and $\pi_{iI} = \begin{cases} 1 \text{ if } t_{iI} > t_{RD(70\%)} \\ 0 \text{ otherwise} \end{cases}$ $l = 2, 3$

f) Acute GvHD times MNAR and relapse/death times MNAR: The shortest 30% of all acute GvHD and relapse/death times are missing:

$\pi_{i1} = \begin{cases} 1 \text{ if } t_{i1} < t_{1(30\%)} \\ 0 \text{ otherwise} \end{cases}$ and $\pi_{iI} = \begin{cases} 1 \text{ if } t_{iI} < t_{RD(30\%)} \\ 0 \text{ otherwise} \end{cases}$ $l = 2, 3$

I expected FCS MI methods to perform badly when event times were MNAR for both event types. Therefore, (to avoid repetition) I did not consider a scenario in which longest times were missing for both event types.

6.2.8. Estimands of interest

For each analysis, the estimands of interest were:

(i) The vector of regression parameters $\boldsymbol{\beta}_{hj}$ for the clock-forward Markov transition intensity models $\alpha_{hj}(t) = \alpha_{hj}^{0}(t) \exp(\boldsymbol{\beta}'_{hj} z_i)$ for all h and j. For compatibility with the DGM, disease status at time of transplant was the only covariate in transitions from transplant. Disease status at time of transplant and number of CB units transplanted were the covariates for the transition from acute GvHD to relapse/death.

(ii) The restricted expected length of stay (RELOS) in each state, restricted to the time period between transplant and two years post-transplant. RELOS was used as a summary of the transition probability distributions.

As described in Chapter 2, Section 2.2.3, RELOS from time 0 to time s post-transplant for state j is defined as:

$$e_j(s) = \int_0^s P_j(t)\, dt$$

where $P_j(t)$ is the probability of being in state j at time t (72). Since all patients are in state 0 initially, $P_j(t)$ is equivalent to the transition probability from state 0 (transplant) to state j at time t (73).

This measure was used (rather than, for example, the transition probability point estimates at one year post-transplant), so that the transition probability distributions over the entire time-period of interest could be taken into account (not just at discrete time-points). RELOS was calculated for patients with reference values of the covariates (those receiving a single cord transplant who were not in relapse at time of transplant). Since RELOS is, in general, a function of baseline intensities and regression parameters, using only these patients enabled any bias in estimation of the baseline intensity functions to be identified separately from any bias associated with regression parameter estimates.
The estimated value of $e_j(2)$ for the j^{th} state and k^{th} simulation, hereafter referred to as $\hat{e}_j^k(2)$, was calculated as the area under the transition probability curve, using the consistent estimator (72):

$$\hat{e}_j^k(2) = \sum_{m=0}^{M} \widehat{P}_j^k\left(t_m^k\right) . \left(t_{m+1}^k - t_m^k\right)$$

where, for the k^{th} simulation, $\widehat{P}_j^k(t)$ is the estimated probability of being in state j at time t and $t_0^k < t_1^k < \ldots < t_M^k \leq t_{M+1}^k$ are the set of ordered event times up to two years post-transplant, across all transitions. For Cox models, the set of event times was the set of all simulated event times for the k^{th} simulation. For Weibull models, the set of event times was specified as the

set of all values of t between zero and two years, in increments of 0.1 days.

(iii) The final estimand of interest was the regression parameter γ_{12} from a clock-forward semi-Markov transition intensity model for α_{12}. This model has the same form as the Markov model specified earlier, with the addition of an extra covariate, d_i, which denotes the time from transplant until acute GvHD:

$$\alpha_{12}(t) = \alpha_{12}^{0}(t) \exp(\gamma_{12} d_i + \boldsymbol{\beta'_{12}} \boldsymbol{z_i})$$

The inclusion of time until entry into current state is used in practice as a test for the Markov assumption (49). In this study, the regression parameter γ_{12} was included to test whether imputation leads to correct conclusions about the Markov assumption (given that, in the fully observed dataset, this assumption was true).

Estimates were calculated by fitting a Cox PH model for each transition, using the censored event times. As described in Chapter 2, Section 2.2.1, Cox models are often used in practice in the analysis of survival data. An advantage of Cox models is that they do not require explicit definition of the baseline intensity function, which is usually unknown for real data. Therefore, it was of interest to assess the performance of MI methods when using a Cox model. However, simulated data were generated from Weibull distributions. Therefore, it was also of interest to determine whether there was a difference in results when fitting a Cox model rather than a Weibull model, due to the disparity between the analysis model and DGM in the former case. Hence, analysis was repeated, fitting a Weibull PH model instead of a Cox model.

In the MCAR scenario, for all three main imputation methods (that is, the methods that were applied in all scenarios, see section 6.2.12), I performed two analyses: first using a Cox model and then a Weibull model. For all other MDMs (to avoid repetition), for all imputation models assessed, the analysis was performed using a Cox model. The analysis was repeated using a Weibull model

for one imputation method only. Cox models were fitted using the 'survival' R package (183) and Weibull models were fitted using the 'flexsurv' R package (188). Transition probabilities were calculated for both models using the 'mstate' R package (172).

6.2.9. Number of simulation datasets

One thousand datasets, each of size 500, were created using the DGM described above. Based on the empirical SE of the estimands using complete data (*i.e.* with no missing event times) when fitting a Cox model (see Section 6.2.17), 1000 datasets allowed estimation of the estimands of interest with acceptable precision (regression parameters: 0.01 in each case; RELOS: 0.1 days for the transplant state, and 1 day for the acute GvHD and relapse/death states; 1% for coverage for γ_{12}). Precision was calculated using the formula suggested by Burton *et al.* (173) with a type 1 error of 5%.

6.2.10. Candidate methods for handling missing event times

The methods identified as the best-performing when event times were missing in a competing risks framework (see Chapter 5, Section 5.4) were considered for handling missing event times in this study. The methods are MI FCS using type 1 PMM, and linear imputation on the original scale (with the latter method applied separately for each sub-group of patients with a different distribution of event times). These methods, and variations, are described below:

(i) Type 1 PMM imputation with no restrictions on the imputed values (PMM).

(ii) PMM, applied separately for patients who did and did not experience acute GvHD before relapse/death (PMMSUBGP). This method was used to assess whether imputation was improved if the heterogeneity of the two distributions of times to relapse/death (described in Section 6.2.8), was explicitly accounted for in the imputation scheme. My hypothesis was that

standard PMM would tend to identify donors with the same set of transitions as the incomplete case and hence there would be little difference in results for PMM and PMMSUBGP. After imputation by PMMSUBGP, analysis was performed using both Cox and Weibull models in all missing data scenarios.

(iii) PMM, accounting for the ordered nature of the event times as specified in the analysis model (PMMCOMP). In this method, PMM was applied separately for patients who did and did not experience acute GvHD before relapse/death. This method proceeds as follows:

a) Impute the first event time, *i.e.* impute relapse/death times for the subgroup of patients who did not experience acute GvHD, and acute GvHD event times for the subgroup who did.
b) Impute the time from acute GvHD to relapse/death for the sub-group who experienced acute GvHD. First, calculate the time from acute GvHD to relapse/death among those with observed times of acute GvHD and relapse/death. Second, draw imputations for the time of acute GvHD (using an imputation model that includes the calculated time from acute GvHD to relapse/death but excludes the relapse/death time) and for the time from acute GvHD to relapse/death (using an imputation model that includes the time of acute GvHD but excludes the relapse/death time). Post-imputation, calculate any missing relapse/death event time as the sum of the (observed or imputed) acute GvHD event time and the (observed or imputed) time from acute GvHD to relapse/death.

This approach has the advantage of compatibility with the analysis model because, for patients who experience acute GvHD and relapse/death, their imputed relapse/death event time is always greater than the (observed or imputed) acute GvHD event time. However, it has the disadvantage that

not all available information is used about patients with a missing acute GvHD time and an observed relapse/death time. This is because the observed relapse/death time is not included in the imputation model. For these patients, both the acute GvHD time, and the time from acute GvHD to relapse/death, will be imputed (and the analysed relapse/death time will have been calculated, post-imputation, from these imputed times). Therefore, SE are expected to be larger for this method than for the other methods. This method was only applied in the MCAR scenario.

(iv) Linear imputation on the original scale, applied separately for patients who did and did not experience acute GvHD before relapse/death (NORMSUBGP). This method has the disadvantage that negative event times may be imputed and such times cannot be handled by the 'mstate' package. As a workaround, negative imputed times were replaced by the value 0.0001 post-imputation. This was expected to lead to under-estimation of the SE, as was seen with restricted range methods in the competing risks simulation study (see Chapter 5, Section 5.4). This method was only applied in scenarios in which event times for only one event type were missing.

As in the competing risks analysis (Chapter 5, Section 5.2.7), each MI method was implemented using the 'mice' R package (93) with the default of five imputations. To explore whether increasing the number of imputations changed the bias and/or SE, PMM was also implemented using 30 imputations for the MCAR scenario (referred to as PMM30IMP). As in the competing risks simulation study (Chapter 5, Section 5.2.7), the default of five donors was used for each PMM method. In scenarios (i) and (ii), only one variable (the time of acute GvHD and relapse/death, respectively) was incomplete. Therefore, the number of cycles per imputation was one. In scenario (iii), the default of five iterations was used. R version 3.5.2 and 'mice' version 3.3.0 was used for all results.

Imputation models included the analysis outcome and all variables either predictive of event times, predictive of event times and missingness of event times, or included as covariates in the analysis model. That is, imputation models included both covariates, event times, and indicators for whether or not each event (acute GvHD and relapse/death) was experienced.

CCA was also performed to confirm that this would result in unbiased estimates when event times were MCAR and biased estimates when event times were MAR or MNAR.

6.2.11. Performance measures of interest

Performance measures of interest for regression parameters β_{hj} and RELOS within two years, $e_j(2)$, were bias, standardised bias and average model-based SE (see Chapter 5, Section 5.4.8, for definitions of these performance measures). Both bias and standardised bias were reported because both the absolute size of the bias and the size, relative to the empirical SE, give useful information in this simulation study. The impact of bias can be easier to visualise using the absolute bias *e.g.* a bias of 10 days in estimating the length of stay in the acute GvHD state may be easier to interpret than a standardised bias of 1.0. On the other hand, standardised bias is useful for comparing bias across the estimands, as these vary in scale. In addition, because standardised bias incorporates the uncertainty in the parameter estimate, it can be useful in evaluating the likely impact of bias on analysis conclusions (173). When comparing different estimates of the same estimand, standardised bias must be interpreted carefully. For example, estimates using two different methods may have the same absolute bias, but the estimate with the greater precision (smaller empirical SE) would have larger standardised bias. For completeness, empirical SE was also reported. The performance measure of interest for the regression parameter γ_{12} was the

coverage of the 95% confidence interval *i.e.* the percentage of within-simulation 95% confidence intervals for $\hat{\gamma}_{12}$ that included the true value.

Model-based SE of the estimated regression parameters was estimated using standard methods for parametric and Cox PH models via the 'flexsurv' (188) and 'survival' (183) R packages. However, based on the current literature, there is not a standard method for calculating the model-based SE for RELOS. Crowther and Lambert (189) used a simulation-based, bootstrap-type method, whereas Grand and Putter (72) used a pseudo-observations approach. Nemes *et al.* (190) reviewed estimators of the SE of restricted mean all-cause survival time (the univariate equivalent of RELOS). They described five different approaches based on the method of moments, bootstrapping, the delta method, estimating equations and pseudo-observations. In this study, I used three of the approaches proposed by Nemes *et al.* to derive estimators of SE[$\hat{e}_j^k(s)$], where SE[$\hat{e}_j^k(s)$] is defined as the standard error of RELOS up to time s, for state j, for the k^{th} simulation. These estimators are described below.

6.3. Description of estimators of the standard error of the restricted expected length of stay in state

The three proposed estimators were based on (i) the method of moments, (ii) bootstrapping and (iii) the delta method. I did not consider methods based on estimating equations or pseudo-observations because these methods could not be easily incorporated into the modelling approach used in this study.

(i) I derived the first estimator, $\widehat{SE}_1[\hat{e}_j^k(s)]$, by extending Royston and Parmar's method of moments approach for the restricted mean all-cause survival time (191). Following the steps in Royston and Parmar's derivation, my argument is as follows:

Define the time of transition from state j, T, with $X = \min(T, s)$ representing the transition time restricted to the time period 0 to s.
The method of moments estimator is defined as $\text{Var}(X) = \text{E}(X^2) - [\text{E}(X)]^2$ where $\text{E}(X^2) = \text{E}(T^2 \mid T \leq s)\ \text{P}(T \leq s) + \text{E}(T^2 \mid T > s)\ \text{P}(T > s)$

Here, $\text{P}(T > t)$ is the state occupation probability $P_j(t)$, the probability of being in state j at time t, and $\text{P}(T \leq t) = 1 - P_j(t)$, the probability of being in a state other than j at time t. Substituting $P_j(t)$ and applying integration by parts for the first term in this expression gives:

$$\text{E}(X^2) = s^2\,[1 - P_j(s)] - \int_0^s 2t\,[1 - P_j(t)]\,dt + s^2\,P_j(s)$$

$$= s^2 - \int_0^s 2t\,dt + \int_0^s 2t\,P_j(t)\,dt$$

$$= \int_0^s 2t\,P_j(t)\,dt$$

As before, $\text{E}(X) = \int_0^s P_j(t)\,dt$

Therefore, $\text{Var}(X) = \text{E}(X^2) - [\text{E}(X)]^2 = \int_0^s 2t\,P_j(t)\,dt - \left[\int_0^s P_j(t)\,dt\right]^2$

For the k^{th} simulation, $\widehat{SE}_1[\hat{e}_j^k(s)] = \sqrt{\frac{\widehat{\text{Var}}(X)}{n}}$, where n is the sample size.

Note that Royston and Parmar assume that the underlying transition intensity models are known, but in my study, these must be estimated.
I adapted the expression for the estimator of $\hat{e}_j^k(s)$, described in Section 5.2.10, to define an estimator of $\widehat{\text{Var}}(X)$ for the k^{th} simulation as follows:

$$\widehat{\text{Var}}(X) = \sum_{m=0}^{M} 2t_m^k.\,\widehat{P}_j^k\left(t_m^k\right).\left(t_{m+1}^k - t_m^k\right) - \left[\sum_{m=0}^{M} \widehat{P}_j^k\left(t_m^k\right).\left(t_{m+1}^k - t_m^k\right)\right]^2$$

where, for the k^{th} simulation, $\widehat{P}_j^k(t)$ is the estimated probability of being in state j at time t and $t_0^k < t_1^k < \ldots < t_M^k \leq t_{M+1}^k$ are the set of ordered event times up to two years post-transplant, across all transitions.

(ii) The second estimator $\widehat{SE}_2[\hat{e}_j^k(s)]$ was the non-parametric bootstrap estimator (192), using 50 bootstraps per simulated dataset. As mentioned previously (Chapter 3, Section 3.4.2), Bartlett and Hughes (108) note that it is valid to use bootstrapping after MI, provided the imputation and analysis models are congenial, but may lead to incorrect coverage if they are not.

(iii) I derived the third estimator, $\widehat{SE}_3[\hat{e}_j^k(s)]$, using the delta method.

For the Weibull PH model, I used the fact that $\hat{e}_j^k(s)$ is a function of the transition intensity model parameter estimates $\widehat{\boldsymbol{\theta}}$, with estimated covariance $\widehat{\boldsymbol{\Sigma}}$, in the application of the delta method. Hence, I defined $\widehat{SE}_3[\hat{e}_j^k(s)]$ as:

$$\widehat{SE}_3[\hat{e}_j^k(s)] = \sqrt{\left(\frac{\partial \hat{e}_j^k(s;\boldsymbol{\theta})}{\partial(\boldsymbol{\theta})}\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}\right)^T \widehat{\boldsymbol{\Sigma}} \left(\frac{\partial \hat{e}_j^k(s;\boldsymbol{\theta})}{\partial(\boldsymbol{\theta})}\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}\right)}$$

I estimated the vector of partial derivatives $\frac{\partial \hat{e}_j^k(s;\boldsymbol{\theta})}{\partial(\boldsymbol{\theta})}$ by calculating the difference in $\hat{e}_j^k(s)$ from increasing or decreasing each parameter estimate in turn by a small value ϵ, that is:

$$\frac{\partial \hat{e}_j^k(s;\boldsymbol{\theta})}{\partial(\boldsymbol{\theta})} = \frac{\hat{e}_j^k(s;\widehat{\boldsymbol{\theta}}+\epsilon) - \hat{e}_j^k(s;\widehat{\boldsymbol{\theta}} - \epsilon)}{2\epsilon}$$

It was not possible to use the same approach for the Cox PH model because the baseline cumulative hazard function in a Cox model is estimated non-parametrically (see Chapter 2, Section 2.2.1). Instead, as for the estimator of the SE of a percentile of event times (Chapter 5, Section 5.3), I adapted Collett's definition of SE for the median all-cause survival time (53), as follows:

$$\widehat{SE}_3[\hat{e}_j^k(s)] = \frac{1}{\widehat{P}_j^k\{\hat{e}_j^k(s)\}} \widehat{SE}[\widehat{P}_j^k\{\hat{e}_j^k(s)\}]$$

The SE of the state probability function at $\hat{e}_j^k(s)$, $\widehat{SE}[\widehat{P}_j^k\{\hat{e}_j^k(s)\}]$, was calculated using the Greenwood-style estimator described by de Wreede *et al.* (193). I

estimated the derivative $\widehat{P}'^{k}_{j}\{\hat{e}^{k}_{j}(s)\}$ by calculating a local gradient, such that:

$$\widehat{P}'^{k}_{j}\{\hat{e}^{k}_{j}(s)\} = \frac{\widehat{P}^{k}_{j}\{\hat{u}^{k}_{j}(s)\} - \widehat{P}^{k}_{j}\{\hat{l}^{k}_{j}(s)\}}{\hat{u}^{k}_{j}(s) - \hat{l}^{k}_{j}(s)}$$

where $\hat{u}^{k}_{j}(s) = \min\{t^{k}_{i} | t^{k}_{i} \geq \hat{e}^{k}_{j}(s) + \epsilon\}$ and $\hat{l}^{k}_{j}(s) = \max\{t^{k}_{i} | t^{k}_{i} \leq \hat{e}^{k}_{j}(s) - \epsilon\}$ for all event times i for the j^{th} event type and small ϵ.

In this study, ϵ was chosen to be 0.00001 for Weibull models and 10 days for Cox models. A large value of ϵ was needed for the Cox models to accommodate sparse event times; ϵ needed to be large to ensure that $\hat{u}^{k}_{j}(s)$ was different from $\hat{l}^{k}_{j}(s)$ for each simulated dataset.

6.3.1. Methods: study to compare estimators of the standard error of the restricted expected length of stay in state

The three proposed estimators of the model-based SE of RELOS were compared using a simulation study. Using each of the three proposed estimators, SE of RELOS (between 0 and 2 years post-transplant) was estimated for each state j for patients with reference values of covariates. The complete simulated data (*i.e.* without any missing event times) from all 1000 simulated datasets (created using the DGM described previously) was used to compare the estimators. Uncensored event times were used so that any differences in SE estimates could not be attributed to increased variation due to censoring. As before, clock-forward, Markov transition intensity models were fitted, using, firstly, a Cox PH model and, secondly, a Weibull PH model. Performance measures of interest were the average within-simulation model-based SE and coverage of each estimator of SE. For reference, bias and empirical SE for the estimates of RELOS were also calculated.

The true values of $e_j(2)$ for each state j were calculated as $\int_0^2 P_j(t)\, dt$, using numerical integration. The transition intensity models specified in the DGM were substituted into standard expressions for $P_0(t)$, $P_1(t)$ and $P_2(t)$ for a Markovian, clock-forward, illness-death model, (49) namely:

$$P_0(t) = \exp\left\{-\int_0^t \alpha_{01}(s) + \alpha_{02}(s)\, ds\right\}$$

$$P_1(t) = \int_0^t \alpha_{01}(s) \exp\left\{-\int_s^t \alpha_{12}(u) du\right\} P_0(s)\, ds$$

$$P_2(t) = \int_0^t \left(\alpha_{01}(s)\left(1 - \exp\left\{-\int_s^t \alpha_{12}(u) du\right\}\right) + \alpha_{02}(s)\right). P_0(s)\, ds$$

In addition, the theoretical values of the extended Royston and Parmar estimator, $SE_1[\hat{e}_j(s)]$, were calculated using the same method.

6.3.2. Results: study to compare estimators of the standard error of the restricted expected length of stay in state

The average within-simulation estimated model-based SE and coverage of the three estimators of the SE of $\hat{e}_j(2)$ for each state j, fitting either a Cox or Weibull model, are shown in Table 6.4. For reference, bias and empirical SE of the estimates of $\hat{e}_j(2)$, and theoretical (analytical) values of the extended Royston and Parmar estimator, $SE_1[\hat{e}_j(2)]$, are also shown.

Table 6.4. Average model-based standard error (ModSE) and coverage (Cov) of each estimator of the SE of RELOS between 0 and 2 years post-transplant.

Model	State	Bias of $\hat{e}_j(2)$	EmpSE of $\hat{e}_j(2)$	Estimator of SE of $\hat{e}_j(2)$						
				Royston & Parmar			Bootstrap		Delta	
				TheorSE	ModSE	Cov	ModSE	Cov	ModSE	Cov
Cox	Tx	-0.05	1.12	0.89	0.94	0.90	1.42	0.96	1.32	0.97
	AgvHD	-0.38	11.12	9.13	9.00	0.90	11.35	0.95	17.10	0.98
	R/death	-10.00	15.17	17.51	16.98	0.93	15.90	0.87	111.60	>0.99
Weibull	Tx	0.02	1.04	0.89	0.88	0.91	0.99	0.92	0.99	0.94
	AgvHD	-0.18	11.04	9.13	9.10	0.90	11.24	0.94	11.26	0.95
	R/death	0.16	11.18	17.51	17.50	>0.99	11.37	0.95	11.40	0.95

$\hat{e}_j(2)$, RELOS between 0 and 2 years post-transplant for state j; EmpSE, empirical SE; TheorSE, theoretical SE for Royston & Parmar estimator; Tx, transplant; AgvHD, acute GvHD; R/death, relapse/death.
Estimates use uncensored event times for patients with reference values of the covariates, based on 1000 simulated datasets.
Monte Carlo SE for bias was 0.04, 0.03 for $\hat{e}_0(2)$, 0.35, 0.35 for $\hat{e}_1(2)$, and 0.48, 0.35 for $\hat{e}_2(2)$ for Cox and Weibull models, respectively.
Monte Carlo SE for coverage was ≤ 0.01 in all cases.

The theoretical (analytical) values of RELOS between 0 and 2 years post-transplant were $e_0(2)$ = 26.3 days, $e_1(2)$ = 140.6 days, and $e_2(2)$ = 563.1 days. Values of RELOS estimated from the simulated data were unbiased for both Cox and Weibull models, except for the relapse/death state when fitting a Cox model (given Monte Carlo 95% CI for bias of (-0.13, 0.03) for the transplant state, (-1.07, 0.31) for the acute GvHD state, and (-10.94, -9.06) for the relapse/death state for Cox models; and (-0.04, 0.08) for the transplant state, (-0.87, 0.51) for the acute GvHD state, and (-0.53, 0.85) for the relapse/death state for Weibull models). The bias in this case can be explained by the large time intervals between individual simulated relapse/death event times.

Estimates of SE using the extended Royston and Parmar (R&P) method, $\widehat{SE}_1[\hat{e}_j^k(s)]$, were very similar to the theoretical calculated values for this estimator, for both Cox and Weibull models. However, the R&P estimated SEs differed from the empirical SEs, as follows. For the transplant and acute GvHD states, R&P estimated SEs were less than empirical SE and, for the relapse/death state, R&P estimated SEs were greater than empirical SE. The difference between the R&P estimated SEs and empirical SE here may be because the Royston and Parmar method treats model parameters as fixed. Hence, the covariance associated with model parameter estimation (which may have a positive or negative effect on SE) is not accounted for. The under-estimation of the SE could explain the under-coverage for transplant and acute GvHD states for both Cox and Weibull models (since the point estimates are unbiased) (75). Similarly, the over-estimation of the SE could explain the over-coverage for the relapse/death state for the Weibull model. The bias in the estimate of RELOS for the relapse/death state for the Cox model could explain the under-coverage in this case (75).

Estimates of SE using the bootstrap method, $\widehat{SE}_2[\hat{e}_j^k(s)]$, were very similar to the empirical SE for all three states, for both Cox and Weibull models. However,

there was under-coverage for the relapse/death state from the Cox model, and for the transplant state from the Weibull model. As before, the bias in the estimate of RELOS for the relapse/death state from the Cox model explains the greater degree of under-coverage in this case. For the Weibull model, the bootstrap estimated SE is slightly lower than the empirical SE for the transplant state, which explains the under-coverage in this case.

Estimates of SE using the delta method, $\widehat{SE}_3[\hat{e}_j^k(s)]$, were also very similar to the empirical SE for all three states from Weibull models, with appropriate coverage. However, for the Cox models, performance of this estimator was poor, with over-coverage for all states. For the transplant state, the delta estimated SE was similar to the empirical SE. However, for acute GvHD and relapse/death states, delta estimated SEs were larger than the empirical SEs; for the relapse/death state, the SE estimate was nearly 10 times larger. This can be explained by the large time intervals between individual simulated relapse/death event times; Royston and Parmar (191) comment that the delta method does not work well with sparse event times. The results for the bootstrap and delta methods suggest that, in contrast to the R&P method, both these methods appropriately incorporate the covariance associated with model parameter estimation.

The best-performing estimators of the SE of RELOS were used in the main simulation study. The bootstrap estimator was used when fitting Cox models and the delta estimator was used when fitting Weibull models.

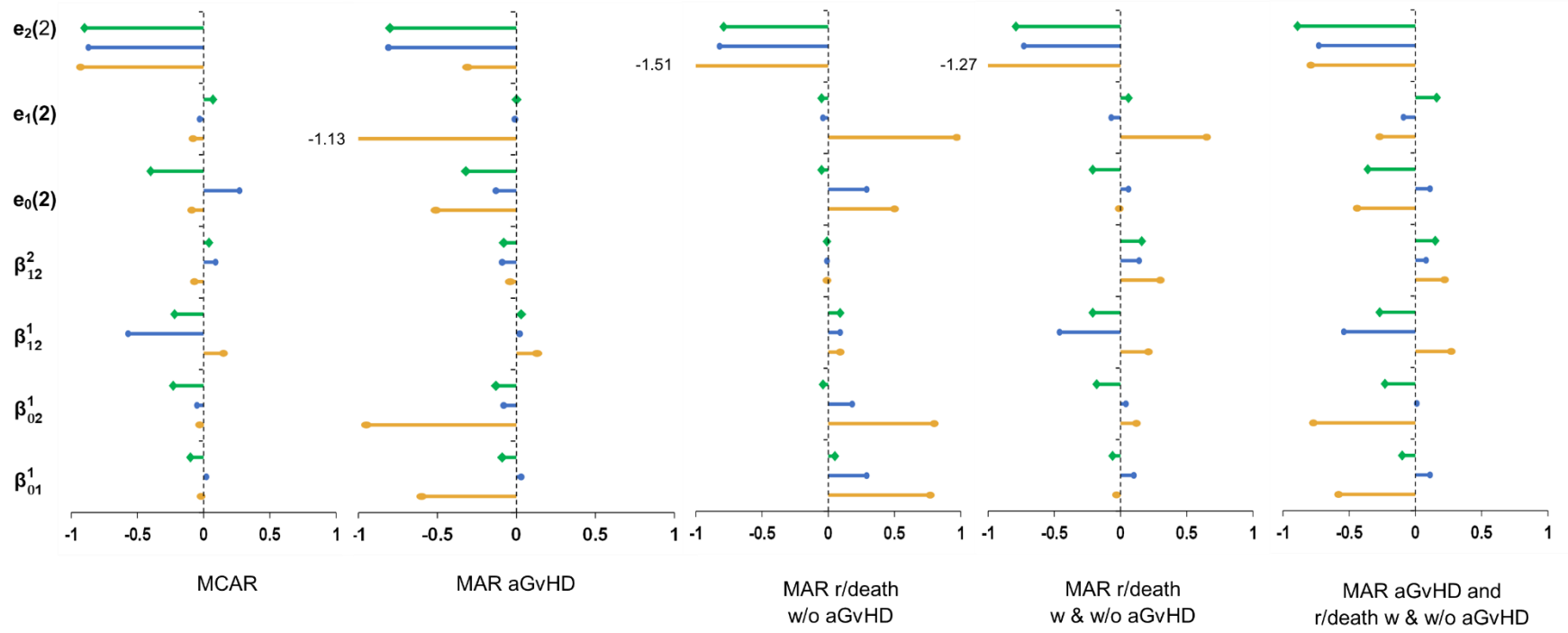
6.4. Results: simulation study comparing imputation methods for handling missing event times in a multi-state model

The true (theoretical) values of the regression parameters in the transition intensity models were β_{01}^1 = -0.8, β_{02}^1 = 1.2, β_{12}^1 = 1.2 and β_{12}^2 = 1.0. The true values

of RELOS between 0 and 2 years were $e_0(2) = 26.3$ days (alive and relapse- and acute GvHD-free post-transplant), $e_1(2) = 140.6$ days (alive and relapse-free following acute GvHD) and $e_2(2) = 563.1$ days (in relapse/died). Estimates based on complete simulated data (*i.e.* without any missing event times) were unbiased for most estimands, for both Cox and Weibull models (*i.e.* bias was 0, or Monte Carlo 95% CI for bias contained 0, see Appendices, Tables A.4a and A.4b). There was slight bias (Monte Carlo 95% CI for bias was (0.01, 0.03)) in estimation of β_{12}^{1} for both Cox and Weibull models. As described in Section 6.2.16, there was substantial bias (Monte Carlo 95% CI for bias was (-14.44, -12.36)) for $e_2(2)$ when fitting a Cox model, due to sparse event times.

As expected, CCA gave unbiased estimates only when event times were MCAR, and model-based SE was larger than when using MI methods. Simulation study results are illustrated in Figures 6.3 and 6.4 using "lollipop" plots and all results are included in the Appendices (Tables A.4a and A.4b). Figures 6.4 and 6.5 show the standardised bias of regression parameters $\boldsymbol{\beta}_{hj}$ and RELOS, $e_j(2)$, for each transition intensity fitted using a Cox model. Results are illustrated for the three main missing data methods (that is, the methods that were applied in all scenarios): CCA, PMM and PMMSUBGP. Bias and model-based SE are not illustrated because these could not be shown on the same scale for all estimands, and because model-based SE was similar for all imputation methods and MDMs. Similarly, coverage of γ_{12} is not illustrated because, with one exception, discussed later, it was similar for all imputation methods and MDMs. Figure 6.4 shows results for scenarios in which MI was expected to work well, that is, when all event times were either completely observed, MCAR or MAR. Conversely, Figure 6.5 shows results for scenarios in which MI was not expected to work well, that is, when some or all event times were MNAR.

Figure 6.4. Standardised bias of regression parameters $\boldsymbol{\beta}_{lm}$ and RELOS between 0 and 2 years, $e_j(2)$, for each transition intensity model given event times MCAR and MAR for imputation methods CCA (yellow oval), PMM (blue circle) and PMMSUBGP (green diamond)



Values are stated for points outside the scale.

State indicators: 0 = transplanted; 1 = acute GvHD (aGvHD); 2 = relapse/death (r/death)

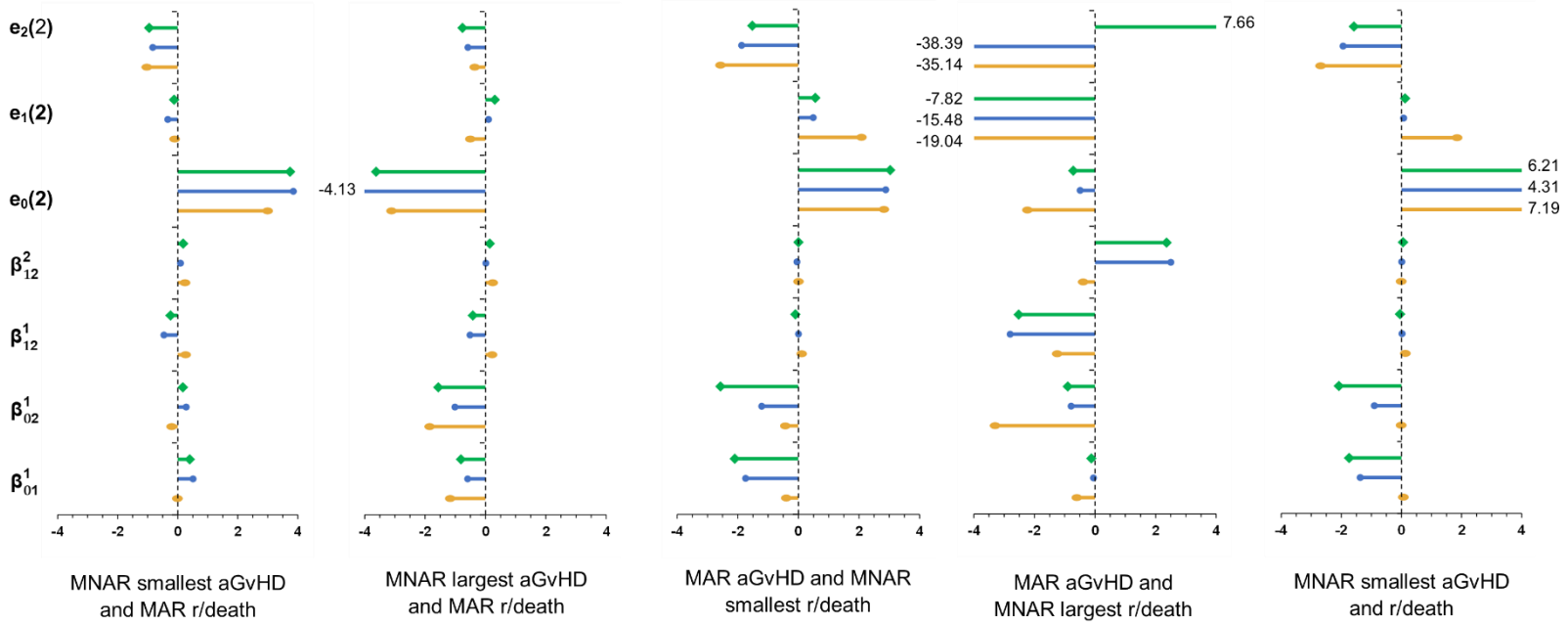
β^1 and β^2 parameters are for disease status at time of transplant and number of CB units transplanted, respectively.

CCA, complete case analysis

PMM, MI by Type 1 predictive mean matching

PMMSUBGP, as for PMM with imputation models fit separately for subjects with (w) and without (w/o) aGvHD

Figure 6.5. Standardised bias of regression parameters $\boldsymbol{\beta}_{lm}$ and RELOS between 0 and 2 years, $e_j(2)$, for each transition intensity model given some or all event times MNAR for imputation methods CCA (yellow oval), PMM (blue circle) and PMMSUBGP (green diamond)



Values are stated for points outside the scale.

State indicators: 0 = transplanted; 1 = acute GvHD (aGvHD); 2 = relapse/death (r/death)

In all scenarios, missingness mechanisms for r/death times were applied to both subjects with and without aGvHD

MNAR smallest/largest means that the smallest or largest event times were missing

β^1 and β^2 parameters are for disease status at time of transplant and number of CB units transplanted, respectively.

CCA, complete case analysis

PMM, MI by Type 1 predictive mean matching

PMMSUBGP, as for PMM with imputation models fit separately for subjects with and without aGvHD

When event times were either MCAR or MAR (Figure 6.4), MI by PMM gave small bias for all estimands (that is, the magnitude of the standardised bias was <0.5), except for $e_2(2)$ (RELOS for the relapse/death state) and the regression parameter β_{12}^1 (for the covariate "in relapse or not at time of transplant" in the transition intensity model from acute GvHD to relapse/death). The bias in the RELOS estimate, $e_2(2)$, remained for all imputation methods and MDMs when fitting a Cox model, so is not further discussed here. Bias for regression parameter β_{12}^1 was large in scenarios when event times for relapse/death after acute GvHD were missing and small when only event times for acute GvHD or relapse/death without acute GvHD were missing. Results were very similar for PMM whether the number of imputations was the default of five or increased to 30 (see Appendices, Tables A.4a and A.4b).

Applying PMM separately to patients who did and did not experience acute GvHD before relapse/death (PMMSUBGP) or accounting for the ordered nature of the event times as specified in the analysis model (PMMCOMP) reduced the bias in regression parameter β_{12}^1. When these methods were used, bias remained small for all other estimands except the RELOS estimate, $e_2(2)$. However, bias in the estimate of RELOS for the acute GvHD state, $e_1(2)$, was slightly larger for PMMSUBGP than for PMM in some scenarios. Model-based SE was slightly smaller for PMMSUBGP compared with PMM, and was larger for PMMCOMP than for other methods with respect to regression parameters β_{12}^1 and β_{12}^2, as expected (see Appendices, Tables A.4a and A.4b). Coverage for γ_{12} was in the range 0.92-0.97 for all methods, with coverage slightly closer to the nominal value for PMMSUBGRP compared with PMM (see Appendices, Tables A.4a and A.4b). Results using PMMSUBGP were very similar for both Cox and Weibull models, except that the bias in the RELOS estimate, $e_2(2)$, was greatly reduced when fitting a Weibull model. MI using a linear imputation model, with replacement of a small number of negative imputed times by the value 0.0001 post-imputation, did not lead to under-estimation of the SE as predicted.

However, it did result in large bias for some estimands, particularly estimates of RELOS.

When some or all event times were MNAR (Figure 6.5), MI using both PMM and PMMSUBGP led to biased estimates. Bias was generally the same or larger than when using CCA. Using MI, bias was larger when the time to the absorbing state (relapse/death) was MNAR than when the time to the intermediate state (acute GvHD) was MNAR, and when largest times were MNAR than when smallest times were MNAR. As described previously, times to relapse/death tend to be longer for patients who experience acute GvHD than for patients who experience relapse/death without acute GvHD. Therefore, MNAR mechanisms where longer times to relapse/death tended to be missing mainly affected patients who experienced acute GvHD before relapse/death. Conversely, MNAR mechanisms where shorter times to relapse/death tended to be missing mainly affected patients who experienced relapse/death without acute GvHD. This may explain why parameter estimates for the acute GvHD to relapse/death transition intensity model, β_{12}^{1} and β_{12}^{2}, are more biased than parameter estimates for the models of transition from transplant, β_{01}^{1} and β_{02}^{1}, when largest relapse/death times are MNAR and vice versa when smallest relapse/death times are MNAR.

In a separate semi-Markov model analysis, the time from transplant until acute GvHD was added as a covariate to the transition intensity model from acute GvHD to relapse/death. This variable was included as a check of the Markov assumption. Coverage for the regression parameter for this covariate, γ_{12}, was in the range 0.92-0.98 in all methods and scenarios, except one. The coverage was 0.66 when applying MI using the PMMSUBGP method and a Weibull analysis model, with acute GvHD times MAR and largest relapse/death times MNAR. To allow further exploration of this outlying value for coverage, performance measures for the regression parameter γ_{12} are shown in Table 6.5 overleaf, for all scenarios in which acute GvHD times were MAR, relapse/death times were

MNAR and the imputation method was PMMSUBGP. As discussed above, MNAR mechanisms in which smallest times of relapse/death times tended to be missing affected mainly patients who experienced relapse/death without acute GvHD. Hence, the regression parameter γ_{12} is unbiased with coverage close to the nominal value in this scenario (Table 6.5). In MNAR mechanisms in which largest relapse/death times tended to be missing, there is little bias when fitting a Cox model. However, the model-based SE is larger, which may explain the slight over-coverage in this case. Bias is large when fitting a Weibull model, which may explain the high degree of under-coverage in this case.

Table 6.5. Performance measures for estimates of the regression parameter γ_{12} in the transition intensity model from acute GvHD to relapse/death when event times are MNAR.

Estimand (true result)		**γ_{12} (0.95)**			
Missing data mechanism	**Imputation method**	**Bias**	**Mod SE**	**Std Bias**	**Cov**
MAR (acute GvHD) & MNAR (smallest relapse/death)	PMMSUBGP	-0.001	0.004	-0.29	0.94
	PMMSUBGP*	<0.001	0.004	0.02	0.94
MAR (acute GvHD) & MNAR (largest relapse/death)	PMMSUBGP	-0.001	0.005	-0.26	0.98
	PMMSUBGP*	0.007	0.005	1.44	0.66

ModSE, average model-based SE; StdBias, standardised bias; Cov, coverage.
Parameter γ_{12} is for the time from transplant until acute GvHD.
Cox models were fit, except for methods indicated with a *, for which Weibull models were fit.
PMMSUBGP, MI by Type 1 predictive mean matching with imputation models fit separately for patients with and without acute GvHD.

6.5. Discussion

In this chapter, a simulation study was described, whose aim was to evaluate the bias and precision of estimates from a MSM in different missing data scenarios, comparing the best previously identified MI methods for handling missing event times. In a competing risks analysis (Chapter 5), type 1 PMM (PMM) and a linear imputation model applied separately for patients who did and did not experience acute GvHD (thus allowing for different distributions of event times, NORMSUBGP), gave similar results, with small bias.

In PMM, missing values are replaced by sampling at random from a donor pool of patients (with observed values) who are 'similar' to the subject with missing

data. In the MSM context, this means that the donor pool tends to contain patients who have experienced the same sequence of events as the incomplete case, hence preserving the original sequence of events for the incomplete case. This may explain the generally small bias in estimates for PMM. In my simulation study, the distribution of relapse/death times was different for patients who did and did not experience acute GvHD. Applying PMM separately for sub-groups of patients who did and did not experience acute GvHD (PMMSUBGP) tended to reduce bias and model-based SE when estimates depended on the time of relapse/death. PMMSUBGRP also improved coverage in a parameter used to test the Markov assumption (by including time of acute GvHD in a separate semi-Markov model). PMMSUBGRP worked well with both semi-parametric (Cox) and parametric (Weibull) PH models for the transition intensities.

I also considered an alternative method for preserving the sequence of events experienced, by including the acute GvHD event time and time from acute GvHD to relapse/death in the imputation model, but not the relapse/death event time (PMMCOMP). PMMCOMP gave results with comparable bias to PMMSUBGP, but larger model-based SE. Due to the loss of information in this approach, with no advantage in terms of bias reduction, I would not recommend this approach.

In this study, the NORMSUBGP approach led to more bias than PMM when estimating transition intensity model parameters and RELOS. This may be because negative event times could be imputed in the NORMSUBGP method. Such times were replaced by the value 0.0001 post-imputation (because negative times cannot be handled by the 'mstate' package). The NORMSUBGP approach could result in an imputed relapse/death time that was smaller than the (observed or imputed) acute GvHD time. Hence, the NORMSUBGP approach was not compatible with the analysis model (in which a subject's acute GvHD

time must be less than their relapse/death time). This may explain the bias in estimates using this method.

Overall, I recommend using type 1 PMM to impute missing event times in a MSM analysis. I recommend first exploring the distribution of event times for each sub-group of patients with a different path through the MSM. Then, type 1 PMM should be applied separately for each sub-group of patients with a different distribution of event times. In my simulation study, the distributions of time of relapse/death were very different for patients who did and did not experience acute GvHD. In analysis of real data, there may be smaller differences between distributions of event times for different sub-groups of patients. Hence, results may be similar, whether PMM is applied for all patients in one imputation model or applied separately for different sub-groups. To assess the sensitivity of results to the imputation method, analysis could be performed using both PMM and PMMSUBGP.

Based on the current literature, there is not a standard method for calculating the model-based SE for RELOS. I proposed three different estimators (an extended version of the Royston and Parmar method (191), the non-parametric bootstrap, and the delta method). Based on my simulation study, comparing the performance of these SE estimators, I recommend using the bootstrap estimator when fitting Cox models and the delta estimator when fitting Weibull models. The delta estimator is not recommended when fitting a Cox model, particularly if there are large time intervals between individual event times. I do not recommend the extended Royston and Parmar method, because (in contrast to the bootstrap and delta methods) my results suggest that it does not appropriately incorporate the covariance associated with model parameter estimation.

Generally, FCS MI techniques that assume MAR are not recommended when data are MNAR. In this study, imputation resulted in biased estimates when

event times were MNAR or a mixture of MAR and MNAR. My results suggest that bias is greater when the time to the absorbing state (relapse/death) is MNAR than when the time to the intermediate state (acute GvHD) is MNAR. This may be because simulated acute GvHD times were mostly between 0 and 100 days post-transplant (consistent with the real dataset). In contrast, relapse/death times were more widely distributed, with different distributions for patients who did and did not experience acute GvHD. Hence, there was more potential for relapse/death times to be imputed incorrectly. The direction and size of the bias in my results may be specific to my DGMs, missingness mechanisms and analysis model. However, in general, the constrained nature of the time to an intermediate event (in an illness-death model, this is bounded by 0 and the time of transition to the absorbing state) may limit the degree of bias even when event times are MNAR. Conversely, the lack of constraint on the maximum time of transition to an absorbing state, and the different pathways through the MSM to that state (each potentially with a different distribution of event times), may increase the degree of bias. In addition, data-generation was based on the Markov assumption, such that the time of relapse/death did not depend on the time of acute GvHD. Violation of this assumption, as in a semi-Markov or non-Markov model, may result in greater bias when intermediate state event times are MNAR.

Generally, parametric analysis models performed as well as semi-parametric models. Furthermore, parametric models resulted in less biased estimates of the expected length of stay in state (RELOS) when there were sparse event times. However, regression parameter estimates from parametric models were more biased than estimates from semi-parametric models when event times were MNAR. Therefore, I recommend careful exploration of likely missingness mechanisms for the dataset of interest, to identify parameters likely to be affected by event times MNAR. I do not recommend using FCS MI methods that assume MAR if any incomplete variable in the dataset is suspected to be MNAR.

Further studies are required to determine whether PMM is a suitable MI method when using more complex analysis models *e.g.* MSM with multiple intermediate states, with reversible or recurring transitions between states or with multiple time scales (68). This could occur in my real data if, for example, myeloid engraftment was included in the MSM because myeloid engraftment can occur before or after onset of acute GvHD. If both event times were missing then both the order of events, as well as the event times themselves, would be unknown. It should be noted that the 'mstate' R package can only handle uni-directional transitions (172).

Although PMM performed well in my study, there is still scope for improvement, by ensuring compatibility between the imputation and analysis model. This could be achieved, for example, by extending the MAR stacked MI approach (119) or the SMC-FCS method (194) to MSM analysis. SMC-FCS methods have been applied to cause-specific semi-parametric hazards models in a competing risks framework (127). These methods could be extended to accommodate the sequential nature of the events in a MSM *e.g.* as in my study, to ensure that imputed times of relapse/death are greater than the (imputed or observed) times of acute GvHD.

CHAPTER 7. APPLICATION OF METHODS TO THE NHS CORD BLOOD BANK COHORT

7.1. Introduction

The overall aim of my research is to provide the first insight into patient outcomes after transplantation using CB donated to the NHS CBB. In this chapter, I describe the incidence of each event of interest (myeloid engraftment, acute and chronic GvHD, relapse, and overall survival) for patients in the NHS CBB cohort. In addition, I will use MSM analysis to identify covariates associated with the events of interest, and to make probability predictions for NHS CBB patients with certain sets of characteristics.

In the NHS CBB dataset, event times are often missing for the post-transplant events of interest. In this case, CCA will be biased because missingness depends on the analysis outcome (event times are incompletely observed for some event types, but completely observed for others). Biased analysis results may lead to incorrect conclusions about the risks of CB transplantation, which will have considerable consequences for both clinicians and patients.

The previous three chapters of my thesis have been focused on exploring the potential missingness mechanisms for event times in the NHS CBB dataset (Chapter 4) and comparing methods for handling missing event times (Chapters 5 and 6). Using simulations studies, I have identified FCS MI strategies for handling missing event times that can be used to correct bias and improve precision (compared with CCA and other naïve methods) in a competing risks analysis (Chapter 5) and a MSM analysis (Chapter 6). I found that FCS MI using

type 1 PMM was the best available method for handling missing event times when times were MAR, conditional on the observed data.

In this chapter, I apply FCS MI methods to impute missing event times and covariate data in the NHS CBB dataset. I describe the results of competing risks and MSM analyses after imputation. As a baseline, I compare results with a complete case analysis (CCA).

7.2. Estimands of interest

In analyses of the NHS CBB cohort, the estimands of interest were:

(i) The cumulative incidence at 100 days and median time of myeloid engraftment (with competing risks of graft failure and death prior to engraftment).

(ii) The cumulative incidence at 100 days and median time of acute GvHD (with competing risks of graft failure and death prior to acute GvHD).

(iii) The cumulative incidence at one year and 10^{th} percentile time of chronic GvHD (with competing risks of graft failure and death prior to chronic GvHD).

(iv) The cumulative incidence at one year and 10^{th} percentile time of relapse (with competing risk of death prior to relapse).

(v) Overall survival at one year post-transplant and lower quartile time of death.

From a MSM analysis:

(vi) HRs of covariates for each transition intensity (for the transitions from transplant to acute GvHD and relapse/death; from acute GvHD to relapse/death).

(vii) The expected length of stay in each state (transplant, acute GvHD, relapse/death) in the first year post-transplant (RELOS).

Differences in the state occupation probability functions will be examined by comparing estimates of RELOS, since RELOS is a function of the state occupation probabilities (see Chapter 6, Section 6.2.10). In addition, state occupation probability functions in the first year post-transplant will be illustrated graphically.

7.3. Statistical methods

Estimates were obtained using the same methods as in the simulation studies described in Chapters 5 and 6. Briefly, the methods are as follows (see Chapter 2, Section 2.2 for further details):

- Estimands (i) to (iv) (cumulative incidence of various events) were obtained using the non-parametric Aalen-Johansen estimator of the cumulative incidence function for each event of interest.

- Estimand (v) (overall survival) was obtained using the non-parametric Kaplan-Meier estimator.

- Estimands (vi) to (viii) (HR, RELOS and state occupation probabilities) were obtained using Cox PH models for each transition intensity of the MSM, assuming a clock-forward Markov model for all transitions (see Chapter 6, Section 6.2.2 for a detailed description of this model). RELOS and state occupation probability functions were calculated from these models for three different patient types, namely, a patient with reference values of all covariates, a low-risk, and a high-risk patient. Characteristics of low-risk and high-risk patients were chosen according to the strongest covariate associations identified in the transition intensity analysis (and also to be representative of patients in the NHS CBB dataset). State occupation probability functions were presented graphically, to illustrate the differences between the three patient types.

The following tests of the transition intensity modelling assumptions were performed:

(i) The PH assumption was tested for each transition intensity model using the global test (*i.e.* testing for proportional hazards across all covariates in combination) proposed by Grambsch and Therneau (195).

(ii) As a test of the Markov assumption, an additional semi-Markov model was fitted for α_{12}, including the time from transplant until acute GvHD as well as all covariates (see Chapter 6, Section 6.2.10).

My simulation study results (see Chapter 6, Section 6.2.17) suggested that estimates of RELOS were biased when event times were sparsely distributed, and Cox models were used for transition intensities. Therefore, as a sensitivity analysis (sensitivity analysis 1), Weibull models were also fitted, and the results compared.

In all MSM analyses, models included all baseline (at time of transplant) covariates shown in the literature to be clinically relevant (see Chapter 1, Table 1.1). All covariates were included in all models because the purpose of my MSM analysis was to use the transition intensity models to make probability predictions for patients with certain sets of characteristics.

Transition intensity model covariates

The clinically relevant baseline characteristics included as covariates in each transition intensity model were: number of CB units transplanted (single or double); patient age (in years, assuming a linear association); disease type (acute leukaemia, other blood cancer, non-malignant disorder); disease status at time of transplant (in remission, relapse, other); conditioning regimen (intensive or not); sex and CMV+ match between donor(s) and recipient; number of HLA mismatches between donor(s) and recipient (well-matched: 0 or 1 mismatches; or not: 2 or more mismatches); TNC dose at infusion (low, <3.0; medium, 3.0-5.0; high, >5.0 $\times 10^7$/kg).

7.3.1. Statistical software

Overall survival estimates and SEs were calculated using the 'survival' R package. Estimates, SE of cumulative incidence functions were calculated using the 'Cuminc' function of the 'mstate' R package, or the 'cifDMnocens' function provided by Pintilie (164) if there was no censoring (this applies for myeloid engraftment), because the 'Cuminc' function requires at least one censored event time. Estimates, SE from Cox transition intensity models, and transition probabilities were calculated using the 'mstate' R package. Estimates, SE from Weibull transition intensity models were calculated using the 'flexsurv' R package. I manually calculated estimates, SE of percentile times and SE of RELOS. For the SE of percentile times, I used the estimator I derived in Chapter 5, Section 5.3. As per my simulation study assessing the performance of this estimator (described in Chapter 5, Section 5.3), here, I used a value of 0.01 for the estimator parameter, ϵ, and a simple gradient function for probability density function estimation. For the SE of RELOS, I used the estimators that I derived in Chapter 6, Section 6.2.14. As per my simulation study comparing the performance of these SE estimators (described in Chapter 6, Section 6.2.15-16), here, I used the bootstrap estimator when fitting Cox models and the delta estimator when fitting Weibull models.

7.4. Methods for handling missing data

7.4.1. Missing event times

As described in Chapter 4, Section 4.4, event times were incompletely observed for some event types in the NHS CBB dataset. These were (with percentage missing, of patients who experienced the event): acute GvHD time (24%), chronic GvHD time (35%), relapse time (25%) and graft failure time (11%). The best-performing method for handling missing event times, identified in the simulation studies (see Chapters 5 and 6), was applied to the NHS CBB dataset. This method was FCS MI using type 1 PMM (PMM). Results from my

simulation studies suggest that bias and SE may be reduced by applying PMM separately for each sub-group (PMMSUBGP) with a different distribution of event times. Therefore, to explore whether PMM should be applied in this way in the NHS CBB dataset, the distribution of times for each event was compared for the five patient sub-groups (as defined in Chapter 4, Section 4.4.1). The sub-groups were patients experiencing (i) acute and chronic GvHD; (ii) acute without chronic GvHD; (iii) chronic without acute GvHD; (iv) relapse without GvHD; and (v) neither relapse nor GvHD.

7.4.2. Missing covariate data

Some covariates were also incomplete, in two ways:

(i) For transplants in which the single donor or both donors (of a double CB transplant) donated to the NHS CBB, some baseline data and/or event times were not reported. Missing values were imputed using FCS MI.

(ii) For double CB transplants, where one of the two donors did not donate via the NHS CBB, none of this donor's data were available. In previous chapters (see Chapter 4, Section 4.3.1 for a full explanation), values based on the known donor's data (for sex and CMV+ match between donor(s) and recipient, number of HLA mismatches between donor(s) and recipient, and TNC dose at infusion), were treated as fully observed values (*i.e.* as though they were based on both donors' data). In this analysis, such values are treated as missing and are imputed using FCS MI.

To assess whether using the known donor's data as a proxy for both donors' data would change the analysis results, a sensitivity analysis (sensitivity analysis 2) was performed.

The sensitivity analysis used a similar approach to that suggested by Cornish *et al.* (196).

The approach was as follows:

- As in the main analysis, for double cord transplants, values based on one of two donors' data (sex and CMV+ match between donor(s) and recipient; number of HLA mismatches between donor(s) and recipient; TNC dose at infusion) were treated as missing and were imputed using FCS MI.
- For each of these variables in turn, the result based on the known donor's data was used in the imputation models for all other variables.

The incomplete covariates were (with percentage missing data for the main analysis and percentage missing if treating the known donor's data as the fully observed result): disease status at time of transplant (30%, 30%); conditioning regimen (4%, 4%); sex match (20%, 1%); CMV+ match (29%, 12%); number of HLA mismatches (29%, 12%) between donor(s) and recipient; TNC dose at infusion (56%, 24%). Covariate data were imputed using standard methods: binary variables (conditioning regimen and number of HLA mismatches) using logistic regression, and categorical variables (disease status at time of transplant, sex match, CMV+ match, and TNC dose at infusion) using multinomial regression.

7.4.3. Imputation models

As is standard for MI (102), imputation models included analysis outcomes, analysis model covariates, and all other variables predictive of missing data. In survival analysis, it is well-established that both the event indicator (a binary variable indicating whether the event was experienced or not) and a representation of the distribution of event times should be included in imputation models (91, 101, 123). In this analysis, both covariate data and event times are imputed. Hence, the actual event times are included in the imputation models, rather than, for example, the baseline hazard function recommended by

White and Royston (123). Since the covariates included in the MSM are also thought to be predictive of event times (based on the clinical literature, see Chapter 1, Table 1.1), the same imputation models were used for all analysis models. Year and country of transplant were also included in each imputation model because they were highly predictive of missingness of event times (see Chapter 4, Section 4.5), and may be (directly or indirectly) predictive of the event times themselves. Country of transplant was also missing for one patient; as before, this was imputed using multinomial regression.

All missing data were imputed by FCS MI using the 'mice' R package (version 3.3.0). In the simulation studies, increasing the number of imputations (from the default number of five) did not change the bias or precision of estimates. However, the relationship between efficiency, reproducibility, and the number of imputations depends on the fraction of missing information (FMI, the ratio of between-imputation variance to the sum of between- and within-imputation variance). It is likely that FMI is greater in the real dataset, in which there are many different patterns of missingness and multiple covariates, than in the simulated datasets (which have a relatively simple structure). Therefore, here, I use a more conservative approach, letting the number of imputations equal the percentage of incomplete cases. As in the competing risks simulation study (Chapter 5, Section 5.2.7), the default of five donors was used for each PMM method. The default of five iterations was used. To obtain final analysis estimates and 95% CI, I calculated the mean and SE of per-imputation estimates of HR and RELOS. SE was calculated using Rubin's rules (88).

I used a different approach to calculate state occupation probability functions, because these rely on specification of the baseline hazard function at all time-points (see Chapter 6, Section 6.2.10). Since the baseline hazard function of a Cox model is estimated non-parametrically (see Chapter 2, Section 2.2.1), it is not possible to combine the baseline hazards for each imputed dataset (whereas, for example, combining model parameter estimates if fitting a parametric model

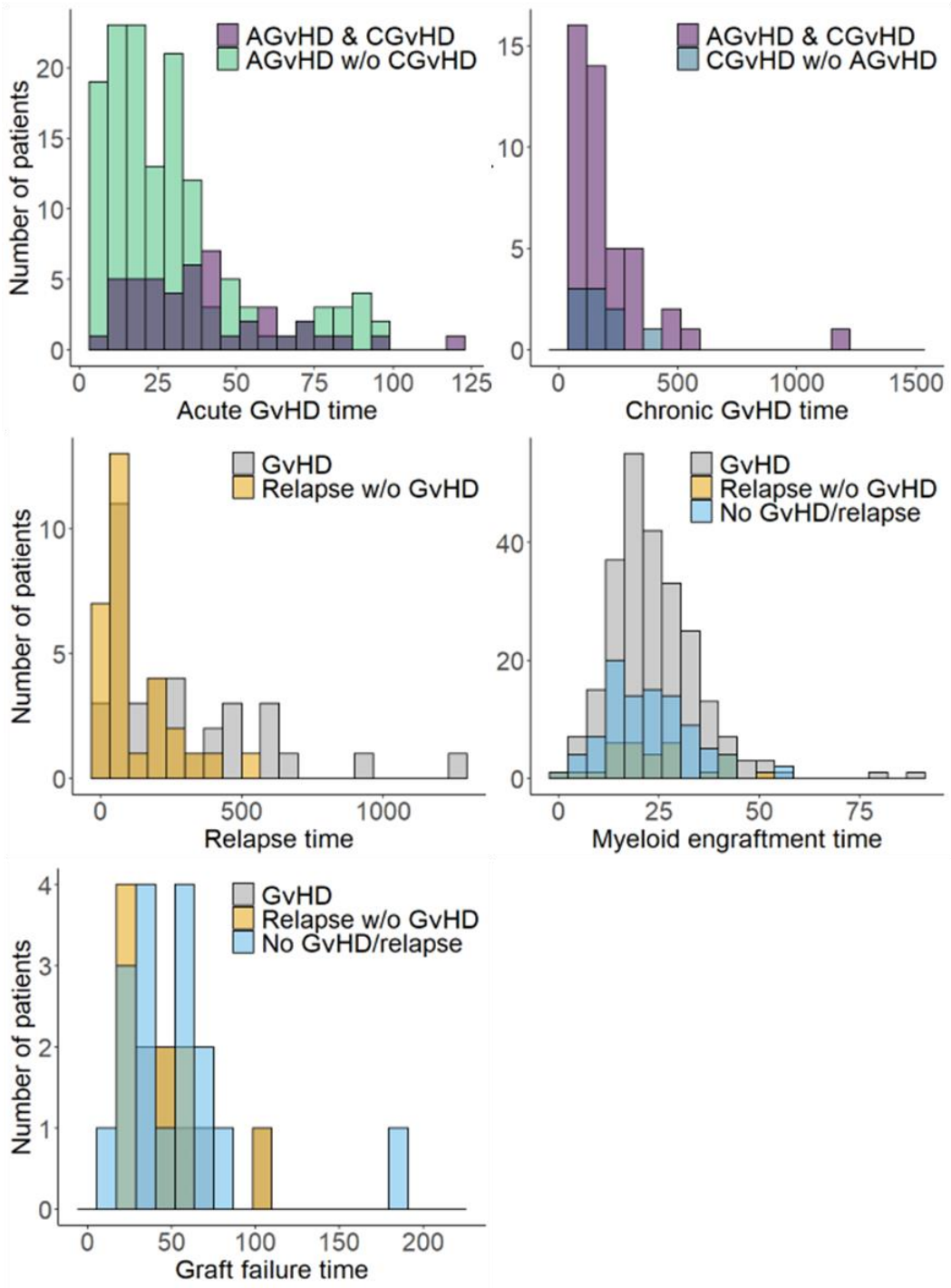
would be straight-forward). Therefore, estimates of state occupation probabilities were calculated from the "stacked" imputed dataset, *i.e.* by combining all imputed datasets and performing one analysis (see Chapter 3, Section 3.4.3). State occupation probability estimates are presented graphically (for illustration purposes), without SE. However, SE could be calculated using the method suggested by Beesley and Taylor (119), if desired. The test of the PH assumption (see Section 7.2) was also performed using the stacked dataset.

Analysis results were compared with CCA. Here, CCA means that each record was complete with respect to both event times and covariate data.

7.5. Investigating the distribution of event times for the NHS CBB dataset

The distribution of observed event times is shown in Figure 7.1, overleaf, for each event type and each sub-group of patients in the NHS CBB dataset who experienced that event. For times of relapse, myeloid engraftment and graft failure, patients who experienced any type of GvHD were grouped together. This was because event times were similar, regardless of the type of GvHD experienced, and/or due to small numbers of patients experiencing the event. Differences in the distributions of event times are described below. It is difficult to draw strong conclusions due to the small numbers in some patient sub-groups.

Figure 7.1. Distribution of event times for the NHS Cord Blood Bank cohort, by event type.



Bar colours represent patients experiencing different combinations of events.
Note that colours may be overlaid.

Patients who experienced both acute and chronic GvHD had slightly longer times to acute and chronic GvHD than patients who experienced only one of acute GvHD or chronic GvHD (comparing the acute GvHD time for patients experiencing acute and chronic GvHD vs. acute GvHD only: median 36 days, IQR 30 days vs. median 24 days, IQR 22 days, respectively; comparing the

chronic GvHD time for patients experiencing acute and chronic GvHD vs. chronic GvHD only: median 159 days, IQR 133 days vs. median 183 days, IQR 103 days, respectively).

Similarly, patients who experienced any type of GvHD had longer times to relapse than patients who experienced relapse without GvHD (comparing the relapse time for patients who experienced GvHD vs. neither relapse nor GvHD: median 213 days, IQR 346 days vs. median 76 days, IQR 145 days, respectively).

There were also differences in the distribution of graft failure times (comparing the graft failure time for patients experiencing GvHD vs. relapse only vs. neither GvHD nor relapse: median 50 days, IQR 24 days vs. median 44 days, IQR 34 days vs. median 48 days, IQR 31 days).

Times of myeloid engraftment were similar for all patient sub-groups (comparing the engraftment time for patients experiencing GvHD vs. relapse only vs. neither GvHD nor relapse: median 22 days, IQR 12 days vs. median 23 days, IQR 12 days vs. median 22 days, IQR 15 days).

In summary, there are (generally small) differences in the distributions of event times across the patient sub-groups. Hence, applying PMM for each sub-group (PMMSUBGP) may reduce bias and SE in model estimates. However, this must be balanced with the requirement for sub-groups to be of sufficient size to allow for random donor selection in the imputation procedure. Therefore, in this analysis, MI FCS using the PMMSUBGP method was applied, using the following sub-groups:

(i) Patients experiencing both acute and chronic GvHD, or chronic GvHD without acute GvHD (N=82). For this group, indicators of chronic GvHD, graft failure and myeloid engraftment were excluded from the imputation model because all/nearly all patients had the same indicator values. Time of graft failure was also excluded because only one patient experienced

graft failure.

(ii) Patients experiencing acute GvHD without chronic GvHD (N=173). For reasons as above, indicators of acute and chronic GvHD and graft failure were excluded from the imputation model for this group.

(iii) Patients experiencing relapse without GvHD, or neither relapse nor GvHD (N=177). For reasons as above, indicators of acute and chronic GvHD and graft failure were excluded from the imputation model for this group.

For the sensitivity analyses described previously (sensitivity analysis 1 and 2, see Section 7.4.2), imputation was performed using the PMMSUBGP method.
To explore whether applying PMM by patient sub-group (PMMSUBGP) changed the results compared with applying PMM for all patients (PMM), an additional sensitivity analysis (sensitivity analysis 3) was performed, in which PMM was applied for all patients in one imputation model. Plots of state occupation probability functions are for illustration only and hence, are presented for the main analysis but not the sensitivity analyses.

7.6. Results: Cumulative incidence and percentile times of the events of interest

Table 7.1 overleaf shows estimates and 95% CI of: the cumulative incidence at 100 days and median time of myeloid engraftment and acute GvHD; the cumulative incidence at one year and 10th percentile time of chronic GvHD and relapse; and the percentage of patients surviving at one year post-transplant and lower quartile time of death. Results are shown for the following missing data methods: CCA, PMMSUBGP (PMM by patient sub-group), PMMSUBGP proxy method (sensitivity analysis 2: PMMSUBGP with observed donor information used in imputation models for all other variables) and PMM (sensitivity analysis 3: PMM with one imputation model fit for all patients). Note that sensitivity

analysis 1 (fitting Weibull rather than Cox transition intensity models) only applies to the MSM analysis, so is not included here.

Only 116 (27%) patients in the NHS CBB dataset had complete data. Therefore, 80 imputations were performed for each MI analysis. As shown in Table 7.1, CCA estimates of the cumulative incidence of myeloid engraftment, acute and chronic GvHD, and overall survival, were lower than any of the MI estimates. The CCA estimate of the cumulative incidence of relapse was higher than any of the MI estimates. In particular, the CCA point estimate of acute GvHD was below the 95% CI of all the MI estimates. The CCA estimate of the median time of myeloid engraftment was the same as the MI estimates, although with wider CI. CCA estimates of the 10^{th} percentile time of chronic GvHD and relapse were higher than any of the MI estimates. In particular, the CCA point estimate of the 10^{th} percentile time of chronic GvHD was above the 95% CI for all the MI estimates. The CCA estimate of the lower quartile time of death was lower than MI estimates.

Table 7.1. Estimate (Est) and 95% confidence interval (CI) of the cumulative incidence and percentile times of myeloid engraftment, acute GvHD, chronic GvHD, relapse and overall survival.

Event type	Estimand	Missing data method							
		CCA (N=116)		PMMSUBGP (N=432)		PMMSUBGP proxy method (N=432)		PMM (N=432)	
		Est	95% CI	Est	95% CI	Est	95% CI	Est	95% CI
Myeloid engraftment	Cumulative incidence at 100 days (%)	84.5	77.9-91.1	86.3	83.1-89.6	86.3	83.1-89.6	86.3	83.1-89.6
	Median time (days)	24	22-26	24	23-25	24	23-25	24	23-25
Acute GvHD	Cumulative incidence at 100 days (%)	48.3	39.2-57.4	54.4	49.5-59.3	54.6	49.8-59.5	54.4	49.7-59.2
	Median time (days)	n/a*	n/a*	78	59-97	77	59-96	76	52-99
Chronic GvHD	Cumulative incidence at one year (%)	14.3	7.8-20.8	18.5	14.3-22.7	18.4	14.3-22.4	17.2	13.4-21.0
	10^{th} percentile time (days)	215	164-266	165	115-214	166	125-207	161	107-214
Relapse	Cumulative incidence at one year (%)	19.4	12.1-26.8	17.4	13.8-21.2	17.4	13.7-21.1	17.5	13.8-21.2
	10^{th} percentile time (days)	96	0-241	90	46-134	88	53-123	91	49-133
Overall survival	Survival at one year (%)	54.5	46.1-64.5	59.3	54.5-64.1	59.3	54.5-64.1	59.3	54.5-64.1
	Lower quartile time of death (days)	102	75-129	108	74-142	108	74-142	108	74-142

* For this method, < 50% patients experienced acute GvHD so the median could not be estimated.
95% CI lower bounds < 0 are truncated at 0.
CCA, complete case analysis.
PMMSUBGP, FCS MI by type 1 predictive mean matching with imputation models fit separately for patients experiencing both acute and chronic GvHD or chronic GvHD without acute GvHD, acute GvHD without chronic GvHD, relapse without GvHD, and neither GvHD nor relapse.
PMMSUBGP proxy method, as for PMMSUBGP, with observed donor information used in imputation models for all other variables.
PMM, FCS MI by Type 1 predictive mean matching with one imputation model fit for all patients.

Generally, 95% CI for CCA estimates were far wider than for any of the MI estimates. However, the CI for the lower quartile time of death was the exception: the CI for the CCA estimate was narrower than the CI for the MI estimates (CCA: 75-129 days vs. MI: 74-142 days). These CIs were estimated manually using my SE estimator (see Chapter 5, Section 5.3). CCA CIs calculated using an alternative method (by linear interpolation using the 'survfit' R package) were wider: 82-189 days. In CCA, less than 50% patients experienced acute GvHD so the median time could not be estimated.

For all estimates, there was little difference between results using the three different MI methods. Results for myeloid engraftment and overall survival were identical for all three MI methods because there were very few missing event times (no missing times of death) for these events. Of the MI methods considered, CIs were widest for the PMM method and narrowest for the PMMSUBGP proxy method.

Estimates from the main analysis (MI using PMMSUBGP) are summarised as follows:

- Most patients achieved myeloid engraftment (86.3%, 95% CI 83.1-89.6%), with median time of engraftment of 24 days (95% CI 23-25 days)
- 54.4% (95% CI 49.5-59.3%) of patients experienced acute GvHD by day 100 post-transplant, with median time of acute GvHD of 78 days (95% CI 59-97 days)
- 18.5% (95% CI 14.3-22.7%) of patients experienced chronic GvHD by one year post-transplant, with 10^{th} percentile time of chronic GvHD of 165 days (95% CI 115-214 days)
- 17.4% (95% CI 13.8-21.2%) of patients experienced relapse by one year post-transplant, with 10^{th} percentile time of relapse of 90 days (95% CI 46-134 days)
- 59.3% of patients were alive at one year post-transplant (95% CI 54.5-64.1%), with lower quartile time of death of 108 days (95% CI 74-142 days).

7.7. Results: Multi-state model analysis

7.7.1. Transition intensity models

Tables 7.2a-c show estimated HRs for all covariates in each (Markov) transition intensity model. The HR for time to acute GvHD from a separate semi-Markov model (included as a test of the Markov assumption) is also shown in Table 7.2c. Results are shown for the following missing data methods: CCA, PMMSUBGP (PMM by patient sub-group), PMMSUBGP Weibull (sensitivity analysis 1: PMMSUBGP, fitting Weibull rather than Cox transition intensity models), PMMSUBGP proxy method (sensitivity analysis 2: PMMSUBGP with observed donor information used in imputation models for all other variables), and PMM (sensitivity analysis 3: PMM with one imputation model fit for all patients).

As per the cumulative incidence analysis described in Section 7.6, CI are widest for CCA estimates and overlap CI for MI estimates. For all estimands, results using the four different MI methods are more similar to each other than to CCA estimates. In particular, in the model for the transition from transplant to acute GvHD, the CCA point estimates of the HR for double cord transplantation, for patient age, and for TNC dose at infusion, were outside the 95% CI for PMMSUBGP estimates. In the model for the transition from transplant to relapse/death, the CCA point estimates of the HR for disease type, and disease status at transplant, were also outside the 95% CI for PMMSUBGP estimates. In the model for the transition from acute GvHD to relapse/death, most CCA point estimates of HRs were outside the 95% CI for PMMSUBGP estimates. However, conclusions about the covariates most strongly associated with time of each transition were similar for CCA and MI. Weibull model (PMMSUBGP Weibull) estimates were similar to Cox model (PMMSUBGP) estimates. However, for the Weibull model for the transition from acute GvHD to relapse/death, the optimisation algorithm used to estimate parameters and SE failed to converge for 29 of the 80 imputed datasets.

Table 7.2a. Hazard ratios (HR) and 95% confidence interval (CI) for covariates in the transition intensity model from transplant to acute GvHD.

Covariate (reference value)	**Missing data method**									
	CCA (N=116)		**PMMSUBGP (N=432)**		**PMMSUBGP Weibull (N=432)**		**PMMSUBGP proxy method (N=432)**		**PMM (N=432)**	
	HR	**95% CI**	**HR**	**95% CI**	**HR**	**95% CI**	**HR**	**95% CI**	**HR**	**95% CI**
Double cord transplant (single)	0.26	0.06-1.11	0.90	0.56-1.46	0.82	0.51-1.34	0.87	0.54-1.40	0.65	0.39-1.08
Patient age (10-year increments)	1.02	0.80-1.31	0.89	0.80-1.00	0.93	0.83-1.04	0.91	0.81-1.01	0.93	0.83-1.04
Disease type (acute leukaemia)										
Other blood cancer[1]	0.66	0.27-1.64	0.89	0.61-1.31	0.84	0.57-1.24	0.88	0.60-1.29	0.88	0.59-1.31
Non-malignant disorder[2]	0.29	0.08-1.13	0.48	0.26-0.89	0.36	0.19-0.68	0.48	0.26-0.87	0.50	0.26-0.95
Disease status at time of transplant (in remission)										
Relapse	0.69	0.23-2.01	0.43	0.23-0.82	0.48	0.25-0.94	0.38	0.21-0.69	0.45	0.23-0.89
Other[3]	1.50	0.49-4.56	1.25	0.75-2.07	1.27	0.75-2.15	1.19	0.73-1.93	1.20	0.70-2.06
Reduced intensity conditioning regimen (intensive)	1.86	0.88-3.93	1.36	0.96-1.93	1.39	0.98-1.99	1.39	0.98-1.96	1.38	0.97-1.96
Donor-recipient CMV match (-/-)										
-/+	1.40	0.65-3.00	1.27	0.82-1.97	1.45	0.92-2.30	1.27	0.82-1.98	1.36	0.87-2.11
+/-	1.45	0.64-3.29	1.29	0.83-2.01	1.49	0.93-2.39	1.39	0.90-2.15	1.46	0.95-2.27
+/+	0.69	0.27-1.75	0.82	0.46-1.45	0.78	0.43-1.40	0.84	0.47-1.49	0.62	0.33-1.17
Donor-recipient sex match (F/F)										
F/M	1.52	0.70-3.31	1.31	0.82-2.09	1.27	0.79-2.05	1.19	0.75-1.90	1.29	0.81-2.06
M/F	1.22	0.49-3.00	1.18	0.72-1.93	1.09	0.67-1.79	1.14	0.70-1.87	1.26	0.77-2.05
M/M	1.12	0.48-2.62	1.10	0.66-1.82	1.06	0.63-1.76	1.08	0.66-1.77	1.09	0.66-1.81

Covariate (reference value)	Missing data method									
	CCA (N=116)		PMMSUBGP (N=432)		PMMSUBGP Weibull (N=432)		PMMSUBGP proxy method (N=432)		PMM (N=432)	
	HR	95% CI	HR	95% CI	HR	95% CI	HR	95% CI	HR	95% CI
Number of donor-recipient HLA mismatches[4] (Well-matched: 0/1)										
Not well-matched: 2 or more	1.42	0.76-2.66	1.11	0.75-1.63	1.20	0.79-1.82	1.10	0.74-1.63	1.33	0.88-2.01
TNC dose at infusion $\times 10^7$/kg (Low: <3.0)										
Medium: 3.0-5.0	2.61	1.22-5.57	1.02	0.69-1.53	1.24	0.82-1.88	1.18	0.75-1.87	1.25	0.78-2.01
High: > 5.0	1.46	0.61-3.47	0.84	0.54-1.32	1.01	0.64-1.58	1.01	0.62-1.62	1.15	0.74-1.79

CMV, cytomegalovirus; HLA, human leucocyte antigen; TNC, total nucleated cells.
Unless otherwise stated, Cox transition intensity models were fitted.
CCA, complete case analysis.
PMMSUBGP, FCS MI by type 1 predictive mean matching with imputation models fit separately for patients experiencing both acute and chronic GvHD or chronic GvHD without acute GvHD, acute GvHD without chronic GvHD, relapse without GvHD, and neither GvHD nor relapse.
PMMSUBGP proxy method, as for PMMSUBGP, with observed donor information used in imputation models for all other variables.
PMM, FCS MI by Type 1 predictive mean matching with one imputation model fit for all patients.
[1] Other blood cancer includes lymphoproliferative and plasma cell disorders, myelodysplastic syndromes and myeloproliferative disorders.
[2] Non-malignant disorder includes histiocytic disorder, solid tumour, bone marrow failure syndrome, haemoglobinopathy, primary immune deficiency and inborn error of metabolism.
[3] Other disease status includes acute, chronic and accelerated phase, refractory disease, transformed to acute leukaemia, blastic crisis, MDS, MDP and non-malignant disorders.
[4] HLA A and B loci at antigenic level and DR-B1 at allelic level

Table 7.2b. Hazard ratios (HR) and 95% confidence interval (CI) for covariates in the transition intensity model from transplant to relapse/death.

Covariate (reference value)	**Missing data method**									
	CCA (N=116)		**PMMSUBGP (N=432)**		**PMMSUBGP Weibull (N=432)**		**PMMSUBGP proxy method (N=432)**		**PMM (N=432)**	
	HR	**95% CI**	**HR**	**95% CI**	**HR**	**95% CI**	**HR**	**95% CI**	**HR**	**95% CI**
Double cord transplant (single)	1.24	0.20-7.81	1.61	0.79-3.28	1.44	0.70-2.98	1.61	0.79-3.25	1.97	0.89-4.40
Patient age (10-year increments)	1.01	0.71-1.43	0.94	0.79-1.10	1.00	0.84-1.18	0.94	0.79-1.11	0.92	0.78-1.09
Disease type (acute leukaemia)										
Other blood cancer[1]	0.42	0.11-1.68	0.84	0.46-1.55	0.76	0.42-1.39	0.86	0.48-1.54	0.73	0.39-1.35
Non-malignant disorder[2]	0.17	0.03-1.15	0.42	0.15-1.14	0.38	0.14-1.04	0.45	0.16-1.26	0.43	0.15-1.26
Disease status at time of transplant (in remission)										
Relapse	5.92	1.51-23.17	2.22	1.11-4.48	2.63	1.31-5.30	1.83	1.01-3.32	3.31	1.67-6.58
Other[3]	3.17	0.61-16.59	1.72	0.77-3.84	1.73	0.77-3.88	1.51	0.68-3.32	1.86	0.77-4.49
Reduced intensity conditioning regimen (intensive)	0.67	0.23-2.00	0.93	0.56-1.54	0.94	0.57-1.57	0.94	0.56-1.59	0.89	0.52-1.51
Donor-recipient CMV match (-/-)										
-/+	1.30	0.42-4.02	1.60	0.74-3.44	1.71	0.78-3.73	1.47	0.66-3.27	1.32	0.58-3.03
+/-	0.81	0.25-2.68	1.15	0.51-2.57	1.34	0.59-3.04	1.09	0.47-2.51	0.77	0.32-1.85
+/+	2.15	0.66-6.99	2.63	1.21-5.70	2.41	1.12-5.20	2.24	0.97-5.15	2.35	1.04-5.31
Donor-recipient sex match (F/F)										
F/M	0.44	0.15-1.28	0.88	0.43-1.78	0.85	0.41-1.77	0.87	0.42-1.80	0.73	0.35-1.53
M/F	0.52	0.16-1.71	0.97	0.47-2.01	0.90	0.43-1.90	0.99	0.48-2.08	0.86	0.41-1.80
M/M	0.40	0.13-1.27	0.92	0.41-2.06	0.90	0.39-2.07	0.93	0.41-2.11	0.85	0.37-1.95

Covariate (reference value)	Missing data method									
	CCA (N=116)		PMMSUBGP (N=432)		PMMSUBGP Weibull (N=432)		PMMSUBGP proxy method (N=432)		PMM (N=432)	
	HR	95% CI	HR	95% CI	HR	95% CI	HR	95% CI	HR	95% CI
Number of donor-recipient HLA mismatches[4] (Well-matched: 0/1)										
Not well-matched: 2 or more	1.50	0.61-3.67	1.49	0.80-2.76	1.63	0.87-3.04	1.42	0.74-2.71	1.65	0.87-3.13
TNC dose at infusion $\times 10^7$/kg (Low: <3.0)										
Medium: 3.0-5.0	1.88	0.63-5.63	1.06	0.55-2.06	1.18	0.61-2.31	1.10	0.57-2.09	1.14	0.56-2.31
High: > 5.0	1.16	0.41-3.25	1.06	0.54-2.06	1.22	0.61-2.44	1.01	0.50-2.02	0.86	0.43-1.72

CMV, cytomegalovirus; HLA, human leucocyte antigen; TNC, total nucleated cells.
Unless otherwise stated, Cox transition intensity models were fitted.
CCA, complete case analysis.
PMMSUBGP, FCS MI by type 1 predictive mean matching with imputation models fit separately for patients experiencing both acute and chronic GvHD or chronic GvHD without acute GvHD, acute GvHD without chronic GvHD, relapse without GvHD, and neither GvHD nor relapse.
PMMSUBGP proxy method, as for PMMSUBGP, with observed donor information used in imputation models for all other variables.
PMM, FCS MI by Type 1 predictive mean matching with one imputation model fit for all patients.
[1] Other blood cancer includes lymphoproliferative and plasma cell disorders, myelodysplastic syndromes and myeloproliferative disorders.
[2] Non-malignant disorder includes histiocytic disorder, solid tumour, bone marrow failure syndrome, haemoglobinopathy, primary immune deficiency and inborn error of metabolism.
[3] Other disease status includes acute, chronic and accelerated phase, refractory disease, transformed to acute leukaemia, blastic crisis, MDS, MDP and non-malignant disorders.
[4] HLA A and B loci at antigenic level and DR-B1 at allelic level

Table 7.2c. Hazard ratios (HR) and 95% confidence interval (CI) for covariates in the transition intensity model from acute GvHD to relapse/death.

Covariate (reference value)	**Missing data method**									
	CCA (N=116)		**PMMSUBGP (N=432)**		**PMMSUBGP Weibull (N=432)**		**PMMSUBGP proxy method (N=432)**		**PMM (N=432)**	
	HR	**95% CI**	**HR**	**95% CI**	**HR**	**95% CI**	**HR**	**95% CI**	**HR**	**95% CI**
Double cord transplant (single)	0.18	0.01-2.23	0.42	0.20-0.88	0.39	0.19-0.82	0.39	0.18-0.83	0.56	0.25-1.26
Patient age (10-year increments)	1.61	1.03-2.50	1.10	0.92-1.31	1.11	0.94-1.32	1.16	0.98-1.37	1.15	0.95-1.38
Disease type (acute leukaemia)										
Other blood cancer[1]	2.09	0.47-9.29	1.09	0.59-2.03	1.14	0.61-2.11	1.04	0.57-1.88	0.92	0.50-1.70
Non-malignant disorder[2]	3.89	0.45-33.42	1.37	0.51-3.69	1.38	0.52-3.66	1.36	0.51-3.63	1.07	0.39-2.90
Disease status at time of transplant (in remission)										
Relapse	2.40	0.54-10.63	1.03	0.38-2.81	0.92	0.36-2.37	1.43	0.57-3.59	2.39	0.86-6.65
Other[3]	1.29	0.19-8.64	1.14	0.50-2.57	1.03	0.47-2.27	1.29	0.60-2.77	1.63	0.70-3.75
Reduced intensity conditioning regimen (intensive)	0.40	0.09-1.85	0.99	0.54-1.83	0.99	0.54-1.81	0.86	0.48-1.55	0.78	0.44-1.41
Donor-recipient CMV match (-/-)										
-/+	0.86	0.29-2.57	1.90	0.94-3.82	1.93	0.97-3.85	1.71	0.85-3.47	1.68	0.81-3.47
+/-	1.11	0.31-3.92	1.29	0.62-2.71	1.27	0.61-2.62	1.33	0.63-2.79	1.33	0.62-2.86
+/+	0.52	0.11-2.46	0.80	0.29-2.24	0.84	0.31-2.26	0.74	0.26-2.09	1.01	0.36-2.87
Donor-recipient sex match (F/F)										
F/M	3.45	0.89-13.31	1.56	0.72-3.38	1.48	0.69-3.16	1.65	0.77-3.51	1.52	0.67-3.44
M/F	3.24	0.72-14.70	1.46	0.63-3.40	1.46	0.64-3.34	1.49	0.64-3.48	1.34	0.55-3.23
M/M	1.67	0.31-8.90	1.49	0.62-3.55	1.43	0.60-3.45	1.36	0.57-3.27	1.49	0.61-3.59

Covariate (reference value)	Missing data method									
	CCA (N=116)		PMMSUBGP (N=432)		PMMSUBGP Weibull (N=432)		PMMSUBGP proxy method (N=432)		PMM (N=432)	
	HR	95% CI	HR	95% CI	HR	95% CI	HR	95% CI	HR	95% CI
Number of donor-recipient HLA mismatches[4] (Well-matched: 0/1)										
Not well-matched: 2 or more	1.93	0.66-5.63	1.76	0.92-3.37	1.92	1.03-3.59	1.73	0.87-3.42	1.52	0.79-2.90
TNC dose at infusion $\times 10^7$/kg (Low: <3.0)										
Medium: 3.0-5.0	0.82	0.24-2.74	0.62	0.36-1.08	0.65	0.38-1.11	0.79	0.45-1.41	0.80	0.41-1.58
High: > 5.0	0.60	0.08-4.52	0.32	0.15-0.70	0.35	0.17-0.72	0.41	0.19-0.90	0.49	0.22-1.09
Model testing Markov assumption (for other covariate results, see Appendices, Table A.5)										
Time of acute GvHD	1.01	0.99-1.03	1.00	0.99-1.01	1.00	0.99-1.01	1.00	0.99-1.01	1.00	0.99-1.01

CMV, cytomegalovirus; HLA, human leucocyte antigen; TNC, total nucleated cells.
Unless otherwise stated, Cox transition intensity models were fitted.
CCA, complete case analysis.
PMMSUBGP, FCS MI by type 1 predictive mean matching with imputation models fit separately for patients experiencing both acute and chronic GvHD or chronic GvHD without acute GvHD, acute GvHD without chronic GvHD, relapse without GvHD, and neither GvHD nor relapse.
PMMSUBGP proxy method, as for PMMSUBGP, with observed donor information used in imputation models for all other variables.
PMM, FCS MI by Type 1 predictive mean matching with one imputation model fit for all patients.
[1] Other blood cancer includes lymphoproliferative and plasma cell disorders, myelodysplastic syndromes and myeloproliferative disorders.
[2] Non-malignant disorder includes histiocytic disorder, solid tumour, bone marrow failure syndrome, haemoglobinopathy, primary immune deficiency and inborn error of metabolism.
[3] Other disease status includes acute, chronic and accelerated phase, refractory disease, transformed to acute leukaemia, blastic crisis, MDS, MDP and non-malignant disorders.
[4] HLA A and B loci at antigenic level and DR-B1 at allelic level

Results from the main transition intensity analysis (MI using PMMSUBGP) are summarised below.

Transition from transplant to acute GvHD

The covariates most strongly associated with hazard of acute GvHD are disease type and disease status at time of transplant (Table 7.2a). The hazard of acute GvHD is more than halved for patients with a non-malignant disease compared with patients with acute leukaemia (HR 0.48, 95% CI 0.26-0.89), and for patients in relapse at time of transplant, compared with patients in remission (HR 0.43, 95% CI 0.23-0.82).

Transitions to relapse/death

The covariates most strongly associated with hazard of relapse/death without acute GvHD are disease status at time of transplant and donor-recipient CMV match (Table 7.2b). For patients in relapse at time of transplant, the hazard of relapse/death is more than doubled compared with patients in remission (HR 2.22, 95% CI 1.11-4.48). The hazard of relapse/death when both donor and recipient are CMV+ is more than double that when both donor and recipient are CMV- (HR 2.63, 95% CI 1.21-5.70). These associations are less apparent in the model for the time of transition from acute GvHD to relapse/ death (HR 1.03, 95% CI 0.38-2.81 for a patient in relapse at time of transplant, and HR 0.80, 95% CI 0.29-2.24 when both donor and recipient are CMV+).

The covariates most strongly associated with the hazard of relapse/death after acute GvHD are the number of CB units received and TNC dose at infusion. The hazard of relapse/death is more than halved for patients receiving a double cord transplant compared with a single cord transplant (HR 0.42, 95% CI 0.20-0.88), and for patients receiving a high TNC dose rather than a low TNC dose (HR 0.32, 95% CI 0.15-0.70).

There was no apparent association between time of acute GvHD and the hazard

of relapse/death after acute GvHD (HR 1.00, 95% CI 0.99-1.01), suggesting there was no violation of the Markov assumption. There was some indication of a violation of the PH assumption, particularly for the model for the transition from transplant to relapse/death (global PH test p-value = 0.14, 0.03, 0.30 for the transitions from transplant to acute GvHD, transplant to relapse/death and acute GvHD to relapse/death, respectively).

7.7.2. Expected length of stay in each state

The expected length of stay in state in the first year post-transplant (RELOS) was calculated for three different patient types. These were: a patient with reference values of all covariates, a low-risk, and a high-risk patient. Based on the strongest associations described above for the transition intensity models, low-risk and high-risk patients are defined as follows:

(i) Low-risk: A patient with a non-malignant disorder, receiving a high TNC dose. Based on the NHS CBB cohort, approximately 14% of patients had these characteristics. Most were children (with a mean age of 4 years), which is to be expected, because patients with non-malignant disorders tend to be treated at a young age (4). These patients had a range of values of other covariates, but most received a single cord transplant and had 0 or 1 HLA mismatches. For simplicity, reference values of all other covariates and the mean age of 4 years were used when calculating RELOS for this patient type.

(ii) High-risk: A patient in relapse at time of transplant, receiving a double cord transplant, with 2 or more HLA mismatches. Although a double cord transplant was associated with a lower hazard of relapse/death after acute GvHD, most patients in relapse at time of transplant in the NHS CBB dataset were adults (adults tend to receive a double cord transplant, see Chapter 4, Section 4.3). Most patients in relapse at time of transplant also had 2 or more donor-recipient HLA mismatches. Furthermore, although a +/+ donor-recipient CMV match was also strongly associated

with the hazard of relapse/death, only one patient in the NHS CBB dataset (on average, based on the set of imputed datasets) had the other high-risk characteristics as well as a +/+ CMV match. Therefore, the reference value (-/- CMV match) for this variable was used instead. Reference values for all other covariates and the mean age of 44 years were used when calculating RELOS for this patient type.

Table 7.3 overleaf shows estimates and 95% CI for RELOS for the three patient types. Results are shown for the following missing data methods: CCA, PMMSUBGP (PMM by patient sub-group), PMMSUBGP Weibull (sensitivity analysis 1: PMMSUBGP, fitting Weibull rather than Cox transition intensity models), PMMSUBGP proxy method (sensitivity analysis 2: PMMSUBGP with observed donor information used in imputation models for all other variables), and PMM (sensitivity analysis 3: PMM with one imputation model fit for all patients).

Estimates vary across the different missing data methods. Generally, MI estimates are more similar than CCA estimates. In particular, CCA estimates of the time spent in relapse/death are higher than MI estimates. For all estimates, CI are wide, and widest for CCA estimates. CI are also wide when fitting a Weibull model (PMMSUBGP Weibull), which may be a consequence of the poor performance of the optimisation algorithm, described above.

Table 7.3. Estimate (Est) and 95% confidence interval (CI) of expected length of stay in each state in the first year post-transplant (RELOS).

RELOS (days)	Missing data method									
	CCA (N=116)		PMMSUBGP (N=432)		PMMSUBGP Weibull (N=432)		PMMSUBGP proxy method (N=432)		PMM (N=432)	
	Est	95% CI	Est	95% CI	Est	95% CI	Est	95% CI	Est	95% CI
Reference: Patient with reference values of covariates, age 24y										
Transplant	151	35-267	158	96-220	219	106-332	158	96-220	174	111-237
Acute GvHD	76	0-173	132	71-193	95	0-196	129	69-189	113	56-170
Relapse/death	123	0-271	68	23-113	51	0-130	70	26-114	70	22-118
Low-risk: Patient with non-malignant disorder, high TNC dose, age 4y*										
Transplant	273	0-365	233	160-306	291	212-370	222	147-297	229	151-307
Acute GvHD	38	0-365	94	25-163	54	0-123	100	30-170	95	24-166
Relapse/death	39	0-365	31	0-71	20	0-61	36	0-80	33	0-79
High-risk: Patient in relapse at transplant, receiving a double cord transplant, with 2 or more HLA mismatches, age 44y*										
Transplant	33	0-365	139	25-253	165	0-335	152	39-265	72	0-167
Acute GvHD	9	0-365	49	0-100	36	0-93	39	0-79	22	0-54
Relapse/death	309	0-365	170	45-295	163	0-360	166	45-287	262	155-369

*With reference values of all other covariates.
CMV, cytomegalovirus; HLA, human leucocyte antigen; TNC, total nucleated cells.
95% CI boundaries outside the range [0,365] were truncated to 0 and 365 for lower and upper bounds respectively.
Unless otherwise stated, Cox transition intensity models were fitted.
CCA, complete case analysis.
PMMSUBGP, FCS MI by type 1 predictive mean matching with imputation models fit separately for patients experiencing both acute and chronic GvHD or chronic GvHD without acute GvHD, acute GvHD without chronic GvHD, relapse without GvHD, and neither GvHD nor relapse.
PMMSUBGP proxy method, as for PMMSUBGP, with observed donor information used in imputation models for all other variables.
PMM, FCS MI by Type 1 predictive mean matching with one imputation model fit for all patients.

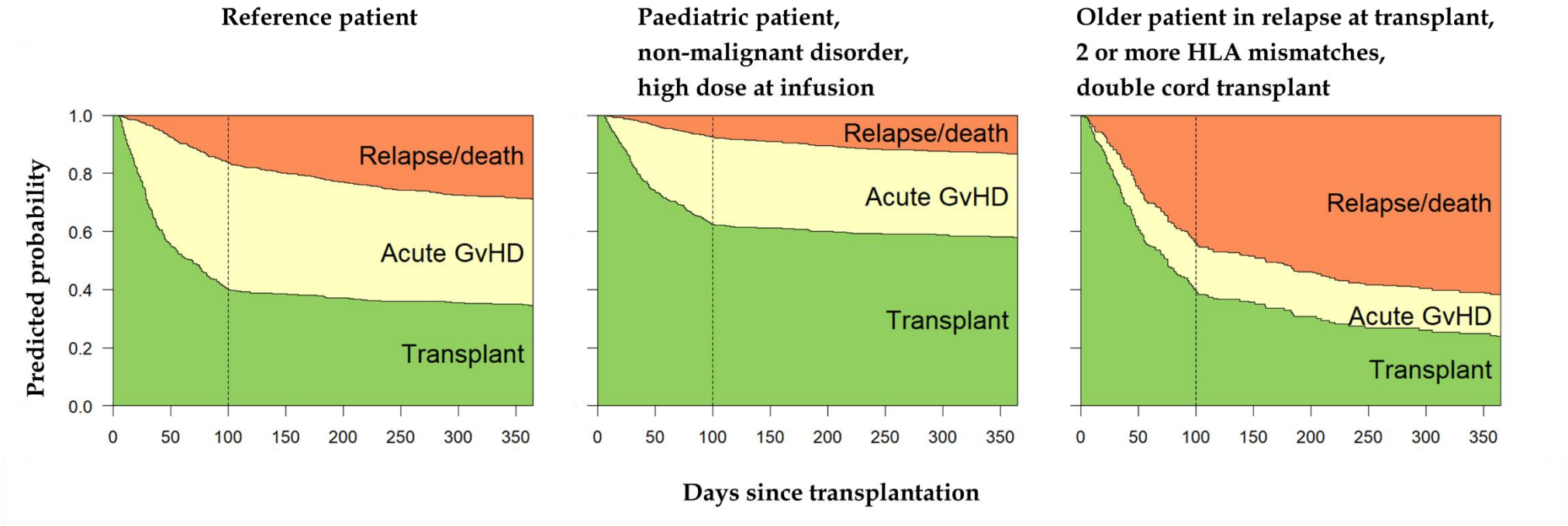
Estimates of RELOS from the main analysis (MI using PMMSUBGP) are summarised as follows: In the first year post-transplant, low-risk patients spend a larger number of days (233 days, 95% CI 160-306 days) alive without acute GvHD nor relapse than patients with reference values of covariates (158 days, 95% CI 96-220 days) and high-risk patients (139 days, 95% CI 25-253 days). Conversely, compared with the other patient types, high-risk patients spend fewer days alive without relapse after acute GvHD (49 days, 95% CI 0-100 days for high-risk patients vs. 132 days, 95% CI 71-193 days for reference patients, and 94 days, 95% CI 25-163 days for low-risk patients) and more days in relapse/death (170 days, 95% CI 45-295 days for high-risk patients vs. 68, 95% CI 23-113 days for reference patients, and 31 days, 95% CI 0-71 days for low-risk patients).

7.7.3. State occupation probabilities

State occupation probabilities during the first year post-transplant, for the three patient types described previously, are presented in Figure 7.2. State occupation probabilities are shown for the main analysis (MI using PMMSUBGP). Note that Cox models fitted using the stacked dataset give identical results (to 1 d.p.) to those in Tables 7.2a-c (which were calculated as the mean of the per-imputation results). Therefore, the state occupation probabilities presented here are consistent with the previous results from per-imputation analyses.

For a patient with reference values of all covariates, there is approximately the same probability of being alive without GvHD nor relapse, alive without relapse after acute GvHD or in relapse/death at one year post-transplant (Figure 7.2). At any point in the first year post-transplant, a low-risk patient is most likely to be alive without GvHD nor relapse, with a small probability of being in relapse/death (14% or less) throughout the first year. In contrast, beyond the acute period post-transplant (day 100 onwards), a high-risk patient is most likely to be in relapse/death. For high-risk patients, the probability of being alive without relapse after acute GvHD is at most 17% during the first year.

Figure 7.2. State occupation probabilities for each state in the first year post-transplant for three patient types in the NHS Cord Blood Bank cohort



Vertical line separates the acute and chronic post-transplant periods.

7.8. Discussion

In this chapter, I applied FCS MI methods to impute missing event times and covariate data in the NHS CBB dataset. Then, I analysed the imputed datasets using competing risks and MSM methods, to obtain estimates of the cumulative incidence of each event of interest (myeloid engraftment, acute and chronic GvHD, relapse, and overall survival), to identify covariates associated with the events of interest, and to make probability predictions for different patient types. In the sections below, I discuss the clinical and statistical implications of my results. The implications of my thesis for statisticians, researchers, clinicians, and patients are further discussed in Chapter 8, Sections 8.2-8.4.

7.8.1. Clinical implications

Cumulative incidence of the events of interest

In Table 7.4 overleaf, the estimated cumulative incidence rates for NHS CBB patients are compared with those reported in other CB transplantation studies (8, 14, 23, 37-40, 197, 198). CB donated to the NHS CBB is used to treat patients with both malignant and non-malignant blood diseases, as well as both adult and paediatric patients. Prognoses and treatment approaches vary between these patient groups. Hence, to allow a direct comparison with the NHS CBB, I have focused on studies of other CB banks, and general reviews of CB transplantation (rather than studies of a particular type of blood disease, or studies restricted to either an adult or paediatric patient group). As shown in Table 7.4, cumulative incidence rates for NHS CBB patients are similar to those reported in other CB transplantation studies. In particular, the cumulative incidence and median time of myeloid engraftment for the NHS CBB are very similar to those reported in other studies. This result will provide reassurance to patients and clinicians that transplantation using CB donated to the NHS CBB is a safe and effective treatment.

It is more difficult to compare cumulative incidence rates for the other events of interest, either because rates vary considerably between studies in the clinical literature (*e.g.* for acute GvHD), or because the time period of interest varies between studies (*e.g.* chronic GvHD, overall survival). Although the median time to engraftment was reported in all studies included in this comparison, percentile times of other events were rarely reported.

Table 7.4. Comparison of cumulative incidence rates for the NHS CBB vs. other CB transplantation studies.

Event type	**Estimand**	**NHS CBB (95% CI)**	**Other CB banks/ CB studies**
Myeloid engraftment	Cumulative incidence at 100 days	86% (83-90)	Varies between 85 and 100%; 68% in CML patients
	Median time (days)	24 (23-25)	Varies between 22 and 30 days
Acute GvHD	Cumulative incidence at 100 days	54% (50-59)	> 75% (any grade) in double cord recipients with chronic GvHD; 47% (95% CI 36-58) (grade 2-4) in CML[1] patients; 13% (95% CI 7-20) (grade 2-4) in paediatric patients;
	Median time (days)	78 (59-97)	Not reported
Chronic GvHD	Cumulative incidence at one year	19% (14-23)	23% (95% CI 11–34), 26% (95% CI 21-30) at 2 years in double cord recipients 7% (95% CI 4-11) at 2 years in single cord recipients
	10th percentile time (days)	165 (115-214)	Median of 152 days in double cord recipients with chronic GvHD; generally not reported
Relapse	Cumulative incidence at one year	17% (14-21)	<20% in AL/MDS patients 31% (95% CI 27-35) at 2 years in double cord recipients 29% (95% CI 22-36) at 2 years in single cord recipients
	10th percentile time (days)	90 (46-134)	Not reported
Overall survival	Survival at one year	59% (55-64)	In paediatric patients with blood cancer: 73% (95% CI 63-80) (single cord) and 65% (95% CI 56-74) (double cord); 56% (95% CI 51-60) at 2 years in double cord recipients 64% (95% CI 58-69) at 2 years in single cord recipients
	Lower quartile time of death (days)	108 (74-142)	Not reported

CML, chronic myeloid leukaemia; AL, acute leukaemia; MDS, myelodysplastic syndromes

Covariates associated with acute GvHD, relapse and death

In the NHS CBB analysis, the covariates most strongly associated with the transition intensity in the model for the transition from transplant to acute GvHD are disease type and disease status at time of transplant. The hazard of acute GvHD is more than halved for patients with a non-malignant disease compared with patients with acute leukaemia. This may indicate a different treatment approach for blood cancers compared with non-malignant disease, because some degree of acute GvHD is desirable in patients with blood cancers in order to achieve a GvL effect (see Chapter 1, Section 1.2.5) (10), but this is not as apparent for non-malignant disorders.

The hazard of acute GvHD is also more than halved for patients in relapse at time of transplant, compared with patients in remission. Conversely, the hazard of relapse/death without acute GvHD is more than doubled for patients in relapse at time of transplant, compared with patients in remission. This is to be expected because patients in relapse at time of transplant are less likely to achieve myeloid engraftment. Hence, GvHD is less likely and the hazard of post-transplant relapse/death is increased (19). In addition to disease status at time of transplant, donor-recipient CMV status match was also strongly associated with the hazard of relapse/death. The hazard of relapse/death when both donor and recipient are CMV+ is more than double that when both donor and recipient are CMV-. This is to be expected because a CMV+ patient status is known to be associated with increased mortality (14).

Finally, the covariates most strongly associated with the hazard of relapse/death after acute GvHD are the number of CB units received and TNC dose at infusion. The hazard of relapse/death after acute GvHD is more than halved for patients receiving a double cord transplant compared with a single cord transplant, and for patients receiving a high TNC dose rather than a low TNC dose. It is likely that these variables are correlated: a double cord transplant is likely to provide a higher dose than a single cord transplant (8). Myeloid engraftment, and hence

GvHD, are more likely and relapse is less likely when the TNC dose is at least 3 $\times 10^7$/kg (4).

In Table 4.5, overleaf, I compare the values associated with increased hazard (for each covariate and transition) in my analysis of the NHS CBB, with those reported in the literature (see Chapter 1, Table 1.1). Note that in Table 4.5 (as in my MSM design), I have assumed that values associated with increased hazard of relapse (as reported in the literature) are also associated with increased hazard of death (because relapse is associated with increased mortality (17, 20)). This assumption does not allow for different covariate associations for different causes of death (*e.g.* death from a transplant-related cause vs. death from relapse). The different covariate associations for transplant-related and relapse-related causes of death could be explored by extending the MSM used here to a four-state model, with two absorbing states: (i) relapse or death due to relapse, and (ii) death due to a transplant-related cause.

For the covariates most strongly associated with the transition intensities in my analysis (described above), the values associated with increased hazard of each event are consistent with the clinical literature (Table 4.5). However, my results were inconclusive for several covariate associations reported in the clinical literature. This was the case for patient age at transplant, conditioning regimen, and donor-recipient sex match. This may be explained by the relatively small size of the NHS CBB dataset combined with the relatively large number of covariates considered, which may have limited the power of my analysis. In addition, in my analysis, the number of donor-recipient HLA mismatches was only weakly associated with the transition intensities. In HSC transplantation in general, donor-recipient HLA matching is very important. However, compared with BM and PB transplantation, CB transplantation requires less stringent HLA matching (Chapter 1, Section 1.2.2). This may explain the weak association between HLA matching and NHS CBB patient outcomes. As above, the relatively small size of the NHS CBB dataset may also be the reason.

Table 7.5. Comparison of covariates associated with acute GvHD, relapse, and death for the NHS CBB vs. other CB transplantation studies.

Covariate	**Values associated with increased hazard of event**			
	Acute GvHD		**Relapse/death**	
	Other studies	**NHS CBB analysis in agreement?**	**Other studies**	**NHS CBB analysis in agreement?**
Number of CB units received	Double cord	Inconclusive results	Single cord	Yes (transition from acute GvHD) No – double cord (transition from transplant, weak association)
Disease status at time of transplant	In remission at time of transplant	Yes, compared with a patient in relapse at time of transplant	In relapse at time of transplant	Yes (transition from transplant) Inconclusive results (transition from acute GvHD)
Conditioning regimen	Intensive conditioning	Inconclusive results	Reduced intensity conditioning	Inconclusive results
TNC dose at infusion	Higher dose	Inconclusive results	Lower dose	Yes (transition from acute GvHD) Inconclusive results (transition from transplant)
GvHD prophylaxis (yes/no)	No prophylaxis	Not considered	Prophylaxis	Not considered
Disease type	Not considered	Acute leukaemia, compared with non-malignant disorder	Not considered	Inconclusive results
Donor-recipient HLA mismatch	2 or more mismatches	Yes (weak association)	2 or more mismatches	Yes (weak association)
Patient age at transplant	Adult patient	Inconclusive results	Adult patient	Inconclusive results
Donor-recipient sex match	Female donor and male recipient	Inconclusive results	Female donor and male recipient	Inconclusive results
Donor-recipient CMV+ match	CMV+ patient	Inconclusive results	CMV+ patient	Yes (transition from transplant) Yes (transition from acute GvHD, weak association)

Contrary to the findings of Lee *et al.* (160), for patients in the NHS CBB dataset there was no apparent association between the time of acute GvHD and the hazard of relapse/death after acute GvHD (based on a semi-Markov model). However, the prognosis after acute GvHD varies according to the grade (severity) (17). Hence, it may be that this association would become more apparent if the MSM used here was extended to a model with two intermediate states: (i) grade 1 (mild) acute GvHD, and (ii) grade 2-4 (moderate to very severe) acute GvHD. An alternative approach that avoids increasing the complexity of the model used here, would be to consider grade 2-4 acute GvHD as the only intermediate state (assuming that mild acute GvHD is unlikely to be related to subsequent patient outcomes).

7.8.2. Statistical implications

As discussed in Chapter 4, Section 4.5, in the NHS CBB dataset, the missingness of event times depended on the analysis outcome. That is, missingness depended on the type of event experienced, as well as patient, donor and transplant characteristics. In the competing risks analysis considered here, these characteristics were not part of the analysis model because the outcomes of interest were the non-parametric estimates of cumulative incidence of various events. In this case, CCA is biased because event times for the competing events are fully observed and only cases of the events of interest are under-represented. In contrast, estimates based on FCS MI should be unbiased (assuming data MAR, conditional on the observed data, and that the imputation model is correctly specified). In the NHS CBB analysis, CCA estimates of the cumulative incidence of myeloid engraftment, acute and chronic GvHD, and overall survival, were lower than any of the MI estimates. The CCA estimate of the cumulative incidence of relapse was higher than any of the MI estimates. For myeloid engraftment, (for which there were no missing event times, and few missing event times for competing events), the CCA estimate of the median time of myeloid engraftment was the same as the MI estimates. However, CCA CI were

wider than MI CI. These results suggest that application of FCS MI methods to the NHS CBB dataset has reduced bias and improved precision in estimates.

In the MSM analyses, many of the patient, donor and transplant characteristics were also covariates in the Cox regression analysis models for the transition intensities. In my analyses of the NHS CBB cohort, CCA estimates were often different from MI estimates. Of particular note, in the model for the transition from transplant to acute GvHD, the CCA estimates of the HR for a double cord transplant were outside the 95% CI for MI estimates. There was a greater proportion of double cord transplants for patients with missing times of acute GvHD than for patients with observed times of acute GvHD (Chapter 4, Section 4.5). This suggests that the under-representation of double cord transplants in the complete case cohort has led to biased results. Note that if missingness of all incomplete variables in the NHS CBB dataset had depended only on the covariates and not the outcome, and all variables predictive of missingness were included as covariates in the analysis model, CCA estimates would be unbiased (86).

Only 116 patients in the NHS CBB dataset had complete data and hence CI were far wider for CCA estimates than for MI estimates. Generally, CIs for MI estimates were widest for the PMM and PMMSUBGP method when fitting a Weibull model, and narrowest for the PMMSUBGP proxy method. For double CB transplants where one of the two donors did not donate via the NHS CBB (hence no data were available for this donor), covariate values that relied on both donors' data were missing (disease status, CMV+ match, number of HLA mismatches between donor(s) and recipient, and dose at infusion, each with between 29% and 56% missing data). In the PMMSUBGP proxy method, the known donor's data was used in imputation models to improve prediction of all other variables (although the missing covariate values that relied on both donors' data were still imputed). This suggests that known donor data can be used within imputation models to improve the precision of estimates.

Results from my simulation studies (Chapter 6, Section 6.2.16-17) suggest that bias in estimates and SE of RELOS will be reduced if a parametric, rather than a semi-parametric, model is used. However, a disadvantage of parametric models in practice is that they seem to be more prone to convergence problems than semi-parametric models. Further work is required to determine if this is also the case for flexible parametric models. There was some indication that the PH assumption did not hold for my transition intensity models, particularly for the model for the transition from transplant to relapse/death. Therefore, my models could be improved by considering hazards that vary over time for each covariate.

CHAPTER 8. DISCUSSION AND CONCLUSIONS

8.1. Introduction

The purpose of this thesis was to provide the first insight into patient outcomes after transplantation using CB donated to the NHS CBB. My research aims were:

(i) To describe the incidence of various post-transplant events (myeloid engraftment, acute and chronic GvHD, relapse, and death) among NHS CBB patients.

(ii) To identify patient, donor, and transplant characteristics associated with the events of acute GvHD, relapse, and death, and to describe the probability of these events for different patient types.

In the NHS CBB dataset, times of acute GvHD, chronic GvHD, and relapse were frequently missing. Missing event times were assumed to have occurred in a known, finite, time period (based on clinical criteria or the known patient follow-up period) and hence, the missing event times were considered interval-censored. Sun (47) suggests a simple method for handling interval-censored event times, substituting a single point from the time interval (usually the mid-point if there is no prior information about the part of the interval where the event is more likely to occur). However, in most circumstances it is advisable to use a more rigorous approach. Methods for handling interval-censored event times can be broadly categorised as either (i) treating the event times as missing data and applying MI strategies or (ii) taking a FML approach.

Based on the available literature, it was not clear which was the best method for handling interval-censored event times, particularly (as in my research) when the analysis is a competing risks model or MSM. Therefore, a substantial element of

this thesis has been focused on comparing methods for handling missing event times. Due to the limitations of the FML methods developed to date, it was not possible to treat FML methods on an equal footing with MI methods. I have focused on MI strategies, because the real data analysed here deviates from the assumptions of the FML methods developed so far, in two ways: (i) event times are a mixture of observed and missing times and (ii) the time interval boundaries are generally the same for all patients and are wide relative to the observed event times. In contrast, the appeal of MI is its flexibility: after imputation, any desired complete data method may be used, a mixture of observed and missing data can be accommodated, and additional data that are predictive of the missing times, but not required for the substantive analysis, can be used during the imputation step to inform the imputed times.

In my research, I have used simulation studies to compare MI and FML methods for handling interval-censored event times. For a competing risks analysis model, I considered the following MI methods: FCS MI by type 1 PMM; FCS MI by log-linear regression with post-imputation back-transformation; FCS MI by linear regression with and without restrictions on the imputed values; and Delord and Genin's method (136) based on sampling from the cumulative incidence function. I compared these MI methods with CCA and the FML method proposed by Bakoyannis *et al.* (64). I subsequently compared the best-performing MI methods (identified in the competing risks study), when the analysis model was a MSM. I found no advantage in using FML methods rather than MI to handle missing event times. Through this research, I have addressed key questions about MI of interval-censored event times. These questions, and the findings and recommendations from my research, are summarised in Table 8.1 overleaf.

Table 8.1. Summary of findings from this thesis comparing MI methods.

Research Question	**Findings from my research**
Is it necessary to constrain the imputed times to lie within the interval boundaries?	Not recommended. Restricting the range of imputed values during the imputation step leads to under-estimation of the SE.
Should skewness be accounted for in the imputation model?	Recommend FCS MI by type 1 PMM (sampling from a set of observed times) rather than transforming data (*e.g.* by imputing on the log scale) to account for skewness.
Are methods sensitive to incompatibility between the imputation scheme and the observed data distribution?	Violation of the constant variance assumption when imputing using a linear regression model led to biased analysis estimates. Recommend careful exploration of the distribution of event times by event type (competing risks analysis) or for each pathway through the MSM (MSM analysis). Recommend a separate imputation model for each sub-group with a different event time distribution.
Are methods sensitive to incompatibility between the imputation scheme and the analysis model?	Type 1 PMM was robust to imputation model misspecification. Imputation by sampling from the cumulative incidence function (competing risks analysis), or strategies to preserve the observed sequence of events (MSM analysis) were no better than standard PMM.
Are methods sensitive to the percentage of missing data?	Type 1 PMM performed well even when 50% of event times were MCAR or MAR.
Are methods sensitive to strength of the relationship between missingness and the observed data?	For type 1 PMM, results were similar for: • different DGMs • event times MCAR, MAR (dependent on event type) and MAR (dependent on event type and covariates) • event times missing for a single or multiple event types Recommend including all variables predictive of the missing event time in the imputation model (as well as the event type indicator and all analysis model covariates). Variables that are only predictive of missingness but not the event time itself should not be included.
How do methods perform given event times MNAR?	FCS MI assuming MAR is not recommended when event times are suspected to be MNAR. Estimates of quantiles of the cumulative incidence function, and expected length of stay in state, were particularly biased.

Note that, unless explicitly stated otherwise, these recommendations assume event times are MCAR or MAR, given the observed data.

In addition to comparing missing data strategies, I derived estimators of SE of (i) percentiles of event times (*e.g.* the median time of acute GvHD) and (ii) RELOS (the restricted expected length of stay in state). I used my simulation studies to assess the performance of these estimators. SE estimator (i) was based on the delta method. I identified the elements of this estimator that minimised bias and optimised coverage. Namely, that it was best to estimate the probability density function used in my definition by a simple gradient function, using a small increment (0.03 or less) around the estimated percentile time. For SE estimation scenario (ii) (estimators of SE of RELOS), I found that the best-performing estimators were the bootstrap estimator (when fitting Cox transition intensity models) and the delta estimator (when fitting Weibull transition intensity models).

Finally, I applied methods for handling missing data in an analysis of patient outcomes following CB transplantation. In the NHS CBB dataset, the missingness of event times depended on the type of event experienced, as well as patient, donor and transplant characteristics. In this case, CCA is biased because event times for the competing events are fully observed and only cases of the events of interest are under-represented. By using MI strategies to handle missing event times, I was able to avoid this source of bias and improve precision of my estimates. I made further gains in precision by using the known donor's data (for double cord transplants in which one set of donor data were missing) in imputation models to improve prediction of all other variables. I found that event rates for NHS CBB patients were comparable to those reported in other CB transplantation studies, and for other CB banks (14, 35, 38, 40). Using MSM analysis, I identified patient characteristics associated with low- and high-risk of post-transplant adverse events. In addition, I estimated the expected number of days spent after acute GvHD, and relapse or death, in the first year following CB transplantation. The implications of this research are explored below.

My research represents the first comparison of MI and FML methods for handling interval-censored event times, as well as the first time that MI methods have been compared in a MSM analysis. It also represents the first analysis of HSC transplantation outcomes using MI methods.

8.2. Implications for statisticians

A strength of this thesis was the use of simulation studies to assess the performance of MI methods. Although theoretical knowledge underpins the use of MI methods, this relies on correct specification of the imputation and analysis models. A benefit of simulation studies is that they can be used to explore the performance of methods in practice, when model assumptions may not hold. By changing various elements of the DGM, MDM, and imputation and analysis models, it is possible to gauge the performance of methods in different scenarios. In addition, their sensitivity to model misspecification can be assessed (75).

It is essential that the simulation study is carefully planned *a priori*. I recommend following the "ADEMP" structure (stating aims, data-generating mechanisms, estimands, methods, and performance measures) suggested by Morris *et al.* (75). In addition, it is vital to check that each element of the simulation design works as expected by first performing the simulation under well-understood, predictable conditions. To ensure that the large sample properties of simulated data are as expected, I recommend performing analysis on all simulated datasets combined, before calculating per-simulation results. For example, in my simulation studies, for consistency with my real data, I used medium-sized simulated datasets (N≈500). However, before obtaining results for each simulation, I first checked that estimates from all 1000 simulated datasets combined matched the true values used in my DGM (or the values I had calculated analytically). For simulation studies of missing data methods, I recommend first checking that full data estimates (*i.e.* without any missing data) are as expected. Then, I recommend checking that estimates when data are

MCAR are as expected (for example, CCA should always give unbiased results in this case). In my simulation study, through these checks, I realised that when event times were randomly missing, but this was conditional on event type (*i.e.* times were missing for one event type only), a CCA would be biased (because in this case, times are actually MAR because missingness depends on the analysis outcome).

In my research, simulation studies were particularly useful for assessing the performance of existing MI methods with a novel analysis model (MSM analysis), and for comparing the performance of estimators of SE. For example, knowledge of the DGM enabled me to identify causes of unexpected bias. By comparing results for event times rounded to 0.1 days and whole days, I determined that some estimates and SEs were sensitive to rounding of simulated event times (because the analytical methods I considered assumed that time was measured on a continuous time scale). Knowledge of the DGM also enabled me to identify a cause of the poor performance of a linear regression imputation model (due to violation of the constant variance assumption).

Rounded event times, and event times distributions that differ across patient sub-groups, are features of my real data. When performing simulation studies, it is important to achieve a balance between using DGMs and missingness scenarios that are representative of real data, while also ensuring that study results are generalisable (so that they are useful to other researchers). The NHS CBB dataset is a complex dataset, containing a large number of variables. It would not be easy, or practical, to generate simulated data from sufficiently complex distributions to re-create every element of the real data. I recommend identifying key elements of the real data that should be included in the simulation design. This can be achieved by careful exploration of the distribution of each analysis outcome and covariate, and its missingness, prior to performing the simulation study. For example, my research focused on missingness of event times, not covariate data. Therefore, in my simulation

studies, I considered various ways in which event times could be missing but assumed all covariate data were observed.

In my simulation studies, I used DGMs based on sampling from the real data, and from parametric distributions. I recommend using both types of DGMs, to assess the sensitivity of results to the choice of DGM. I also recommend comparing results from simulation studies with those from real data analysis, because performance of methods may differ. For example, in my simulation study, I found that use of parametric PH transition intensity models gave less biased estimates of RELOS than when using Cox models (Chapter 6, Section 6.2.17). However, when applying parametric models using the real NHS CBB dataset, I found that model convergence was frequently an issue, which made use of parametric models impractical (Chapter 7, Section 7.7.1).

8.3. Implications for researchers

When an analysis dataset includes variables with missing data, I recommend first performing a detailed exploration of the potential missingness mechanism for each incomplete variable in the analysis model. This can be achieved by comparing the distribution of all other variables for cases with missing and observed values of the incomplete variable, or by using logistic regression analysis (where the regression outcome is a binary missingness indicator for the incomplete variable). This is a similar approach to that described in the "TARMOS" framework (guidelines for the treatment and reporting of missing data in observational studies) (199). If missingness depends on the analysis outcome, then in most circumstances, CCA will be biased (79). For example, CCA will be biased in a competing risks or MSM analysis, if event times are missing only for certain event types.

I also recommend consulting clinicians, and other researchers who have expert knowledge of the data, to identify possible missingness mechanisms (200).

Knowledge of the predictors of the incomplete variable can be used to explore the plausibility of a MAR rather than MNAR missing data scenario. For example, if there is no difference in the distribution of a known predictor of event times (when comparing cases with missing and observed event times), a MNAR scenario is less likely. This is because, if event times were MNAR, a relationship between the missingness indicator and predictor would be induced. In summary, if some event times are missing, MI methods that assume MAR can be used to correct bias and improve precision compared with CCA, unless data MNAR is suspected.

When using MI, it is recommended that the number of imputations should equal at least the percentage of incomplete cases, to ensure adequate reproducibility and efficiency (87). If there is only one incomplete variable, only one iteration is required per imputation. In general, convergence is generally achieved within a small number of iterations (90). Convergence can be assessed by plotting the estimate against the iteration number within each imputation.

In general, I recommend following the guidance I have summarised in Table 8.1 above, by using FCS MI with type 1 PMM (PMM) to handle missing event times. When applying PMM methods, it is recommended to use at least five donors in the donor pool, and possibly more if the analysis dataset is very large (92). I have found that PMM can be used to handle missing event times in complex survival analysis models, such as competing risks and MSM analyses. In theory, MI assuming MAR is valid only when the imputation model is correctly specified (and all incomplete variables are not MNAR). However, I have found that PMM is robust to model misspecification. It was the least biased of all considered methods in multiple scenarios. PMM is also straight-forward to apply in many software packages. In addition, other incomplete variables (*e.g.* incomplete covariates) can be imputed within the same imputation model.

I found that the performance of PMM was improved by applying it separately for each sub-group of patients with a different distribution of event times. Therefore, when using PMM to impute event times, I recommend careful exploration of the data and distribution of event times for each sub-group of patients. In a competing risks analysis, I recommend comparing the distribution of event times for each event type. In MSM analysis, I recommend comparing the distribution of event times for each group of patients with a different path through the MSM. Applying PMM in separate imputation models for each sub-group of patients with a different distribution of event times may reduce bias and SE in model estimates. However, this must be balanced with the requirement for sub-groups to be of sufficient size to allow for random donor selection in the imputation procedure. In analysis of real data, there may only be small differences between distributions of event times for different sub-groups of patients. To assess the sensitivity of results to the imputation method, I recommend performing analysis using both a single imputation model and separate models for each sub-group of patients.

PMM is an appropriate method for handling missing event times when some event times are observed. It is particularly useful when the interval in which missing event times can occur is wide, relative to the distribution of observed times, because the FML methods developed to-date do not perform well in this case. It is applicable to any study with a mixture of missing and observed event times. For example, in a study of obesity and the risk of stillbirth (201), pregnancies that resulted in birth or stillbirth but with missing gestational age were excluded from analyses. Based on study criteria, missing gestational ages occurred between 20 and 42 weeks. In this scenario, PMM could be used to improve bias and precision of estimates.

The patient outcome data used in my study come from transplant centres in many different countries, because CB from the NHS CBB can be used for any suitably-matched patient worldwide. I have found that data completeness varies

between transplant centres and countries. For similar reasons, it is likely that analysis of other HSC transplant registries, or any other dataset using events data from multiple settings, will require handling of missing event times. Therefore, the MI strategies I describe are widely applicable.

PMM cannot be used when all event times are interval-censored. This may be due to the design of follow-up reporting. For example, after corneal transplantation, hospitals were asked whether any post-transplant surgery had been performed since the previous follow-up report but were not asked for the date of surgery (202). In this example, time of surgery was interval-censored for all patients. Similarly, in a study of risk factors for self-harm with and without suicidal intent (203), exposure data were collected annually or bi-annually from birth but participants were only asked at age 16/17 years whether they had harmed with suicidal intent at some point during their lifetime. Here the outcome of interest could have occurred at any point from birth to 16/17 years. Valid use of PMM would require further data collection to obtain exact event times for a suitable sample of patients. A likelihood-based method, such as those proposed by Bakoyannis *et al.* (64) for competing risks analysis, and Machado and van den Hout (152) for MSM analysis, may be a useful alternative in this scenario.

8.4. Implications for clinicians and patients

In HSC transplantation studies, and patient outcome studies in general, simple methods are often used to handle missing event times. These include CCA (44), the replacement of all missing times with the mean of the observed times (45), or the substitution of the censoring interval midpoint (204). Such methods can lead to bias and under-coverage (46, 47). For the NHS CBB patient outcomes analysis, CCA is biased because missingness of event times depends on the analysis outcome and covariates. In this scenario, my simulation study results suggest that CCA will tend to under-estimate the cumulative incidence and consequently

over-estimate the median event time (because event times for the competing events are generally fully observed and only cases of the events of interest are missing). In addition, my results suggest that CCA will result in biased estimates of the probability of experiencing acute GvHD and relapse following CB transplantation. Biased analysis results may lead to incorrect conclusions about the risks of CB transplantation, which will have considerable consequences for both clinicians and patients. Through my research, I have identified MI strategies for handling missing event times that can be used to correct bias and improve precision (compared with CCA and other naïve methods). This will ensure that accurate information is available to inform decision-making for both clinicians and patients.

My analysis of the NHS CBB dataset provides the first insight into patient outcomes following transplantation using CB donated in the UK. My analysis demonstrates that outcomes for NHS CBB patients are comparable to those for other CB banks worldwide, and for CB transplants in general. My analysis conclusions will reassure clinicians and patients that CB donated to the NHS CBB is a safe and effective treatment, and can be used to inform treatment choice, clinical guidelines and commissioning.

Currently, there is no information available for patients, from any of the main UK HSC transplantation organisations, about the frequency and likelihood of adverse events after CB transplantation (5, 11, 27). It is vital that patients have access to clear and easily understandable information about the risks and benefits of CB transplantation. An advantage of the MSM approach I have used is that it facilitates effective communication to patients, because the probability of adverse events at any given time can be illustrated graphically (see Chapter 7, Figure 7.3, for example), or summarised as the number of days spent in each state (Chapter 7, Table 7.3). Hence, the analysis described in this thesis can be used to fill the gap in available patient information.

8.5. Strengths and Limitations

A particular strength of my research has been the incorporation of robust methods for handling missing data alongside complex methods for modelling multiple events. I have compared MI strategies using in-depth simulation studies, providing valuable insight into the best approach for handling missing event times. In addition, I have extended previous research by describing in detail how to simulate MSM data. I have also derived new estimators of SE (of percentiles of event times and RELOS) and examined their performance using simulation studies.

Another strength of my research has been my use of detailed baseline and follow-up data, enabling me to provide the first insight into transplantation outcomes using CB donated to a UK CB bank. In the NHS CBB dataset, event times were commonly missing for transplants performed outside Europe (for transplants in the USA in particular). The standard approach in transplantation studies would be to exclude these transplants from analysis. Using MI methods, I have been able to include these transplants, and hence, provide more generalisable results (*i.e.* my results are not specific to UK/European transplantation protocols). My results are further generalisable because my analysis included patients with both malignant and non-malignant blood diseases, as well as both adult and paediatric patients. Prognoses and treatment approaches vary between these patient groups. Therefore, an additional strength of my research has been the application of MSMs, allowing prediction of transplant outcomes for specific patient types (*e.g.* paediatric patients treated for non-malignant blood disorders).

A limitation of my research has been the relatively small size of the NHS CBB dataset (N=432). This has limited the power of my analysis and may explain, for

example, why I did not detect a strong association between donor-recipient sex match and patient outcomes.

A further limitation of the work presented here is that some data were missing in my real dataset. I was not able to contact transplant centres directly, so there was no way to obtain the missing data. I performed extensive investigations to overcome this issue. However, the MI methods I applied are valid only if all incomplete variables are not MNAR, and the imputation model is correctly specified. In reality, there may be some unmeasured predictors of missingness. In addition, MNAR cannot truly be ruled out in any real dataset with many incomplete variables.

The MSM considered in my research represents an extremely simplified version of the event history for patients after HSC transplantation. In addition to the events reported in the NHS CBB dataset (see Chapter 4, Figure 4.1 for illustration of all event combinations experienced by patients in the NHS CBB dataset), patients can experience repeat recurrence of infections and other adverse events after transplant, as well as multiple transplants and periods of disease remission and relapse (10). Clinical inference would be strengthened, and important clinical questions could be answered, if a more detailed event history was modelled for each patient. However, this does rely on the availability of additional post-transplant data, which will almost certaintly include some missing data. A more complex analysis model will increase the complexity of any imputation model and the likelihood of imputation model misspecification.

8.6. Further work

As further work, the methods I have used in my research could be applied to other UK HSC transplantation registries, *e.g.* the British Bone Marrow registry (BBMR) (205). The larger size of the BBMR patient dataset (because more transplants are performed using BM than CB) would be an advantage. This

would provide further insight into the risks and benefit of HSC transplantation in a UK context. Furthermore, my methods could be applied to the UK organ transplantation registry (32). For example, MSM analysis could be used to determine the risks and benefits of sequential bilateral corneal transplantation (202), or to calculate the relative probability of death while waiting for a lung transplant compared with death following a lung transplant (31).

The three-state, unidirectional Markov model considered in my research is the simplest MSM that includes an intermediate state. Further work is needed to determine if the conclusions of this research still hold for more complex MSMs with multiple intermediate states, with reversible or recurring transitions between states, or with multiple time scales (68). There was some indication that the proportional hazards assumption did not hold for my transition intensity models, particularly for the model for the transition from transplant to relapse/death. Therefore, my analysis could be improved by including time-dependent regression parameters, or by using the dynamic landmarking approach (206).

Although I considered large datasets in some limited situations in my simulation studies, the main objective of my simulation studies was to inform my analysis of the real dataset. Therefore, the size of my simulated datasets, and choice of parameters in DGMs and MDMs, were guided by the real data. To assess the sensitivity of methods to these factors, a useful extension of my research would be to repeat my simulation studies using a range of sample sizes, covariate associations and event rates.

Although PMM performed well in my study, there is still scope for improvement, for example, by development of methods that are explicitly compatible with a competing risks or MSM analysis. This could be achieved, for example, through an extension of the MAR stacked MI approach of Beesley and Taylor (119) or the SMC-FCS method (194) to direct modelling of the cumulative

incidence function and to MSM. Alternatively, another method proposed by Beesley and Taylor (147) could be extended, combining an FML approach with full imputation (Beesley and Taylor use "improper" imputation within their EM algorithm). A particular benefit of their approach is that (unlike other FML methods developed so far), a mixture of exact and interval-censored event times can be accommodated. Finally, in my study, I found that parametric transition intensity models resulted in less biased estimates of expected length of stay in each state than semi-parametric models. Therefore, it would also be useful to extend compatible methods to parametric analysis models.

As a further work, sensitivity analyses could be performed to determine the extent to which analysis conclusions will change if the MAR assumption does not hold. This could be achieved by considering various MNAR missingness mechanisms. The simplest approach would be to set all missing event times to the smallest or largest of the observed times for each event type (or possibly use the average follow-up time instead of the largest observed time for missing times of relapse or chronic GvHD). A more sophisticated alternative would be the "pattern mixture" approach (85), in which imputation model parameters, for each incomplete covariate or outcome, are changed by a value δ, where δ represents the hypothesised difference between the distribution of observed and missing event times. This can be implemented using an FCS MI approach (207). Other possible approaches are to use expert knowledge to inform likely missingness mechanisms (200), to use proxy data in the imputation model (196), or to use a weighting approach post-imputation to correct any bias *e.g.* as in the weighted analysis of stacked MI approach (208).

8.7. Final conclusions

In summary, this thesis provides the first insight into patient outcomes after HSC transplantation using CB donated to a UK CB bank. This information will benefit NHSBT, health statisticians, patients, and clinicians. In my thesis, I have

demonstrated the gains, in terms of bias and precision, in using MI strategies to handle missing event times in complex survival models. The methods described in my thesis will be useful more generally for statisticians and researchers in a wide range of contexts.

References

1. Antony Nolan. Blood cancers and blood disorders [cited 2017 25 August]. Available from: https://www.anthonynolan.org/patients-and-families/blood-cancers-and-blood-disorders.
2. Blood Cancer UK. Facts and information about blood cancer [cited 2021 11 September]. Available from: https://bloodcancer.org.uk/news/blood-cancer-facts/.
3. Wodnar-Filipowicz A. Biological properties of haematopoietic stem cells. In: Apperley J, Carreras E, Gluckman E, Masszi T, editors. Haematopoietic Stem Cell Transplantation: The EBMT Handbook. 6th ed: European School Hematology; 2012.
4. Gluckman E. Choice of the donor according to HLA typing and stem cell source. In: Apperley J, Carreras E, Gluckman E, Masszi T, editors. Haematopoietic Stem Cell Transplantation: The EBMT Handbook. 6th ed: European School Hematology; 2012.
5. British Soc of Blood and Marrow Transplantation. UK & ROI Transplant Activity 2019 [cited 2021 8 September]. Available from: https://bsbmtct.org/activity/2019/.
6. BMDW [cited 2021 14 June]. Available from: www.bmdw.org.
7. Human Tissue Authority. Cord blood banking: A guide for parents February 2016 [cited 2021 7 July]. Available from: https://archive.hta.gov.uk/sites/default/files/Cord%20Blood%20Guide%202019_2.pdf.
8. Danby R, Rocha V. Improving engraftment and immune reconstitution in umbilical cord blood transplantation. Front Immunol. 2014;5(68):1-19.
9. Hwang WYK, Samuel M, Tan D, Koh LP, Lim W, Linn YC. A Meta-Analysis of Unrelated Donor Umbilical Cord Blood Transplantation versus Unrelated Donor Bone Marrow Transplantation in Adult and Pediatric Patients. Biol Blood Marrow Transplant. 2007;13:444-53.
10. Mackall C, Fry T, Gress R, Peggs K, Storek J, Toubert A. Background to hematopoietic cell transplantation, including post transplant immune recovery. Bone Marrow Transplant. 2009;44:457-62.
11. Anthony Nolan. Aligned Registry Demographics 2019-2020 [cited 2021 9 September]. Available from: https://www.anthonynolan.org/sites/default/files/2021-01/2266ST_StateOfTheRegistry_1920_v6.pdf.
12. Horowitz MM, Confer DL. Evaluation of Hematopoietic Stem Cell Donors. Hematology Am Soc Hematol Educ Program 2005;1:469–75.
13. Lou X, Zhao C, Chen H. Unrelated donor umbilical cord blood transplant versus unrelated hematopoietic stem cell transplant in patients with acute leukemia: A metaanalysis and systematic review. Blood Rev. 2018;32:192-202.
14. Barker JN, Kurtzberg J, Ballen K, Boo M, Brunstein C, Cutler C, et al. Optimal Practices in Unrelated Donor Cord Blood Transplantation for Hematologic Malignancies. Biol Blood Marrow Transplant. 2017;23:882-96.

15. Ustun C, Giannotti F, Zhang M-J, Wang H-L, Brunstein C, Labopin M, et al. Outcomes of UCB transplantation are comparable in FLT3+ AML: Results of CIBMTR, Eurocord and EBMT collaborative analysis. Leukemia. 2017;31(6):1408-14.
16. Carreras E. Early complications after HSCT. In: Apperley J, Carreras E, Gluckman E, Masszi T, editors. Haematopoietic Stem Cell Transplantation: The EBMT Handbook. 6th ed. Paris, France: European School Hematology; 2012.
17. Apperley J, Masszi T. Graft-versus-host disease. In: Apperley J, Carreras E, Gluckman E, Masszi T, editors. Haematopoietic Stem Cell Transplantation: The EBMT Handbook. 6th ed. Paris, France: European School Hematology; 2012.
18. Cancer Research UK. Why Some Cancers Come Back [cited 2021 10 September]. Available from: https://www.cancerresearchuk.org/about-cancer/what-is-cancer/why-some-cancers-come-back.
19. Gisselbrecht C, Van Den Neste E. How I manage patients with relapsed/refractory diffuse large B cell lymphoma. Br J Haematol. 2018;182:633-43.
20. Styczyński J, Tridello G, Koster L, Iacobelli S, Biezen Av, Werf Svd, et al. Death after hematopoietic stem cell transplantation: changes over calendar year time, infections and associated factors. Bone Marrow Transplant. 2020;55:126-36.
21. Eapen M, Rubinstein P, Zhang M-J, Stevens C, Kurtzberg J, Scaradavou A, et al. Outcomes of transplantation of unrelated donor umbilical cord blood and bone marrow in children with acute leukaemia: a comparison study. The Lancet. 2007;369:1947-54.
22. Isobe M, Konuma T, Kato S, Tanoue S, Mizusawa M, Maki Oiwa-Monna, et al. Development of Pre-Engraftment Syndrome, but Not Acute Graft-versus-Host Disease, Reduces Relapse Rate of Acute Myelogenous Leukemia after Single Cord Blood Transplantation. Biol Blood Marrow Transplant. 2019;25:1187-96.
23. Baron F, Labopin M, Ruggeri A, Volt F, M. Mohty, Blaise D, et al. Cord blood transplantation is associated with good outcomes in secondary Acute Myeloid Leukaemia in first remission. J Intern Med. 2019;285:446-54.
24. Wang L, Gu Z-y, Liu S-f, Ma D-x, Zhang C, Liu C-j, et al. Single- Versus Double-Unit Umbilical Cord Blood Transplantation for Hematologic Diseases: A Systematic Review. Transfus Med Rev. 2019;33:51-60.
25. Gabelli M, Veys P, Chiesa R. Current status of umbilical cord blood transplantation in children. Br J Haematol. 2020;190:650-83.
26. Eapen M, Klein oP, Ruggeri A, Spellman S, Lee SJ, Anasetti C, et al. Impact of allele-level HLA matching on outcomes after myeloablative single unit umbilical cord blood transplantation for hematologic malignancy. Blood. 2014;123(1):133-40.
27. NHS Blood and Transplant. NHS Cord Blood Bank. [Cited 2018 Jul 16]. Available from: http://www.nhsbt.nhs.uk/cordblood/.
28. National Health Service. The NHS in England. [Cited 2018 Jul 16]. Available from: http://www.nhs.uk/NHSEngland/thenhs/about/Pages/authoritiesandtrusts.aspx.

29. Shaw B, Veys P, Pagliuca A, Addada J, Cook G, Craddock C, et al. Recommendations for a standard UK approach to incorporating umbilical cord blood into clinical transplantation practice: conditioning protocols and donor selection algorithms. Bone Marrow Transplant. 2009;44:7-12.
30. NHS England Specialised Services Clinical Reference Group for Blood and Marrow Transplantation. Clinical Commissioning Policy: Allogeneic Haematopoietic Stem Cell Transplantation (HSCT) (All Ages): Revised. Electronic format only; 2015.
31. Winton Centre for Risk and Evidence Communication. Communicating the risks & benefits around transplant surgery 2021 [cited 2021 14 September]. Available from: https://wintoncentre.maths.cam.ac.uk/projects/communicating-risks-and-benefits-around-transplant-surgery/.
32. NHS Blood and Transplant. Statistics and Reports [cited 2021 10 September]. Available from: https://www.odt.nhs.uk/statistics-and-reports/.
33. Donaldson C, Buchanan R, Webster J, Laundy V, Horsley H, Barron C, et al. Development of a district Cord Blood Bank: a model for cord blood banking in the National Health Service. Bone Marrow Transplant. 2000;25:899-905.
34. Davey S, Armitage S, Rocha V, Garnier F, Brown J, Brown CJ, et al. The London Cord Blood Bank: analysis of banking and transplantation outcome. Br J Haematol. 2004;125:358-65.
35. Guerra SG, Hamilton-Jones S, Brown CJ, Navarrete CV, Chong W. Next generation sequencing of 11 HLA loci characterises a diverse UK cord blood bank. Hum Immunol. 2020;81:269-79.
36. Agence de la Biomedecine. The Eurocord Registry [cited 2017 29 August]. Available from: https://www.agence-biomedecine.fr/Registre-Eurocord.
37. Nagamura-Inoue T, Kai S, Azuma H, MTakanashi, Isoyama K, Kato K, et al. Unrelated cord blood transplantation in CML: Japan Cord Blood Bank Network analysis. Bone Marrow Transplant. 2008;42:241-51.
38. Guerra-Marquez A, Penaflor K, Mayani H. Cord Blood Banking and Transplantation in a National Program: Thirteen Years of Experience. Arch Med Res. 2020;51:54-62.
39. Ong S-Y, Phipps C, Chu P, Prasath A, Ho A, Sobak S, et al. Cord blood unit factors influencing transplant outcomes from the Asian multiethnic Singapore Cord Blood Bank. Bone Marrow Transplant. 2015;50:1256-8.
40. Serafini M, Poli F, Torelli R, Fleischhauer K, Zino E, Arias AEd, et al. The impact of immunogenetics and clinical factors on the outcome of unrelated cord blood transplantation: The Milano Cord Blood Bank data. Transfus Apher Sci. 2010;42:283-8.
41. Gluckman E, Rocha V, Garnier F, Ionescu I, Chevret S. Report of the Eurocord Registry. Experimental Hematology 28 2000. p. 78.
42. Lesaffre E, Komárek A. An overview of methods for interval-censored data with an emphasis on applications in dentistry. Stat Methods Med Res. 2005;14:539–52.
43. De Gruttola V, Lagakos SW. Analysis of Doubly-Censored Survival Data, with Application to AIDS. Biometrics. 1989;45:1-11.

44. Hanna J. Khoury, Tao Wang, Michael T. Hemmer, Daniel Couriel, Amin Alousi, Corey Cutler, et al. Improved survival after acute graft-versus-host disease diagnosis in the modern era. Haematologica. 2017;102(5):958-66.
45. Van Rood JJ, Stevens CE, Smits J, Carrier C, Carpenter C, Scaradavou A. Reexposure of cord blood to noninherited maternal HLA antigens improves transplant outcome in hematological malignancies. Proc Natl Acad Sci U S A. 2009;106(47):19952-7.
46. Little RJA. Regression With Missing X's: A Review. Journal of the American Statistical Association. 1992;87(420):1227-37.
47. Sun J. The Statistical Analysis of Interval-censored Failure Time Data. New York, NY: Springer; 2007.
48. Nikolajeva O, Rocha V, Danby R, Ruggeri A, Volt F, Baudoux E, et al. Umbilical Cord Blood Cytomegalovirus Serostatus Does Not Have an Impact on Outcomes of Umbilical Cord Blood Transplantation for Acute Leukemia. Biol Blood Marrow Transplant. 2017;23:1729–35.
49. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: Competing risks and multi-state models. Stat Med. 2007;26:2389–430.
50. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association. 1958;53(282):457-81.
51. Cox DR. Regression Models and Life-Tables. Journal of the Royal Statistical Society, Series B. 1972;34(2):187-220.
52. Aalen O. A model for nonparametric regression analysis of counting processes. Klonecki W, Kozek A, Rosinski J, editors. New York: Springer; 1980.
53. Collett D. Modelling Survival Data in Medical Research 3rd ed. Florida, USA: Chapman and Hall /CRC; 2015.
54. Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemotherapy Reports. 1966;50(3):163-70.
55. Royston P, Parmar MK. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Stat Med. 2002;21:2175-97.
56. Rutherford MJ, Crowther MJ, Lambert PC. The use of restricted cubic splines to approximate complex hazard functions in the analysis of time-to-event data: a simulation study. Journal of Statistical Computation and Simulation. 2015;85(4):777-93.
57. Breslow N, Crowley J. A Large Sample Study of the Life Table and Product-Limit Estimates under Random Censorship. The Annals of Statistics. 1974;2:437-53.
58. Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. 2nd edition. New York: John Wiley and Sons; 2002.
59. Omer AK, Weisdorf DJ, Lazaryan A, Shanley R, Blazar BR, MacMillan ML, et al. Late Acute Graft Versus Host Disease After Allogeneic Hematopoietic Stem Cell Transplantation. Biol Blood Marrow Transplant. 2016;22(5):879-83.
60. Aalen O, Johansen S. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. Scandinavian Journal of Statistics 1978;5:141-50.

61. Gray RJ. A class of K-sample tests for comparing the cumulative incidence of a competing risk. Annals of Statistics. 1988;16:1141-54.
62. Fine JP, Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. Journal of the American Statistical Association. 1999;94(446):496-509.
63. Putter H, Schumacher M, Houwelingen HCv. On the relation between the cause-specific hazard and the subdistribution rate for competing risks data: The Fine–Gray model revisited. Biom J. 2020;62:790-807.
64. Bakoyannis G, Yu M, Yiannoutsos CT. Semiparametric regression on cumulative incidence function with interval-censored competing risks data. Stat Med. 2017;36:3683-707.
65. Austin PC, Steyerberg EW, Putter H. Fine-Gray subdistribution hazard models to simultaneously estimate the absolute risk of different event types: Cumulative total failure probability may exceed 1. Stat Med. 2021;40:4200-12.
66. Klein JP, Keiding N, Copelan EA. Plotting summary predictions in multistate survival models: Probabilities of relapse and death in remission for bone marrow transplantation patients. Stat Med. 1994;13:2315-32.
67. Andersen PK, Pohar Perme M. Inference for outcome probabilities in multi-state models. Lifetime Data Anal. 2008;14(4):405-31.
68. Iacobelli S, Carstensen B. Multiple time scales in multi-state models. Stat Med. 2013;32(30):5315-27.
69. Titman AC, Putter H. General tests of the Markov property in multi-state models. Biostatistics. 2020;00(00):1-17.
70. Feller W. On Boundaries and Lateral Conditions for the Kolmogorov Differential Equations. Annals of Mathematics. 1957;65(3):527-70.
71. Datta S, Satten GA. Validity of the Aalen-Johansen estimators of stage occupation probabilities and Nelson–Aalen estimators of integrated transition hazards for non-Markov models. Statistics & Probability Letters. 2001;55(4):403–11.
72. Grand MK, Putter H. Regression models for expected length of stay. Stat Med. 2016;35:1178-92.
73. Andersen PK, Keiding N. Multi-state models for event history analysis. Stat Methods Med Res. 2002;11:91-115.
74. Putter H, Spitoni C. Non-parametric estimation of transition probabilities in non-Markov multi-state models: The landmark Aalen–Johansen estimator. Stat Methods Med Res. 2018;27(7):2081-92.
75. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med. 2019;38:2074-102.
76. Wei LJ, Lin DY, Weissfeld L. Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions. Journal of the American Statistical Association. 1989;84(408):1065-73.
77. Beyersmann J, Latouche Ae, Buchholz A, Schumacher M. Simulating competing risks data in survival analysis. Statist Med. 2009;28:956-71.
78. Schafer JL, Graham JW. Missing Data: Our View of the State of the Art. Psychol Methods. 2002;7(2):147-77.

79. Little RJA, Zhang N. Subsample ignorable likelihood for regression analysis with missing data. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2011;60(4):591-605.
80. Bartlett JW, Harel O, Carpenter JR. Asymptotically Unbiased Estimation of Exposure Odds Ratios in Complete Records Logistic Regression. Am J Epidemiol. 2015;182(8):730-6.
81. Galati JC, Seaton KA. MCAR is not necessary for the complete cases to constitute a simple random subsample of the target sample. Stat Methods Med Res. 2016;25(4):1527-34.
82. Little RJA, Rubin DB. Statistical Analysis with Missing Data. Hoboken, New Jersey: Wiley; 2002.
83. Perkins NJ, Cole SR, Harel O, Tchetgen EJT, Sun B, Mitchell EM, et al. Principled Approaches to Missing Data in Epidemiologic Studies. Am J Epidemiol. 2018;187(3):568-75.
84. Carpenter JR, Kenward MG, Vansteelandt S. A comparison of multiple imputation and doubly robust estimation for analyses with missing data. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2006;169(3):571-84.
85. Carpenter JR, Smuk M. Missing data: A statistical framework for practice. Biom J. 2021;63(5):915-47.
86. Hughes R, Heron J, Sterne J, Tilling K. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. Int J Epidemiol. 2019:1-11.
87. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine 2011;30:377–99.
88. Rubin DB. Multiple imputation for nonresponse in surveys. New York: Wiley; 1987.
89. Madley-Dowd P, Hughes R, Tilling K, Heron J. The proportion of missing data should not be used to guide decisions on multiple imputation. J Clin Epidemiol. 2019;110:63-73.
90. Van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res. 2007;16:219–42.
91. Van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med. 1999;18: 681—94.
92. Morris TP, White IR, Royston P. Tuning multiple imputation by predictive mean matching and local residual draws. BMC Medical Research Methodology 2014;14(75).
93. Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software. 2011;45(3):1-67.
94. Hughes RA, White IR, Seaman SR, Carpenter JR, Tilling K, Sterne JAC. Joint modelling rationale for chained equations. BMC Med Res Methodol. 2014;14:28+.
95. Seaman SR, Hughes RA. Relative efficiency of joint-model and full-conditional-specification multiple imputation when conditional models are compatible: The general location model. Stat Methods Med Res. 2016;Epub.
96. Little RJA. Missing Data Adjustments in Large Surveys. Journal of Business Economics and Statistics. 1988;6:287–301.

97. Rodwell L, Lee KJ, Romaniuk H, Carlin JB. Comparison of methods for imputing limited-range variables: a simulation study. BMC Med Res Methodol. 2014;14:57.
98. Lee KJ, Carlin JB. Multiple imputation in the presence of non-normal data. Stat Med. 2017;36:606-17.
99. Belin TR, Hu M-Y, Young AS, Grusky O. Performance of a general location model with an ignorable missing-data assumption in a multivariate mental health services study. Stat Med. 1999;18:3123-35.
100. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Stat Med. 2010;29:2920-31.
101. Bartlett JW, Seaman SR, White IR, Carpenter JR. Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. Stat Methods Med Res. 2015;24(4):462-87.
102. Von Hippel PT. How to impute interactions, squares, and other transformed variables. Sociol Methodol. 2009;39:265–91.
103. Meng X-L. Multiple-Imputation Inferences with Uncongenial Sources of Input. Statistical Science. 1994;9(4):538-73.
104. Morris TP, White IR, Carpenter JR, Stanworth SJ, Royston P. Combining fractional polynomial model building with multiple imputation. Stat Med. 2015;34(25):3298-317.
105. Hughes RA, Sterne JAC, Tilling K. Comparison of imputation variance estimators. Stat Methods Med Res. 2016;25(6):2541-57.
106. Robins JM, Wang N. Inference for imputation estimators. Biometrika. 2000;87(1):113–24.
107. White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometricia. 19980;48:817-30.
108. Bartlett JW, Hughes RA. Bootstrap inference for multiple imputation under uncongeniality and misspecification. Stat Methods Med Res. 2020;29(12):3533-46.
109. Vansteelandt S, Carpenter JR, Kenward MG. Analysis of Incomplete Data Using Inverse Probability Weighting and Doubly Robust Estimators. Methodology. 2010;6(1):37-48.
110. Liu J, Gelman A, Hill J, Su Y-S, Kropko J. On the Stationary Distribution of Iterative Imputations. Biometrika. 2013;Epub.
111. Royston P. Multiple Imputation of Missing Values: Update. The Stata Journal. 2005;5(2):188-201.
112. Morris TP, White IR, Royston P, Seaman SR, Wood WA. Multiple imputation for an incomplete covariate that is a ratio. Stat Med. 2014;33(88-104).
113. Seaman SR, Bartlett JW, White IR. Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. BMC Med Res Methodol. 2012;12:46.
114. Tilling K, Williamson EJ, Spratt M, Sterne JAC, Carpenter JR. Appropriate inclusion of interactions was needed to avoid bias in multiple imputation. J Clin Epidemiol. 2016;80:107-15.
115. Van Buuren S. Flexible imputation of missing data. Boca Raton, FL: Chapman & Hall/CRC; 2012.

116. Vink G, Van Buuren S. Multiple imputation of squared terms. Sociological Methods and Research. 2013;42(4):598-607.
117. Goldstein H, Carpenter JR, Browne WJ. Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2014;177(2):553-64.
118. Diaz-Ordaz K, Kenward MG, Grieve R. Handling missing values in cost effectiveness analyses that use data from cluster randomized trials. Journal of the Royal Statistical Society Series A. 2014;177(2):457-74.
119. Beesley LJ, Taylor JMG. A stacked approach for chained equations multiple imputation incorporating the substantive model. Biometrics. 2020:1-13.
120. Royston P. Multiple imputation of missing values: further update of ice, with an emphasis on interval censoring. The Stata Journal. 2007;7(4):445–64.
121. Von Hippel PT. Should a Normal Imputation Model be Modified to Impute Skewed Variables? Sociological Methods & Research. 2012;42(1):105-38.
122. Kwon TY, Park Y. A new multiple imputation method for bounded missing values. Statistics and Probability Letters. 2015;107:204–9.
123. White IR, Royston P. Imputing missing covariate values for the Cox model. Stat Med. 2009;28:1982–98.
124. Von Hippel PT. Regression with missing Ys: an improved strategy for analyzing multiply imputed data. Sociol Methodol. 2007;37:83-117.
125. Sullivan TR, Salter AB, Ryan P, Lee KJ. Bias and Precision of the "Multiple Imputation, Then Deletion" Method for Dealing With Missing Outcome Data. Am J Epidemiol. 2015;182(6):528–34.
126. Keogh RH, Morris TP. Multiple imputation in Cox regression when there are time-varying effects of covariates. Stat Med. 2018;37:3661-78.
127. Bartlett JW, Taylor JMG. Missing covariates in competing risks analysis. Biostatistics. 2016;17(4):751-63.
128. Jackson D, White IR, Seaman S, Evans H, Baisley K, Carpenter J. Relaxing the independent censoring assumption in the Cox proportional hazards model using multiple imputation. Stat Med. 2014;33(27):4681-94.
129. Hsu C-H, Taylor JMG, Murray S, Commenges D. Survival analysis using auxiliary variables via non-parametric multiple imputation. Stat Med. 2006;25:3503-17.
130. Robins JM, Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. Biometrics. 2000;56:779-88.
131. Xiang F, Murray S, Liu X. Analysis of transplant urgency and benefit via multiple imputation. Stat Med. 2014;33:4655-70.
132. Hsu C-H, Taylor JMG, Murray S, Commenges D. Multiple imputation for interval censored data with auxiliary variables. Stat Med. 2007;26:769-81.
133. Pan W. A Multiple Imputation Approach to Cox Regression with Interval-Censored Data. Biometrics. 2000;56:199-203.
134. Wei GCG, Tanner MA. A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. Journal of the American Statistical Association. 1990;85(411):699-704.

135. Chen L, Sun J. A multiple imputation approach to the analysis of interval-censored failure time data with the additive hazards model. Computational Statistics and Data Analysis. 2010;54 1109-16.
136. Delord M, Génin E. Multiple imputation for competing risks regression with interval-censored data. Journal of Statistical Computation and Simulation. 2015;86(11):2217-28.
137. Kim J, Kim J, Kim SW. Additive-multiplicative hazards regression models for interval-censored semi-competing risks data with missing intermediate events. BMC Med Res Methodol. 2019;19(49).
138. Hyun N, Katki HA, Graubard BI. Sample-weighted semiparametric estimation of cause-specific cumulative risk and incidence using left- or interval-censored data from electronic health records. Stat Med. 2020;39:2387-402.
139. Gomez G, Calle ML, Oller R, Langohr K. Tutorial on methods for interval-censored data and their implementation in R. Statistical Modelling. 2009;9(4):259–97.
140. Hudgens MG, Satten GA, Longini Jr IM. Nonparametric Maximum Likelihood Estimation for Competing Risks Survival Data Subject to Interval Censoring and Truncation. Biometrics 2001;57:74-80.
141. Hudgens MG, Li C, Fine JP. Parametric likelihood inference for interval censored competing risks data. Biometrics. 2014;70(1):1-9.
142. Mitra D, Das U, Das K. Analysis of interval-censored competing risks data under missing causes. Journal of Applied Statistics. 2020;47(3):439-59.
143. Li C. Cause-Specific Hazard Regression for Competing Risks Data Under Interval Censoring and Left Truncation. Comput Stat Data Anal. 2016;104:197-208.
144. Li C. The Fine–Gray Model Under Interval Censored Competing Risks Data. J Multivar Anal. 2016;143:327-44.
145. Frydman H, Szarek M. Nonparametric Estimation in a Markov "Illness-Death" Process from Interval Censored Observations with Missing Intermediate Transition Status. Biometrics. 2009;65:143-51.
146. Foucher Y, Giral M, Soulillou JP, Daures JP. A semi-Markov model for multistate and interval-censored data with multiple terminal events. Application in renal transplantation. Stat Med.26(30):5381-93.
147. Beesley LJ, Taylor JMG. EM algorithms for fitting mutlistate cure models. Biostatistics. 2019;20(3):416-32.
148. Jackson CH. Multi-State Models for Panel Data: The msm Package for R. Journal of Statistical Software. 2011;38(8):1-28.
149. Healy B, Gruttola VD. Hidden Markov models for settings with interval-censored transition times and uncertain time origin: application to HIV genetic analyses. Biostatistics. 2007;8(2):438–52.
150. Joly P, Commenges D, Helmer C, Letenneur L. A penalized likelihood approach for an illness–death model with interval-censored data: application to age-specific incidence of dementia. Biostatistics. 2002;3(3):433-43.
151. Machado RJM, Van den Hout A. Flexible multistate models for interval-censored data: Specification, estimation, and an application to ageing research. Stat Med. 2018;37:1636-49.

152. Machado RJM, Hout Avd, Marra G. Penalised maximum likelihood estimation in multi-state models for interval-censored data. Computational Statistics and Data Analysis. 2021;153:107057.
153. Hladun R, Badell I, González M, Martínez A, Sánchez de Toledo J, Olivé M. Análisis de la supervivencia de los ninos con inmunodeficiencias primarias que han recibido un trasplante de progenitores hematopoyéticos en Espana. Anales de Pediatría. 2015;82:62-7.
154. Knol MJ, Janssen KJM, Donders ART, Egberts ACG, Heerdink ER, Grobbee DE, et al. Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. J Clin Epidemiol. 2010;63:728-36.
155. Hernán MA, Sauer BC, Hernández-Díaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. J Clin Epidemiol. 2016;79:70-5.
156. Glimelius I, Smedby KE, Albertsson-Lindblad A, Crowther MJ, Eloranta S, Jerkeman M, et al. Unmarried or less-educated patients with mantle cell lymphoma are less likely to undergo a transplant, leading to lower survival. Blood Advances. 2021;5(6):1638-47.
157. Collins LM, Schafer JL, Kam C-M. A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures. Psychol Methods. 2001;6(4):330-51.
158. Fisher R. On the interpretation of $\chi 2$ from contingency tables, and the calculation of P. Journal of the Royal Statistical Society. 1922;85(1):87-94.
159. Collett D. Modelling Binary Data. 2nd ed. Florida, USA: Chapman and Hall/CRC; 2002.
160. Lee J-H, Je-Hwan Lee, Kim D-Y, Yun S-C, Kim S-D, Choi Y, et al. Pre-engraftment graft-versus-host disease after allogeneic hematopoietic cell transplantation for acute leukemia. Eur J Haematol. 2011;86:172-81.
161. Shimoni A, Labopin M, Savani B, Volin L, Ehninger G, Kuball J, et al. Long-term survival and late events after allogeneic stem cell transplantation from HLA-matched siblings for acute myeloid leukemia with myeloablative compared to reduced-intensity conditioning: a report on behalf of the acute leukemia working party of European group for blood and marrow transplantation. Journal of Haematology & Oncology. 2016;9(118).
162. Kumar S, Mahmood S, Lacy M, Dispenzieri A, Hayman S, Buadi F, et al. Impact of early relapse after auto-SCT for multiple myeloma. Bone Marrow Transplant. 2008;42:413-20.
163. Lymphoma Association. Allogeneic stem cell transplants 2011 [cited 2016 1 September]. Available from: http://www.nhs.uk/ipgmedia/national/Lymphoma%20Association/Assets/AllogeneicstemcelltransplantsLA84pages.pdf.
164. Pintilie M. Competing Risks: A Practical Perspective. New York: Wiley; 2006.
165. Grand MK, de Witte TJM, Putter H. Dynamic prediction of cumulative incidence functions by direct binomial regression. Biometric journal. 2018;60:734-47.

166. Morgan BJT. Elements of Simulation. 1st ed. London, UK: Chapman & Hall; 1984.
167. Aalen O. Nonparametric estimation of partial transition probabilities in multiple decrement models. Annals of Statistics. 1978;6:534-45.
168. Johansen S. The Product Limit Estimator as Maximum Likelihood Estimator. Scand J Statist. 1978;5:195-9.
169. Marubini E, Valsecchi MG. Analysing Survival Data from Clinical Trials and Observational Studies. New York: Wiley; 1995.
170. Braun TM, Yuan Z. Comparing the small sample performance of several variance estimators under competing risks. Stat Med. 2007;26:1170-80.
171. Allignol A, Schumacher M, Beyersmann J. A note on variance estimation of the Aalen-Johansen estimator of the cumulative incidence function in competing risks, with a view towards left-truncated data. Biometrical journal Biometrische Zeitschrift. 2010;52(1):126-37.
172. de Wreede LC, Fiocco M, Putter H. mstate: An R Package for the Analysis of Competing Risks and Multi-State Models. Journal of Statistical Software. 2011;38(7):1-30.
173. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Stat Med. 2006;25:4279-92.
174. Park J, Bakoyannis G, Yiannoutsos C. Semiparametric competing risks regression under interval censoring using the R package intccr. Comput Methods Programs Biomed. 2019;173:167-76.
175. Delord M. MIICD: Multiple Imputation for Interval Censored Data. R package version 2.4. 2017 [cited 2018 15 June]. Available from: https://CRAN.R-project.org/package=MIICD.
176. Hojo T, Pearson K. Distribution of the Median, Quartiles and Interquartile Distance in Samples From a Normal Population. Biometrika. 1931;23(3/4):315-63.
177. Brookmeyer R, Crowley J. A Confidence Interval for the Median Survival Time. Biometrics. 1982;38(1):29-41.
178. White IR. simsum: Analyses of simulation studies including Monte Carlo error. The Stata Journal. 2010;10(3):369-85.
179. Maathuis M. MLEcens: Computation of the MLE for bivariate (interval) censored data. R package version 0.1-4. [Cited 2019 Jul 24]. Available from: https://CRAN.R-project.org/package=MLEcens.
180. Keiding N, Klein JP, Horowitz MM. Multi-state models and outcome prediction in bone marrow transplantation. Statist Med. 2001;20:1871-85.
181. van den Broek BTA, Page K, Paviglianiti A, Hol J, Allewelt H, Volt F, et al. Early and late outcomes after cord blood transplantation for pediatric patients with inherited leukodystrophies. Blood Advances. 2018;2(1):49-60.
182. Wagner JE, Eapen M, Carter S, Wang Y, Schultz KR, Wall DA, et al. One-Unit versus Two-Unit Cord-Blood Transplantation for Hematologic Cancers. N Engl J Med. 2014;371(18):1685-94.
183. Therneau T. A Package for Survival Analysis in R 2020 [cited 2020 19 October]. Available from: https://CRAN.R-project.org/package=survival.
184. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. Stat Med. 2005;24:1713-23.

185. Delignette-Muller ML, Dutang C. fitdistrplus: An R Package for Fitting Distributions. Journal of Statistical Software. 2015;64(4):1-34.
186. Crowther MJ, Lambert PC. Simulating biologically plausible complex survival data. Stat Med. 2013;32:4118-34.
187. Brilleman SL. simsurv: Simulate Survival Data 2019 [cited 2020 10 February]. Available from: https://CRAN.R-project.org/package=simsurv.
188. Jackson C. flexsurv: A Platform for Parametric Survival Modeling in R. Journal of Statistical Software. 2016;70(8):1-33.
189. Crowther MJ, Lambert PC. Parametric multistate survival models: Flexible modelling allowing transition-specific distributions with application to estimating clinically useful measures of effect differences. Stat Med. 2017;36:4719-42.
190. Nemes S, Bülow E, Gustavsson A. A Brief Overview of Restricted Mean Survival Time Estimators and Associated Variances. Stats. 2020;3:107-19.
191. Royston P, Parmar MK. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. BMCMedical Research Methodology. 2013;13(152):1-15.
192. Efron B. Bootstrap methods: another look at the jackknife. The Annals of Statistics. 1979;7(1):1-26.
193. Wreede LCd, Fiocco M, Putter H. The mstate package for estimation and prediction in non- and semi-parametric multi-state and competing risks models. Computer Methods and Programs in Biomedicine 2010;99:261–74.
194. Bartlett JW. smcfcs: Multiple Imputation of Covariates by Substantive Model Compatible Fully Conditional Specification. R package version 1.2.1. 2016 [cited 2016 19 September]. Available from: https://CRAN.R-project.org/package=smcfcs.
195. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika. 1994;81(3):515-26.
196. Cornish R, Macleod J, Carpenter J, Tilling K. Multiple imputation using linked proxy outcome data resulted in important bias reduction and efficiency gains: A simulation study. Emerging Themes in Epidemiology. 2017;14(14):1-13.
197. Hayashi H, Ruggeri A, Volt F, Cornelissen JJ, Socié G, Sengeloev H, et al. Chronic graft-versus-host disease features in double unit cord blood transplantation according to National Institutes of Health 2005 cGVHD Consensus criteria. Bone Marrow Transplant. 2018;53:417–21
198. Lazaryan A, Weisdorf DJ, DeFor T, Brunstein CG, MacMillan ML, Bejanyan N, et al. Risk Factors for Acute and Chronic Graft-versus-Host Disease after Allogeneic Hematopoietic Cell Transplantation with Umbilical Cord Blood and Matched Related Donors. Biol Blood Marrow Transplant. 2016;22:134-40.
199. Lee KJ, Tilling K, Cornish RP, Little RJ, Bell ML, Goetghebeur E, et al. Framework for the Treatment And Reporting of Missing data in Observational Studies: The TARMOS framework. J Clin Epidemiol. 2021;134:79-88.
200. Mason AJ, Gomes M, Grieve R, Ulug P, Powell JT, Carpenter J. Development of a practical approach to expert elicitation for randomised controlled trials with missing health outcomes: Application to the IMPROVE trial. Clinical Trials. 2017;14(4):357-67.

201. Yao R, Ananth CV, Park BY, Pereira L, Plante LA. Obesity and the risk of stillbirth: a population-based cohort study. Am J Obstet Gynecol. 2014;210(5):457.e1-.e9.
202. Steger B, Curnow E, Cheeseman R, Romano V, Kaye A, Jones M, et al. Sequential Bilateral Corneal Transplantation and Graft Survival on behalf of the National Health Service Blood and Transplant Ocular Tissue Advisory Group and contributing ophthalmologists Am J Ophthalmol. 2016;170:50-7.
203. Mars B, Heron J, Crane C, Hawton K, Kidger J, Lewis G, et al. Differences in risk factors for self-harm with and without suicidal intent: Findings from the ALSPAC cohort. J Affect Disord. 2014;168:407-14.
204. Manzini G, Ettrich TJ, Kremer M, Kornmann M, Henne-Bruns D, Eikema DA, et al. Advantages of a multi-state approach in surgical research: how intermediate events and risk factor profile affect the prognosis of a patient with locally advanced rectal cancer. BMC Med Res Methodol. 2018;18(23).
205. NHS Blood and Transplant. British Bone Marrow Registry [cited 2021 23 September]. Available from: https://www.nhsbt.nhs.uk/british-bone-marrow-registry/.
206. van Houwelingen HC, Putter H. Dynamic predicting by landmarking as an alternative for multi-state modeling: an application to acute lymphoid leukemia data. Lifetime Data Anal.14(4):447-63.
207. Tompsett DM, Leacy F, Moreno-Betancur M, Heron J, White IR. On the use of the not-at-random fully conditional specification (NARFCS) procedure in practice. Stat Med. 2018;37(15):2338–53.
208. Beesley LJ, Taylor JMG. Accounting for not-at-random missingness through imputation stacking. arXiv:210107954. 2021.

Appendices

Table A.1a. Combinations of events, in event time order, for patients with fully observed event times in the NHS CBB dataset.

Events experienced	N
AgvHD & CGvHD	
AgvHD engraftment CGvHD	6
AgvHD engraftment CGvHD relapse death	2
Engraftment AgvHD CGvHD relapse death	3
Engraftment AgvHD CGvHD relapse	4
Engraftment AgvHD CGvHD death	8
Engraftment AgvHD CGvHD	19
AgvHD w/o CGvHD	
AgvHD engraftment relapse death	9
AgvHD engraftment relapse	2
AgvHD engraftment death	23
AgvHD engraftment	30
Engraftment AgvHD relapse death	7
Engraftment AgvHD relapse	2
Engraftment AgvHD death	14
Engraftment AgvHD	38
AgvHD relapse g.failure death	1
AgvHD g.failure death	3
AgvHD g.failure relapse death	2
AgvHD death	1
CGVHD w/o AgvHD	
CGvHD g.failure relapse death	1
Engraftment CGvHD	6
Engraftment CGvHD death	2
Relapse w/o GvHD	
Relapse g.failure death	3
Relapse death	2
Engraftment relapse death	17
Engraftment relapse	4
G.failure relapse death	2
G.failure relapse	2
Neither GvHD nor relapse	
Engraftment death	30
Engraftment	65
G.failure death	11
G.failure	6
Death only	18
ALL PATIENTS	**342**

AgvHD = acute GvHD; CGVHD = chronic GvHD; engraftment = myeloid engraftment; g.failure = graft failure; w/o = without.

Table A.1b. Combinations of events, in event time order, for cases with at least one missing event time.
** indicates events with missing event times, which are listed first.*

Events experienced	N
AgvHD & CGvHD	
AgvHD* CGvHD* relapse* engraftment death	1
AgvHD* CGvHD* engraftment	18
AgvHD* engraftment CGvHD death	1
AgvHD* engraftment CGvHD	2
CGvHD* AgvHD engraftment death	1
CGvHD* AgvHD engraftment relapse death	1
CGvHD* engraftment AgvHD	3
AgvHD w/o CGvHD	
AgvHD* engraftment death	9
AgvHD* engraftment	20
AgvHD* engraftment relapse death	2
AgvHD* relapse* engraftment death	4
Engraftment* AgvHD death	1
Engraftment* AgvHD	1
G.failure* AgvHD relapse death	1
Relapse* AgvHD engraftment death	1
Relapse* AgvHD engraftment	1
Relapse* engraftment AgvHD death	1
CGVHD w/o AgvHD	
CGvHD* engraftment	5
Relapse w/o GvHD	
Relapse* engraftment death	10
Relapse* g.failure* death	1
Relapse* g.failure death	1
Relapse* g.failure	1
Relapse* death	1
Neither GvHD nor relapse	
Engraftment* death	1
G.failure*	2
ALL PATIENTS	**90**

AgvHD = acute GvHD; CGVHD = chronic GvHD; engraftment = myeloid engraftment; g.failure = graft failure; w/o = without.

Table A.2. Study 1: Standardised bias (StdBias) and average model-based SE (ModSE) of cumulative incidence at largest event time and lower quartile time of acute GvHD given event times (a) MCAR, (b) MAR (dependent on event type only), (c) MAR (dependent on event type, the number of cords received and time to myeloid engraftment), (d) shortest times MNAR, (e) longest times MNAR.

Estimand (true result)		**Cumulative incidence (55.58%)**						**Lower quartile (26.00 days)**					
Proportion of missing times		**10%**		**30%**		**50%**		**10%**		**30%**		**50%**	
Missing data mechanism	**Imputation method**	**Std Bias**	**Mod SE**	**Std Bias**	**Mod SE**	**Std Bias**	**Mod SE**	**Std Bias**	**Mod SE**	**Std Bias**	**Mod SE**	**Std Bias**	**Mod SE**
Complete data		**StdBias=0.00; ModSE= 2.39%**						**StdBias=-0.15; ModSE= 1.83 days**					
MCAR	PMM	0.12	2.43	0.38	2.52	0.63	2.67	0.10	1.92	0.66	2.01	0.99	2.17
	LOGNORM	0.01	2.39	0.03	2.39	0.06	2.42	-0.29	1.98	-0.54	2.23	-0.72	2.56
	MICI	0.43	2.41	1.37	2.44	2.33	2.46	-0.36	2.01	-0.71	2.26	-0.95	2.40
	RESNORM	0.12	2.41	0.42	2.47	0.83	2.57	0.48	1.86	1.98	2.09	3.07	3.01
	NORM	0.32	2.46	0.75	2.56	1.00	2.67	-0.33	2.19	-0.69	3.05	-1.05	4.45
	PMMNOAUX	0.00	2.39	0.02	2.40	0.06	2.44	0.08	2.06	0.43	2.88	0.77	4.90
	NORMNOAUX	0.28	2.45	0.60	2.50	0.71	2.54	-0.29	2.22	-0.56	3.13	-0.90	4.74
	CCA	0.00	2.52	-0.01	2.86	0.00	3.38	-0.14	1.93	-0.13	2.16	-0.16	2.58
	INTCCR	0.39	2.53	1.26	2.52	2.30	2.56	-0.46	2.09	-0.86	2.06	-1.24	2.19
MAR (event type only)	PMM	0.10	2.43	0.44	2.58	1.02	2.84	0.15	1.92	0.84	1.98	1.59	2.19
	LOGNORM	0.01	2.39	0.04	2.42	0.09	2.48	-0.29	1.98	-0.53	2.26	-0.73	2.60
	MICI	0.00	2.39	0.00	2.39	-0.01	2.39	-0.21	1.98	-0.37	2.16	-0.56	2.35
	RESNORM	0.04	2.41	0.23	2.49	0.64	2.69	0.33	1.90	1.46	2.09	2.32	2.74
	NORM	0.37	2.47	1.07	2.60	1.78	2.75	-0.32	2.20	-0.72	3.02	-1.29	4.44
	PMMNOAUX	0.00	2.39	0.03	2.41	0.08	2.47	0.08	2.03	0.49	2.86	0.87	4.93
	NORMNOAUX	0.34	2.46	0.95	2.56	1.52	2.62	-0.27	2.21	-0.60	3.12	-0.97	4.89
	MID	0.00	2.39	0.00	2.39	0.00	2.39	0.94	1.70	2.49	2.15	2.87	4.26
	MED	0.00	2.39	0.00	2.39	0.00	2.39	1.05	0.86	1.28	0.14	1.07	0.09
	CCA	-1.02	2.47	-3.04	2.63	-5.11	2.76	0.41	1.83	1.67	2.05	1.70	4.90
	INTCCR	-0.07	2.56	-0.05	2.57	-0.09	2.60	-0.18	2.15	-0.50	2.26	-0.63	2.45

Estimand (true result)		Cumulative incidence (55.58%)						Lower quartile (26.00 days)					
Proportion of missing times		10%		30%		50%		10%		30%		50%	
Missing data mechanism	Imputation method	Std Bias	Mod SE	Std Bias	Mod SE	Std Bias	Mod SE	Std Bias	Mod SE	Std Bias	Mod SE	Std Bias	Mod SE
Complete data		StdBias=0.00; ModSE= 2.39%						StdBias=-0.15; ModSE= 1.83 days					
MAR (event type + covars)	PMM	0.02	2.39	0.08	2.41	0.23	2.45	-0.39	2.03	-0.97	2.23	0.06	2.58
	LOGNORM	-0.01	2.39	0.00	2.39	0.01	2.39	-1.13	2.04	-4.09	1.88	-3.91	2.36
	MICI	-0.01	2.39	-0.01	2.39	-0.01	2.39	-0.65	2.02	-1.90	2.02	-1.96	2.09
	RESNORM	0.00	2.39	0.01	2.39	0.07	2.41	-0.12	1.96	-0.04	2.21	0.93	3.12
	NORM	0.20	2.43	0.55	2.50	0.95	2.57	-1.12	2.25	-3.97	2.59	-5.02	8.01
	PMMNOAUX	0.00	2.39	0.02	2.41	0.07	2.47	-0.39	2.10	-0.92	3.25	0.07	6.11
	NORMNOAUX	0.33	2.46	0.94	2.56	1.52	2.62	-0.73	2.23	-2.22	2.78	-1.97	4.72
	MID	-0.01	2.39	-0.01	2.39	-0.01	2.39	0.60	1.74	1.89	2.43	2.89	4.56
	MED	-0.01	2.39	-0.01	2.39	-0.01	2.39	0.40	0.87	-0.71	0.14	-0.54	0.09
	CCA	-1.02	2.47	-3.05	2.62	-5.11	2.75	-0.07	1.88	0.15	2.14	1.10	7.00
	INTCCR	-0.08	2.60	-0.13	2.67	-0.14	2.74	-0.61	2.19	-1.65	2.28	-1.28	2.87
MNAR (shortest times missing)	PMM	0.16	2.45	0.75	2.71	1.28	2.94	1.94	1.48	4.14	1.72	4.38	3.15
	LOGNORM	0.01	2.39	0.08	2.41	0.23	2.54	1.41	1.50	3.55	1.72	4.81	2.89
	MICI	0.00	2.39	0.00	2.39	0.00	2.39	1.26	1.50	3.27	1.46	5.63	1.36
	RESNORM	0.07	2.42	0.49	2.60	0.92	2.80	1.81	1.50	4.22	1.84	5.11	2.84
	NORM	0.49	2.50	1.37	2.68	2.06	2.82	1.42	1.63	3.01	1.98	3.99	3.64
	PMMNOAUX	0.01	2.39	0.05	2.42	0.13	2.51	1.74	1.52	3.93	2.09	3.96	4.65
	NORMNOAUX	0.34	2.46	0.97	2.56	1.57	2.63	1.34	1.67	3.06	2.04	3.85	3.82
	MID	0.00	2.39	0.00	2.39	0.00	2.39	2.14	1.48	4.34	2.24	4.78	2.99
	MED	0.00	2.39	0.00	2.39	0.00	2.39	2.87	0.62	5.44	0.13	5.05	0.13
	CCA	-1.01	2.47	-3.04	2.63	-5.10	2.75	2.06	1.45	4.24	1.91	2.81	6.85
	INTCCR	-0.04	2.55	-0.03	2.52	-0.05	2.55	1.16	1.78	4.01	1.57	4.29	1.74

Estimand (true result)		Cumulative incidence (55.58%)						Lower quartile (26.00 days)					
Proportion of missing times		**10%**		**30%**		**50%**		**10%**		**30%**		**50%**	
Missing data mechanism	**Imputation method**	**Std Bias**	**Mod SE**	**Std Bias**	**Mod SE**	**Std Bias**	**Mod SE**	**Std Bias**	**Mod SE**	**Std Bias**	**Mod SE**	**Std Bias**	**Mod SE**
Complete data		**StdBias=0.00; ModSE= 2.39%**						**StdBias=-0.15; ModSE= 1.83 days**					
MNAR (longest times missing)	PMM	0.31	2.62	0.83	2.90	0.87	2.87	-1.02	2.02	-3.05	1.84	-6.92	1.25
	LOGNORM	0.01	2.40	0.02	2.41	0.01	2.39	-1.44	1.97	-6.73	1.31	-10.51	1.26
	MICI	-0.01	2.39	-0.01	2.39	-0.01	2.39	-1.32	1.99	-6.10	1.28	-9.06	1.08
	RESNORM	0.17	2.49	0.54	2.68	0.57	2.69	-0.72	1.98	-1.91	2.02	-4.55	1.67
	NORM	0.47	2.54	1.27	2.73	1.44	2.75	-1.42	2.17	-5.99	1.74	-11.27	2.03
	PMMNOAUX	-0.01	2.39	0.01	2.40	0.04	2.44	-1.09	2.18	-3.20	3.20	-3.87	4.37
	NORMNOAUX	0.33	2.46	0.93	2.56	1.48	2.62	-1.43	2.20	-6.25	1.82	-9.43	2.28
	MID	-0.01	2.39	-0.01	2.39	-0.01	2.39	-0.15	1.83	-0.16	1.81	-0.17	3.70
	MED	-0.01	2.39	-0.01	2.39	-0.01	2.39	-0.40	0.89	-3.56	0.13	-7.99	0.08
	CCA	-1.02	2.47	-3.05	2.62	-5.11	2.75	-0.77	1.91	-2.17	1.82	-4.97	1.40
	INTCCR	-0.13	2.74	-0.17	2.75	-0.15	2.76	-1.20	2.31	-4.19	1.75	-7.48	1.18

PMM, MI by Type 1 predictive mean matching; PMMNOAUX, as for PMM excluding the auxiliary variable from the imputation model; LOGNORM, MI by log-normal imputation with post-imputation back-transformation; MICI, Delord and Genin's MI method; RESNORM, MI by normal regression with restrictions on the imputed values; NORM, MI by normal regression; NORMNOAUX, as for NORM excluding the auxiliary variable from the imputation model; MID, replacement with interval mid-point; MED, replacement with median; CCA, complete case analysis; INTCCR, semi-parametric maximum likelihood method of Bakoyannis *et al*.

In all cases, Monte Carlo SE for bias was <0.1% for cumulative incidence and <0.3 days for the lower quartile.

Table A.3. Study 2: Standardised bias (StdBias) and average model-based SE (ModSE) of cumulative incidence at 100 days and median time of acute GvHD given event times (a) MCAR, (b) MAR (dependent on event type only), (c) MAR (dependent on event type and the number of CB units transplanted, (d) shortest times MNAR, (e) longest times MNAR.

Estimand (true result)		**Cumulative incidence (63.11%)**						**Median (44.00 days)**					
Complete data		**StdBias=0.02; ModSE= 2.15%**						**StdBias=0.19; ModSE=3.60 days**					
Proportion of missing times		**10%**		**30%**		**50%**		**10%**		**30%**		**50%**	
Missing data mechanism	**Imputation method**	**Std Bias**	**Mod SE**	**Std Bias**	**Mod SE**	**Std Bias**	**Mod SE**	**Std Bias**	**Mod SE**	**Std Bias**	**Mod SE**	**Std Bias**	**Mod SE**
MCAR	PMM	0.03	2.17	0.02	2.20	0.04	2.26	0.20	3.73	0.24	4.17	0.31	4.94
	LOGNORM	-0.02	2.17	-0.12	2.22	-0.23	2.27	0.28	3.78	0.41	4.24	0.52	4.87
	MICI	0.19	2.15	0.52	2.15	0.88	2.15	-0.09	3.61	-0.09	3.60	-0.18	3.69
	RESNORM	0.11	2.15	0.27	2.15	0.44	2.14	0.85	4.11	2.02	4.78	3.06	5.04
	NORM	-0.12	2.19	-0.40	2.26	-0.65	2.36	0.67	4.19	1.50	5.38	2.12	6.71
	PMMNOAUX	0.02	2.16	0.03	2.18	0.07	2.21	0.19	3.71	0.19	3.97	0.21	4.26
	NORMNOAUX	-0.12	2.19	-0.38	2.26	-0.65	2.36	0.65	4.14	1.47	5.39	2.11	6.61
	CCA	0.04	2.27	0.03	2.58	0.08	3.04	0.20	3.78	0.21	4.39	0.19	5.53
	INTCCR	0.09	2.16	0.47	2.23	0.83	2.49	0.09	4.12	-0.15	4.25	-0.32	5.71
MAR (event type only, 5 imputations)	PMM	0.02	2.17	0.01	2.20	-0.01	2.27	0.21	3.73	0.26	4.17	0.34	5.02
	LOGNORM	-0.03	2.17	-0.16	2.22	-0.33	2.29	0.27	3.82	0.44	4.37	0.62	5.17
	MICI	0.10	2.15	0.27	2.15	0.44	2.14	0.13	3.63	0.03	3.71	0.06	3.89
	RESNORM	0.10	2.15	0.27	2.15	0.44	2.14	0.86	4.16	2.01	4.92	3.25	5.31
	RESNORM500	0.10	2.15	0.27	2.15	0.44	2.14	0.86	4.16	2.01	4.92	3.25	5.31
	NORM	-0.14	2.19	-0.53	2.28	-1.04	2.43	0.67	4.25	1.48	5.69	2.16	7.48
	PMMNOAUX	0.02	2.17	0.01	2.19	0.01	2.22	0.20	3.71	0.22	4.03	0.24	4.60
	NORMNOAUX	-0.14	2.19	-0.53	2.28	-1.04	2.44	0.67	4.23	1.48	5.67	2.18	7.59
	NORMSUBGRP	0.02	2.16	0.02	2.18	0.02	2.20	0.20	3.70	0.20	3.94	0.19	4.13
	CCA*	-0.98	2.26	-2.99	2.48	-5.02	2.71	0.92	4.74	n/a	n/a	n/a	n/a
	INTCCR	0.05	2.15	0.25	2.17	0.41	2.20	0.17	4.10	0.05	4.16	-0.04	4.42

Estimand (true result)		Cumulative incidence (63.11%)						Median (44.00 days)					
Complete data		StdBias=0.02; ModSE= 2.15%						StdBias=0.19; ModSE=3.60 days					
Proportion of missing times		10%		30%		50%		10%		30%		50%	
Missing data mechanism	Imputation method	Std Bias	Mod SE	Std Bias	Mod SE	Std Bias	Mod SE	Std Bias	Mod SE	Std Bias	Mod SE	Std Bias	Mod SE
MAR (event type only, 50 imputations)	PMM	0.02	2.16	0.01	2.19	-0.01	2.25	0.21	3.70	0.26	4.08	0.34	4.88
	LOGNORM	-0.03	2.17	-0.17	2.21	-0.34	2.27	0.27	3.79	0.44	4.30	0.63	5.04
	MICI	0.10	2.15	0.27	2.15	0.44	2.14	0.13	3.61	0.02	3.68	0.06	3.83
	RESNORM	0.10	2.15	0.27	2.15	0.44	2.14	0.87	4.14	2.07	4.85	3.32	5.20
	NORM	-0.14	2.18	-0.54	2.26	-1.05	2.39	0.68	4.21	1.50	5.59	2.22	7.29
	PMMNOAUX	0.02	2.16	0.02	2.18	0.02	2.21	0.18	3.72	0.20	3.94	0.23	4.38
	NORMNOAUX	-0.14	2.18	-0.54	2.26	-1.05	2.39	0.68	4.21	1.50	5.58	2.23	7.30
	CCA*	-0.98	2.26	-2.99	2.48	-5.02	2.71	0.92	4.74	n/a	n/a	n/a	n/a
MAR (event type + covar)	PMM	0.02	2.17	-0.01	2.24	-0.19	2.50	0.19	3.78	0.30	4.77	0.68	7.12
	LOGNORM	-0.02	2.17	-0.17	2.23	-0.37	2.37	0.25	3.85	0.44	4.54	0.56	6.03
	MICI	0.11	2.15	0.27	2.15	0.45	2.14	0.11	3.66	0.01	3.72	0.04	3.89
	RESNORM	0.11	2.15	0.27	2.15	0.45	2.14	0.85	4.17	2.01	4.98	2.94	5.35
	NORM	-0.14	2.19	-0.53	2.31	-1.01	2.63	0.66	4.30	1.43	5.99	1.68	8.89
	PMMNOAUX	0.03	2.16	0.03	2.19	0.02	2.22	0.17	3.76	0.19	4.03	0.23	4.60
	NORMNOAUX	-0.13	2.19	-0.52	2.28	-1.03	2.44	0.65	4.25	1.47	5.63	2.18	7.57
	NORMSUBGRP	0.02	2.17	0.03	2.18	0.03	2.20	0.17	3.74	0.18	3.94	0.17	4.12
	CCA*	-0.98	2.26	-2.98	2.48	-5.03	2.71	0.90	4.76	n/a	n/a	n/a	n/a
	INTCCR	0.05	2.16	0.23	2.17	0.42	2.18	0.17	4.16	0.04	4.17	-0.07	4.34
MNAR (shortest times missing)	PMM	-0.07	2.17	-0.36	2.25	-0.85	2.48	0.84	3.81	2.12	4.51	3.25	5.62
	LOGNORM	-0.13	2.18	-0.66	2.27	-1.68	2.51	0.95	3.92	2.43	4.90	3.54	6.92
	MICI	0.02	2.15	0.02	2.15	0.02	2.15	0.78	3.71	1.97	3.82	3.25	3.88
	RESNORM	0.02	2.15	0.02	2.15	0.02	2.15	1.41	4.14	3.43	4.75	5.56	4.66
	NORM	-0.25	2.19	-1.01	2.33	-2.14	2.59	1.25	4.30	2.90	5.92	4.14	7.71

Estimand (true result)		Cumulative incidence (63.11%)						Median (44.00 days)					
Complete data		StdBias=0.02; ModSE= 2.15%						StdBias=0.19; ModSE=3.60 days					
Proportion of missing times		10%		30%		50%		10%		30%		50%	
Missing data mechanism	Imputation method	Std Bias	Mod SE	Std Bias	Mod SE	Std Bias	Mod SE	Std Bias	Mod SE	Std Bias	Mod SE	Std Bias	Mod SE
MNAR (shortest times missing)	PMMNOAUX	-0.07	2.17	-0.34	2.23	-0.83	2.44	0.84	3.80	2.04	4.38	3.25	5.44
	NORMNOAUX	-0.24	2.19	-1.01	2.33	-2.16	2.61	1.25	4.25	2.88	5.81	4.16	7.79
	CCA*	-1.06	2.26	-3.19	2.48	-5.31	2.71	1.40	4.77	n/a	n/a	n/a	n/a
	INTCCR	-0.03	2.15	0.00	2.17	-0.01	2.18	0.80	4.15	2.10	4.19	3.56	4.28
MNAR (largest times missing)	PMM	0.84	2.13	0.84	2.13	0.84	2.13	-2.46	2.42	-10.31	1.35	-20.04	1.04
	LOGNORM	0.78	2.14	0.76	2.14	0.77	2.14	-2.37	2.46	-9.55	1.49	-18.54	1.17
	MICI	0.84	2.13	0.84	2.13	0.84	2.13	-2.45	2.42	-10.42	1.32	-20.35	0.82
	RESNORM	0.84	2.13	0.84	2.13	0.84	2.13	-1.48	2.68	-5.05	2.40	-0.50	6.71
	NORM	0.69	2.15	0.44	2.20	0.12	2.26	-1.79	2.69	-7.01	1.97	-3.51	6.51
	PMMNOAUX	0.84	2.13	0.84	2.13	0.84	2.13	-2.46	2.41	-10.41	1.31	-20.48	0.81
	NORMNOAUX	0.69	2.15	0.44	2.20	0.12	2.27	-1.78	2.70	-7.09	1.97	-3.63	6.50
	CCA*	-0.25	2.24	-2.39	2.47	-4.54	2.72	-1.28	2.77	n/a	n/a	n/a	n/a
	INTCCR	0.79	2.22	0.74	2.23	0.75	2.33	-2.11	2.67	-9.07	1.52	-17.10	1.34

*In these complete case analyses, less than 50% patients experienced acute GvHD so the median time to acute GvHD could not be estimated.
PMM, MI by Type 1 predictive mean matching; PMMNOAUX, as for PMM excluding the auxiliary variable from the imputation model; LOGNORM, MI by log-normal imputation with post-imputation back-transformation; MICI, Delord and Genin's MI method; RESNORM, MI by normal regression with restrictions on the imputed values and boundary comparison performed up to 200 times; RESNORM500, MI by normal regression with restrictions on the imputed values and boundary comparison performed up to 500 times; NORM, MI by normal regression; NORMNOAUX, as for NORM excluding the auxiliary variable from the imputation model; NORMSUBGRP, as for NORMNOAUX with the imputation model limited to cases of acute GvHD; CCA, complete case analysis; INTCCR, semi-parametric maximum likelihood method of Bakoyannis *et al*.

In all cases, Monte Carlo SE for bias was <0.1% for cumulative incidence and <0.3 days for the median.

Table A.4a. Bias, average model-based SE (ModSE), empirical SE (EmpSE) and standardised bias (StdBias) of regression parameters $\boldsymbol{\beta}_{lm}$ and coverage (Cov) for regression parameter γ_{12} for each transition intensity model, given various missing data mechanisms and imputation methods.

Estimand (true result)		β_{01}^{1} **(-0.8)**				β_{02}^{1} **(1.2)**				β_{12}^{1} **(1.2)**				β_{12}^{2} **(-1.0)**				γ_{12} **(0)**
Missing data mechanism	**Imputation method**	**Bias**	**Mod SE**	**Emp SE**	**Std Bias**	**Bias**	**Mod SE**	**Emp SE**	**Std Bias**	**Bias**	**Mod SE**	**Emp SE**	**Std Bias**	**Bias**	**Mod SE**	**Emp SE**	**Std Bias**	**Cov**
Complete data (Cox)		0.00	0.20	0.20	0.02	0.00	0.16	0.16	-0.01	0.02	0.21	0.21	0.09	0.00	0.13	0.13	-0.01	0.94
Complete data (Weibull)		0.00	0.20	0.20	0.01	0.00	0.16	0.16	0.02	0.02	0.21	0.21	0.11	-0.01	0.13	0.13	-0.04	0.95
MCAR	CCA	-0.01	0.29	0.30	-0.02	-0.01	0.23	0.23	-0.03	0.05	0.31	0.30	0.15	-0.01	0.18	0.19	-0.07	0.95
	CCA*	-0.01	0.29	0.29	-0.02	0.01	0.23	0.22	0.05	0.06	0.30	0.31	0.20	-0.02	0.18	0.19	-0.09	0.95
	PMM	0.00	0.25	0.21	0.02	-0.01	0.19	0.16	-0.05	-0.13	0.31	0.23	-0.57	0.01	0.16	0.15	0.09	0.95
	PMM*	0.00	0.27	0.22	0.01	-0.03	0.23	0.17	-0.16	-0.13	0.33	0.25	-0.54	0.00	0.20	0.16	0.03	0.95
	PMM30IMP	-0.01	0.24	0.21	-0.03	-0.01	0.19	0.15	-0.09	-0.13	0.30	0.22	-0.58	0.01	0.16	0.14	0.08	0.96
	PMMSUBGP	-0.02	0.24	0.23	-0.10	-0.04	0.18	0.17	-0.23	-0.06	0.29	0.26	-0.22	0.01	0.16	0.15	0.04	0.94
	PMMSUBGP*	-0.03	0.24	0.23	-0.12	-0.03	0.18	0.16	-0.17	-0.05	0.29	0.25	-0.18	-0.02	0.16	0.15	-0.11	0.95
	PMMCOMP	0.02	0.23	0.22	0.09	-0.01	0.17	0.17	-0.03	-0.07	0.37	0.32	-0.22	-0.01	0.20	0.20	-0.06	0.96
	NORMSUBGP	0.00	0.26	0.21	0.02	-0.11	0.20	0.15	-0.68	-0.15	0.32	0.24	-0.64	0.05	0.17	0.15	0.31	0.95
MAR (acute GvHD only)	CCA	-0.15	0.24	0.25	-0.60	-0.15	0.16	0.16	-0.95	0.03	0.25	0.25	0.13	-0.01	0.16	0.16	-0.04	0.93
	PMM	0.01	0.22	0.20	0.03	-0.01	0.16	0.16	-0.08	0.01	0.21	0.21	0.02	-0.01	0.13	0.13	-0.09	0.97
	PMMSUBGP	-0.02	0.22	0.22	-0.09	-0.02	0.16	0.16	-0.13	0.01	0.21	0.21	0.03	-0.01	0.13	0.13	-0.08	0.94
	PMMSUBGP*	-0.02	0.22	0.22	-0.08	-0.01	0.16	0.16	-0.08	0.01	0.21	0.21	0.05	-0.02	0.13	0.13	-0.16	0.95
	NORMSUBGP	0.06	0.22	0.19	0.32	0.01	0.16	0.15	0.04	0.01	0.22	0.21	0.06	-0.01	0.13	0.13	-0.10	0.93
MNAR (smallest acute GvHD only)	CCA	0.00	0.24	0.24	0.01	-0.06	0.16	0.16	-0.39	0.03	0.25	0.25	0.14	0.00	0.15	0.16	-0.01	0.94
	PMM	0.11	0.23	0.22	0.51	0.04	0.16	0.15	0.26	0.02	0.22	0.22	0.09	0.00	0.13	0.13	0.00	0.96
	PMMSUBGP	0.09	0.23	0.23	0.38	0.04	0.15	0.15	0.23	0.02	0.22	0.22	0.10	0.00	0.13	0.13	0.00	0.96
	PMMSUBGP*	0.03	0.23	0.26	0.11	0.04	0.15	0.15	0.27	0.02	0.22	0.22	0.11	0.00	0.13	0.13	-0.02	0.95
	NORMSUBGP	0.12	0.23	0.22	0.54	0.04	0.15	0.15	0.25	0.02	0.23	0.22	0.10	0.00	0.13	0.13	-0.01	0.95
MAR (relapse/ death only, 0→2 transition)	CCA	0.17	0.20	0.20	0.77	0.16	0.19	0.20	0.80	0.02	0.21	0.21	0.09	0.00	0.13	0.13	-0.01	0.94
	PMM	0.06	0.25	0.21	0.29	0.03	0.20	0.17	0.18	0.02	0.21	0.21	0.09	0.00	0.13	0.13	-0.01	0.94
	PMMSUBGP	0.01	0.21	0.21	0.05	-0.01	0.17	0.16	-0.04	0.02	0.21	0.21	0.09	0.00	0.13	0.13	-0.01	0.94
	PMMSUBGP*	0.01	0.21	0.22	0.05	0.00	0.16	0.17	0.03	0.02	0.21	0.21	0.11	-0.01	0.13	0.13	-0.04	0.95
	NORMSUBGP	0.00	0.21	0.21	0.01	0.00	0.17	0.17	0.00	0.02	0.21	0.21	0.09	0.00	0.13	0.13	-0.01	0.94

Estimand (true result)		β_{01}^{1} (-0.8)				β_{02}^{1} (1.2)				β_{12}^{1} (1.2)				β_{12}^{2} (-1.0)				γ_{12} (0)
Missing data mechanism	**Imputation method**	**Bias**	**Mod SE**	**Emp SE**	**Std Bias**	**Bias**	**Mod SE**	**Emp SE**	**Std Bias**	**Bias**	**Mod SE**	**Emp SE**	**Std Bias**	**Bias**	**Mod SE**	**Emp SE**	**Std Bias**	**Cov**
MAR (relapse/ death only, both transitions)	CCA	-0.01	0.24	0.25	-0.03	0.02	0.19	0.19	0.12	0.06	0.26	0.27	0.21	0.05	0.15	0.16	0.30	0.94
	PMM	0.02	0.26	0.22	0.10	0.01	0.22	0.17	0.04	-0.12	0.34	0.26	-0.46	0.02	0.17	0.16	0.14	0.94
	PMMSUBGP	-0.01	0.22	0.22	-0.06	-0.03	0.18	0.17	-0.18	-0.06	0.31	0.28	-0.21	0.03	0.16	0.16	0.16	0.94
	PMMSUBGP*	-0.02	0.22	0.21	-0.10	-0.02	0.18	0.16	-0.13	-0.05	0.30	0.28	-0.18	0.00	0.16	0.16	-0.03	0.94
	NORMSUBGP	-0.08	0.23	0.23	-0.36	-0.03	0.17	0.17	-0.16	-0.17	0.32	0.24	-0.70	0.20	0.15	0.13	1.51	0.96
MNAR (smallest relapse/death only, both transitions)	CCA	0.05	0.21	0.22	0.23	0.00	0.32	0.33	-0.01	0.02	0.22	0.22	0.09	0.00	0.13	0.13	-0.01	0.94
	PMM	-0.37	0.26	0.22	-1.69	-0.30	0.44	0.25	-1.21	0.01	0.23	0.22	0.07	0.00	0.13	0.13	-0.02	0.95
	PMMSUBGP	-0.45	0.22	0.22	-2.07	-0.58	0.23	0.22	-2.61	-0.01	0.23	0.22	-0.03	0.00	0.13	0.13	0.03	0.94
	PMMSUBGP*	-0.45	0.22	0.22	-2.10	-0.57	0.25	0.29	-1.93	0.01	0.22	0.22	0.05	-0.04	0.13	0.13	-0.27	0.92
	NORMSUBGP	-0.49	0.23	0.22	-2.22	-0.59	0.23	0.25	-2.38	-0.08	0.25	0.22	-0.39	0.02	0.13	0.13	0.19	0.94
MAR (acute GvHD & relapse/death)	CCA	-0.18	0.30	0.32	-0.58	-0.15	0.19	0.19	-0.77	0.09	0.32	0.32	0.27	0.04	0.19	0.19	0.22	0.94
	PMM	0.02	0.28	0.23	0.11	0.00	0.22	0.18	0.01	-0.14	0.35	0.26	-0.54	0.01	0.16	0.15	0.08	0.92
	PMMSUBGP	-0.02	0.24	0.23	-0.10	-0.04	0.18	0.16	-0.23	-0.08	0.32	0.27	-0.27	0.02	0.17	0.15	0.15	0.93
	PMMSUBGP*	-0.03	0.24	0.23	-0.13	-0.03	0.18	0.16	-0.19	-0.06	0.31	0.27	-0.22	-0.02	0.17	0.16	-0.14	0.95
MNAR (smallest acute GvHD) & MAR (relapse/death)	CCA	0.00	0.29	0.29	-0.01	-0.04	0.19	0.20	-0.20	0.08	0.31	0.30	0.26	0.05	0.18	0.19	0.24	0.94
	PMM	0.12	0.30	0.24	0.51	0.05	0.22	0.17	0.28	-0.12	0.36	0.27	-0.46	0.01	0.17	0.16	0.09	0.94
	PMMSUBGP	0.10	0.26	0.24	0.40	0.03	0.17	0.16	0.17	-0.07	0.33	0.29	-0.24	0.03	0.17	0.15	0.18	0.96
	PMMSUBGP*	0.04	0.26	0.28	0.15	0.04	0.17	0.16	0.23	-0.05	0.33	0.27	-0.19	0.00	0.17	0.16	-0.03	0.95
MNAR (largest acute GvHD) & MAR (relapse/ death)	CCA	-0.38	0.30	0.33	-1.17	-0.40	0.20	0.22	-1.85	0.07	0.32	0.33	0.22	0.04	0.18	0.19	0.24	0.94
	PMM	-0.15	0.28	0.25	-0.59	-0.21	0.27	0.21	-1.01	-0.14	0.35	0.27	-0.51	0.00	0.17	0.15	0.01	0.97
	PMMSUBGP	-0.19	0.22	0.23	-0.81	-0.31	0.19	0.20	-1.56	-0.11	0.31	0.27	-0.42	0.02	0.17	0.15	0.14	0.95
	PMMSUBGP*	-0.31	0.23	0.26	-1.20	-0.21	0.18	0.17	-1.24	-0.11	0.30	0.26	-0.42	-0.03	0.17	0.16	-0.16	0.95
MAR (acute GvHD) & MNAR (smallest relapse/death)	CCA	-0.11	0.25	0.27	-0.40	-0.14	0.32	0.33	-0.43	0.03	0.26	0.27	0.11	0.00	0.16	0.17	0.00	0.94
	PMM	-0.38	0.26	0.22	-1.74	-0.31	0.43	0.26	-1.21	0.00	0.23	0.22	0.00	-0.01	0.13	0.13	-0.05	0.94
	PMMSUBGP	-0.46	0.23	0.22	-2.10	0.61	0.24	0.24	-2.57	-0.02	0.23	0.22	-0.10	0.00	0.13	0.13	0.00	0.94
	PMMSUBGP*	-0.47	0.23	0.22	-2.13	-0.59	0.27	0.31	-1.90	0.00	0.23	0.22	0.00	-0.04	0.13	0.14	-0.31	0.94

Estimand (true result)		β_{01}^{1} (-0.8)				β_{02}^{1} (1.2)				β_{12}^{1} (1.2)				β_{12}^{2} (-1.0)				γ_{12} (0)
Missing data mechanism	**Imputation method**	**Bias**	**Mod SE**	**Emp SE**	**Std Bias**	**Bias**	**Mod SE**	**Emp SE**	**Std Bias**	**Bias**	**Mod SE**	**Emp SE**	**Std Bias**	**Bias**	**Mod SE**	**Emp SE**	**Std Bias**	**Cov**
MAR (acute GvHD) & MNAR (largest relapse/death)	CCA	-0.16	0.27	0.27	-0.60	-0.52	0.16	0.16	-3.30	-0.37	0.28	0.30	-1.25	-0.09	0.23	0.24	-0.39	0.95
	PMM	-0.01	0.23	0.21	-0.05	-0.13	0.18	0.16	-0.79	-0.57	0.25	0.20	-2.80	0.32	0.17	0.13	2.50	0.98
	PMMSUBGP	-0.03	0.23	0.22	-0.12	-0.15	0.18	0.16	-0.90	-0.54	0.25	0.22	-2.52	0.31	0.18	0.13	2.36	0.98
	PMMSUBGP*	-0.03	0.23	0.22	-0.12	-0.14	0.18	0.16	-0.87	-0.48	0.23	0.18	-2.73	0.08	0.16	0.15	0.53	0.66
MNAR (smallest acute GvHD & smallest relapse/death)	CCA	0.01	0.24	0.24	0.06	0.00	0.32	0.33	-0.01	0.03	0.25	0.26	0.13	0.00	0.15	0.16	-0.01	0.94
	PMM	-0.31	0.28	0.23	-1.37	-0.22	0.39	0.24	-0.90	0.00	0.24	0.22	0.02	0.00	0.13	0.13	0.01	0.94
	PMMSUBGP	-0.41	0.24	0.23	-1.74	-0.46	0.23	0.22	-2.08	-0.01	0.24	0.22	-0.05	0.01	0.13	0.13	0.05	0.95
	PMMSUBGP*	-0.44	0.26	0.26	-1.67	-0.45	0.25	0.29	-1.54	0.00	0.23	0.22	-0.01	-0.01	0.13	0.13	-0.06	0.95

State indicators: 0 = transplanted; 1 = acute GvHD; 2 = relapse/death.

Parameters β_{lm}^{1}, β_{lm}^{2}, γ_{12} are for disease status at time of transplant, number of CB units transplanted and time from transplant until acute GvHD, respectively.

Monte Carlo SE for bias/coverage ranges from 0.004 to 0.015 for all estimands.

*Cox models were fit, except for methods indicated with a *, for which Weibull models were fit.

CCA, complete case analysis;

PMM, MI by Type 1 predictive mean matching;

PMM30IMP, as for PMM, with number of imputations = 30;

PMMSUBGP, as for PMM with imputation models fit separately for patients with and without acute GvHD;

PMMCOMP, as for PMM, imputing acute GvHD time and time from acute GvHD to relapse/death with post-imputation calculation of relapse/death time;

NORMSUBGP, MI by normal regression with imputation models fit separately for patients with and without acute GvHD.

Table A.4b. Bias, average model-based SE (ModSE), empirical SE (EmpSE) and standardised bias (StdBias) of RELOS between 0 and 2 years, $e_m(2)$, for each transition intensity model given various missing data mechanisms and imputation methods.

Estimand (true result)		$e_0(2)$ **(26.3)**				$e_1(2)$ **(140.6)**				$e_2(2)$ **(563.1)**			
Missing data mechanism	**Imputation method**	**Bias**	**Mod SE**	**Emp SE**	**Std Bias**	**Bias**	**Mod SE**	**Emp SE**	**Std Bias**	**Bias**	**Mod SE**	**Emp SE**	**Std Bias**
Complete data (Cox)		-0.1	1.4	1.1	-0.05	-0.5	11.4	11.3	-0.04	-13.4	17.8	16.8	-0.79
Complete data (Weibull)		0.0	1.0	1.0	0.02	-0.2	11.4	11.3	-0.02	0.2	11.5	11.4	0.01
MCAR	CCA	-0.2	2.7	1.8	-0.09	-1.4	16.3	16.6	-0.08	-26.9	29.2	28.7	-0.93
	CCA*	0.0	1.4	1.4	0.01	-0.4	16.2	16.4	-0.02	0.3	16.4	16.6	0.02
	PMM	0.4	2.1	1.5	0.27	-0.5	15.4	13.5	-0.03	-19.4	21.9	22.2	-0.87
	PMM*	-0.2	3.4	2.2	-0.08	-1.4	21.9	16.4	-0.08	-9.8	66.3	38.3	-0.26
	PMM30IMP	0.4	2.2	1.4	0.28	-0.4	15.0	13.0	-0.03	-19.1	22.3	21.2	-0.90
	PMMSUBGP	-0.6	2.1	1.4	-0.40	1.0	15.5	14.1	0.07	-20.1	26.3	22.2	-0.90
	PMMSUBGP*	-0.5	1.7	1.3	-0.35	2.6	15.9	13.5	0.20	-2.9	24.6	15.9	-0.19
	PMMCOMP	0.0	1.8	1.4	-0.01	-2.6	19.7	18.8	-0.14	-19.5	28.9	25.8	-0.76
	NORMSUBGP	4.1	4.5	2.3	1.81	8.6	19.8	14.1	0.61	-30.3	53.3	26.4	-1.15
MAR (acute GvHD only)	CCA	-0.7	2.4	1.4	-0.51	-13.6	11.9	12.0	-1.13	-6.7	22.7	21.4	-0.31
	PMM	-0.2	1.6	1.4	-0.13	-0.1	11.5	11.3	-0.01	-13.6	17.8	16.8	-0.81
	PMMSUBGP	-0.5	1.8	1.4	-0.32	0.0	11.5	11.3	0.00	-13.5	17.9	16.8	-0.80
	PMMSUBGP*	-0.4	1.3	1.3	-0.27	0.8	11.4	11.3	0.07	-0.4	11.6	11.4	-0.04
	NORMSUBGP	5.5	3.8	1.9	2.83	-3.5	11.5	11.1	-0.32	-15.9	17.9	16.9	-0.95
MNAR (smallest acute GvHD only)	CCA	5.0	1.7	1.4	3.59	-12.1	12.5	12.6	-0.96	-11.9	22.5	20.8	-0.57
	PMM	5.6	1.4	1.4	4.11	-3.9	11.4	11.2	-0.35	-15.7	17.9	16.8	-0.93
	PMMSUBGP	5.5	1.7	1.5	3.76	-3.8	11.4	11.2	-0.34	-15.6	17.9	16.8	-0.93
	PMMSUBGP*	5.8	1.2	1.4	4.18	-3.8	11.4	11.1	-0.34	-2.0	11.6	11.4	-0.17
	NORMSUBGP	7.1	2.1	1.5	4.84	-4.8	11.4	11.1	-0.43	-16.2	17.8	16.8	-0.97
MAR (relapse/ death only, 0→2 transition)	CCA	0.6	1.3	1.2	0.50	11.6	12.2	12.0	0.97	-26.1	18.4	17.3	-1.51
	PMM	0.4	1.8	1.2	0.29	-0.5	11.4	11.3	-0.04	-13.8	17.9	16.8	-0.82
	PMMSUBGP	-0.1	1.4	1.2	-0.05	-0.5	11.4	11.3	-0.05	-13.3	17.8	16.8	-0.79
	PMMSUBGP*	0.0	1.0	1.1	0.04	-0.2	11.4	11.3	-0.02	0.2	11.5	11.4	0.02
	NORMSUBGP	0.4	1.3	1.1	0.34	1.2	11.6	11.4	0.10	-15.5	17.9	16.8	-0.92

Estimand (true result)		$e_0(2)$ (26.3)				$e_1(2)$ (140.6)				$e_2(2)$ (563.1)			
Missing data mechanism	**Imputation method**	**Bias**	**Mod SE**	**Emp SE**	**Std Bias**	**Bias**	**Mod SE**	**Emp SE**	**Std Bias**	**Bias**	**Mod SE**	**Emp SE**	**Std Bias**
MAR (relapse/death only, both transitions)	CCA	0.0	2.5	1.4	-0.01	10.7	16.9	16.4	0.65	-28.4	24.1	22.4	-1.27
	PMM	0.1	1.9	1.3	0.06	-1.2	19.6	16.9	-0.07	-16.6	25.1	22.8	-0.73
	PMMSUBGP	-0.3	1.5	1.2	-0.21	1.1	19.3	17.5	0.06	-18.3	24.9	23.0	-0.79
	PMMSUBGP*	-0.2	1.1	1.1	-0.19	2.0	19.6	17.5	0.11	-1.8	19.7	17.6	-0.10
	NORMSUBGP	0.2	1.6	1.2	0.20	35.8	16.1	15.1	2.38	-47.4	20.9	16.3	-2.91
MNAR (smallest relapse/death only, both transitions)	CCA	4.1	1.4	1.3	3.19	38.9	13.7	14.0	2.77	-56.9	19.0	18.3	-3.11
	PMM	12.9	7.6	4.4	2.92	5.5	11.9	12.1	0.45	-32.3	19.1	17.7	-1.82
	PMMSUBGP	5.8	2.5	1.7	3.47	6.1	11.8	12.0	0.51	-25.9	17.8	17.2	-1.50
	PMMSUBGP*	5.9	1.2	1.4	4.35	8.5	11.8	12.3	0.69	-14.4	11.7	12.5	-1.16
	NORMSUBGP	5.9	2.2	1.5	4.01	9.5	12.0	12.2	0.78	-28.5	18.1	16.5	-1.74
MAR (acute GvHD & relapse/death)	CCA	-0.8	4.4	1.9	-0.44	-4.8	17.5	17.5	-0.27	-22.1	30.1	27.8	-0.79
	PMM	0.1	2.2	1.4	0.11	-1.5	19.6	16.2	-0.09	-16.4	25.0	22.6	-0.73
	PMMSUBGP	-0.5	2.2	1.5	-0.36	2.6	20.1	16.3	0.16	-20.1	25.5	22.5	-0.89
	PMMSUBGP*	-0.4	1.4	1.3	-0.35	1.9	19.7	15.9	0.12	-1.5	19.8	16.0	-0.09
MNAR (smallest acute GvHD) & MAR (relapse/death)	CCA	5.2	2.9	1.7	2.99	-1.9	18.5	18.7	-0.10	-27.8	29.3	27.1	-1.03
	PMM	6.1	2.5	1.5	3.97	-5.4	20.5	16.3	-0.33	-18.7	25.8	22.5	-0.83
	PMMSUBGP	5.6	1.7	1.5	3.74	-1.9	20.0	16.1	-0.12	-21.5	25.4	22.6	-0.95
	PMMSUBGP*	5.8	1.3	1.4	4.16	-2.1	20.4	16.6	-0.13	-3.7	20.5	16.8	-0.22
MNAR (largest acute GvHD) & MAR (relapse/death)	CCA	-6.1	2.1	2.0	-3.11	-8.9	18.3	17.7	-0.50	-9.8	30.4	27.5	-0.36
	PMM	-6.4	2.9	1.6	-4.13	1.6	21.2	16.5	0.10	-13.0	26.8	22.5	-0.58
	PMMSUBGP	-6.3	2.0	1.7	-3.62	5.0	20.1	16.0	0.31	-16.7	25.6	21.9	-0.76
	PMMSUBGP*	-8.3	0.9	0.9	-9.07	9.3	19.9	16.2	0.57	-1.0	20.0	16.3	-0.06
MAR (acute GvHD) & MNAR (smallest relapse/death)	CCA	4.5	2.6	1.6	2.82	33.2	15.1	15.9	2.08	-57.4	24.1	22.4	-2.57
	PMM	13.4	7.7	4.7	2.88	5.9	11.9	12.0	0.49	-33.2	19.1	17.8	-1.87
	PMMSUBGP	5.6	3.2	1.8	3.03	6.6	11.9	11.9	0.56	-26.1	17.9	17.2	-1.52
	PMMSUBGP*	5.6	1.4	1.5	3.69	9.2	11.9	12.3	0.74	-14.8	11.8	12.5	-1.18
MAR (acute GvHD) & MNAR (largest relapse/death)	CCA	-2.9	1.4	1.3	-2.23	-104.9	3.6	5.5	-19.04	-443.5	10.9	12.6	-35.14
	PMM	-0.6	1.4	1.3	-0.49	-95.1	4.7	6.1	-15.48	-454.9	5.1	11.8	-38.39
	PMMSUBGP	-1.0	1.5	1.2	-0.82	-94.1	4.7	6.2	-15.27	-455.0	18.8	14.9	-30.53
	PMMSUBGP*	-1.0	1.5	1.3	-0.72	-73.1	9.0	9.4	-7.82	74.1	9.2	9.7	7.66

Estimand (true result)		$e_0(2)$ (26.3)				$e_1(2)$ (140.6)				$e_2(2)$ (563.1)			
Missing data mechanism	**Imputation method**	**Bias**	**Mod SE**	**Emp SE**	**Std Bias**	**Bias**	**Mod SE**	**Emp SE**	**Std Bias**	**Bias**	**Mod SE**	**Emp SE**	**Std Bias**
MNAR (smallest acute GvHD & smallest relapse/death)	CCA	11.6	1.9	1.6	7.19	30.0	15.7	16.2	1.85	-60.5	23.9	22.5	-2.69
	PMM	19.4	7.5	4.5	4.31	0.8	11.8	11.9	0.07	-34.1	19.0	17.5	-1.94
	PMMSUBGP	11.8	3.1	1.9	6.21	1.4	11.8	11.8	0.12	-27.1	17.8	17.1	-1.58
	PMMSUBGP*	12.0	1.4	1.7	7.23	3.2	11.8	12.0	0.27	-15.2	11.9	12.3	-1.23

State indicators: 0 = transplanted; 1 = acute GvHD; 2 = relapse/death.

Monte Carlo SE for bias ranges from 0.03 to 0.14 for $e_0(2)$, from 0.17 to 0.59 for $e_1(2)$ and from 0.31 to 1.21 for $e_2(2)$.

*Cox models were fit, except for methods indicated with a *, for which Weibull models were fit.

CCA, complete case analysis;

PMM, MI by Type 1 predictive mean matching;

PMM30IMP, as for PMM, with number of imputations = 30;

PMMSUBGP, as for PMM with imputation models fit separately for patients with and without acute GvHD;

PMMCOMP, as for PMM, imputing acute GvHD time and time from acute GvHD to relapse/death with post-imputation calculation of relapse/death time;

NORMSUBGP, MI by normal regression with imputation models fit separately for patients with and without acute GvHD.

Table A.5. Hazard ratios (HR) and 95% confidence interval (CI) for covariates in the transition intensity model from acute GvHD to relapse/death for the NHS Cord Blood Bank cohort, including time from transplant to acute GvHD.

Covariate (reference value)	**Missing data method**									
	CCA (N=116)		**PMMSUBGP (N=432)**		**PMMSUBGP Weibull (N=432)**		**PMMSUBGP proxy method (N=432)**		**PMM (N=432)**	
	HR	**95% CI**	**HR**	**95% CI**	**HR**	**95% CI**	**HR**	**95% CI**	**HR**	**95% CI**
Double cord transplant (single)	0.21	0.02-2.55	0.42	0.20-0.88	0.39	0.19-0.83	0.39	0.18-0.84	0.56	0.24-1.26
Patient age (10-year increments)	1.52	0.96-2.39	1.10	0.92-1.31	1.11	0.93-1.32	1.16	0.98-1.38	1.15	0.95-1.38
Disease type (acute leukaemia)										
Other blood cancer[1]	1.58	0.36-7.00	1.09	0.59-2.03	1.13	0.61-2.09	1.04	0.57-1.90	0.93	0.50-1.73
Non-malignant disorder[2]	2.32	0.25-21.65	1.37	0.51-3.72	1.35	0.51-3.59	1.35	0.50-3.68	1.08	0.39-3.00
Disease status at time of transplant (in remission)										
Relapse	2.40	0.53-10.96	1.03	0.38-2.82	0.92	0.36-2.31	1.43	0.57-3.61	2.39	0.86-6.66
Other[3]	2.01	0.27-15.11	1.14	0.50-2.60	1.05	0.48-2.32	1.30	0.60-2.81	1.61	0.69-3.76
Reduced intensity conditioning regimen (intensive)	0.40	0.09-1.82	0.99	0.53-1.83	1.00	0.54-1.82	0.86	0.48-1.55	0.79	0.44-1.41
Donor-recipient CMV match (-/-)										
-/+	1.01	0.33-3.06	1.90	0.94-3.83	1.96	0.98-3.90	1.71	0.85-3.48	1.68	0.81-3.48
+/-	1.19	0.33-4.26	1.29	0.62-2.70	1.27	0.62-2.62	1.32	0.63-2.79	1.33	0.61-2.86
+/+	0.60	0.12-2.92	0.80	0.29-2.25	0.85	0.32-2.30	0.74	0.26-2.11	1.01	0.36-2.86
Donor-recipient sex match (F/F)										
F/M	4.81	1.11-20.76	1.56	0.71-3.43	1.52	0.71-3.26	1.66	0.77-3.55	1.52	0.67-3.44
M/F	4.53	0.92-22.16	1.46	0.63-3.42	1.49	0.65-3.39	1.49	0.64-3.50	1.33	0.55-3.22
M/M	1.90	0.35-10.26	1.49	0.62-3.58	1.44	0.60-3.48	1.36	0.57-3.29	1.47	0.61-3.58
Number of donor-recipient HLA mismatches[4] (Well-matched: 0/1)										
Not well-matched: 2 or more	1.63	0.56-4.76	1.76	0.92-3.38	1.90	1.02-3.54	1.72	0.86-3.44	1.52	0.79-2.91

Covariate (reference value)	Missing data method									
	CCA (N=116)		PMMSUBGP (N=432)		PMMSUBGP Weibull (N=432)		PMMSUBGP proxy method (N=432)		PMM (N=432)	
	HR	95% CI	HR	95% CI	HR	95% CI	HR	95% CI	HR	95% CI
TNC dose at infusion $\times 10^7$/kg (Low: <3.0)										
Medium: 3.0-5.0	0.88	0.25-3.01	0.63	0.36-1.09	0.64	0.37-1.09	0.80	0.45-1.43	0.80	0.40-1.58
High: > 5.0	0.54	0.07-4.02	0.32	0.15-0.70	0.34	0.16-0.72	0.41	0.19-0.91	0.49	0.22-1.09
Time of acute GvHD	1.01	0.99-1.03	1.00	0.99-1.01	1.00	0.99-1.01	1.00	0.99-1.01	1.00	0.99-1.01

CMV, cytomegalovirus; HLA, human leucocyte antigen; TNC, total nucleated cells.
Unless otherwise stated, Cox transition intensity models were fitted.
CCA, complete case analysis.
PMMSUBGP, FCS MI by type 1 predictive mean matching with imputation models fit separately for patients experiencing both acute and chronic GvHD or chronic GvHD without acute GvHD, acute GvHD without chronic GvHD, relapse without GvHD, and neither GvHD nor relapse.
PMMSUBGP proxy method, as for PMMSUBGP, with observed donor information used in imputation models for all other variables.
PMM, FCS MI by Type 1 predictive mean matching with one imputation model fit for all patients.
[1] Other blood cancer includes lymphoproliferative and plasma cell disorders, myelodysplastic syndromes and myeloproliferative disorders.
[2] Non-malignant disorder includes histiocytic disorder, solid tumour, bone marrow failure syndrome, haemoglobinopathy, primary immune deficiency and inborn error of metabolism.
[3] Other disease status includes acute, chronic and accelerated phase, refractory disease, transformed to acute leukaemia, blastic crisis, MDS, MDP and non-malignant disorders.
[4] HLA A and B loci at antigenic level and DR-B1 at allelic level

A.6. R Code to generate simulation study results

1. Required libraries

```
library(survival)
library(mstate)
library(MIICD)
library(mice)
library(intccr)
library(doParallel)
library(simsurv)
library(flexsurv)
```

2. Data Generating Mechanisms

2.1. Generating competing risks data for a non-parametric analysis

```
#Create simulation datasets based on average of 65% GvHD, 10% graft failures,
#25% deaths
    sample500=data.frame()
    for (i in 1:1000)
    {
     sample=rbind(cbind.data.frame(agvhd_status_nocens=1,runif=runif(325)),
                cbind.data.frame(agvhd_status_nocens=2,runif=runif(50)),
                cbind.data.frame(agvhd_status_nocens=3,runif=runif(125)))
      #define additional runif for ordering
       sample$rorder=runif(500)
       sample500=rbind.data.frame(sample500,sample)
    }
    #Now simulate times and round times to match observed data
    sample500$agvhd_time=ifelse(sample500$agvhd_status_nocens == 1,
                        ceiling(exp(qnorm(p=sample500$runif,mean=3.26,sd=0.71))),
                        ifelse(sample500$agvhd_status_nocens == 2,

    ceiling(exp(qnorm(p=sample500$runif,mean=3.76,sd=0.66))),

    ceiling(exp(qnorm(p=sample500$runif,mean=4.34,sd=1.3)))))
    #Censoring event times at greater than one year
    sample500$agvhd_status=ifelse(sample500$agvhd_time<=365,sample500$ag
    vhd_status_nocens,0)
    sample500$agvhd_time[sample500$agvhd_status==0]=365
    #also add one covariate - number of cords received - with 45% double cord
    sample500$doublecord=rbinom(500000,1,0.45)
    #Indicator of agvhd
    sample500$agvhd=ifelse(sample500$agvhd_status==1,1,0)
    #Order all obs randomly before adding a sample number for each of the 1000
    datasets
    sample500=sample500[order(sample500$rorder),]
    sample500$sampno=rep(1:1000, 500)
```

2.2. Generating events data for a three-state Markov MSM

```
##First competing risks experiment
#hazard of aGvHD
k_agvhd=1.5
l_agvhd=36

#hazard of relapse/death
k_rel_death=0.9
l_rel_death=120
#also add two covariates
#1. in relapse or not
#2. number of cords received
relapse=rbinom(500000,1,0.2)
doublecord=rbinom(500000,1,0.45)

# add sampno
sampno=rep(1:1000, 500)

covs_all = data.frame(id = 1:500000, relapse, doublecord, sampno)

#simulate fup to 5 years for all pts (censor after all times drawn)
#1. not in relapse
h_all = function(t,x,betas)#function needs this structure in simsurv
  ((k_agvhd/l_agvhd)*(t/l_agvhd)^(k_agvhd-1)
   +
    (k_rel_death/l_rel_death)*(t/l_rel_death)^(k_rel_death-1))

covs=data.frame(id = covs_all[covs_all$relapse==0,1])
simdata = simsurv(hazard = h_all,maxt=365,x=covs,seed=4752)

#assume order of rows is unchanged and drop id from this dataset and merge
back in id from covs
simdata_1=cbind(simdata[,2:3],covs_all[covs_all$relapse==0,])

#now calculate prob of agvhd for binomial draws to determine whether event
of interest or competing event
simdata_1$h_all=h_all(t=simdata_1$eventtime)

h_agvhd = function(t)
  ((k_agvhd/l_agvhd)*(t/l_agvhd)^(k_agvhd-1))

simdata_1$h_agvhd=h_agvhd(t=simdata_1$eventtime)
```

```
#now calculate prob of agvhd at event time
simdata_1$p_agvhd=simdata_1$h_agvhd/simdata_1$h_all

#2. in relapse
#use simsurv
h_all2 = function(t,x,betas)#function needs this structure in simsurv
  (exp(-0.8)*(k_agvhd/l_agvhd)*(t/l_agvhd)^(k_agvhd-1)
   +
    exp(1.2)*(k_rel_death/l_rel_death)*(t/l_rel_death)^(k_rel_death-1))

covs2=data.frame(id = covs_all[covs_all$relapse==1,1])
simdata2 = simsurv(hazard = h_all2,maxt=365,x=covs2,seed=78947)

#assume order of rows is unchanged and drop id from this dataset and merge
#back in id from covs
simdata2_1=cbind(simdata2[,2:3],covs_all[covs_all$relapse==1,])

#now calculate prob of agvhd for binomial draws to determine whether event
#of interest or competing event
simdata2_1$h_all=h_all2(t=simdata2_1$eventtime)

h_agvhd2 = function(t)
  (exp(-0.8)*(k_agvhd/l_agvhd)*(t/l_agvhd)^(k_agvhd-1))

simdata2_1$h_agvhd=h_agvhd2(t=simdata2_1$eventtime)

#now calculate prob of agvhd at event time
simdata2_1$p_agvhd=simdata2_1$h_agvhd/simdata2_1$h_all

##### Combine all rows #####
simdata_all=rbind(simdata_1,simdata2_1)

###Now run binomial experiment to determine event type
for (i in 1:500000) {
simdata_all$agvhd_status[i]=ifelse(rbinom(1,1,simdata_all$p_agvhd[i])==1,1,
0)
}

#rename eventtime for consistency with other progs
names(simdata_all)[names(simdata_all) == 'eventtime'] = 'agvhd_time'

#Need to create relapse time and status cols
simdata_all$relapse_or_death_time=simdata_all$agvhd_time
simdata_all$relapse_or_death_status=ifelse(simdata_all$agvhd_status==0,1,0
)
```

```
#order by id
simdata_all=simdata_all[order(simdata_all$id),]

####Now calculate event times for relapse/death after aGvHD
#Create transition 1->2 times
#Based on conditional survival model
###For simplicity, calculate for all cases, but discard rows for cases with
###agvhd_status=0
#Weibull params
k_rel_death12=0.8
l_rel_death12=160

#First calculate S(t0) ie at time of acute GvHD
simdata_all$S_t0=exp(-
((simdata_all$agvhd_time/l_rel_death12)^k_rel_death12)*exp(1.2*simdata_al
l$relapse - simdata_all$doublecord))
#Generate times using inverse transform sampling and
#discard relapse times with agvhd_status=0 i.e. already in relapse/death
#state
simdata_all$runif=runif(500000)
simdata_all$rel_death_time=ifelse(simdata_all$agvhd_status==1,
                    160*((-log(simdata_all$runif*simdata_all$S_t0)/
                    exp(1.2*simdata_all$relapse -
simdata_all$doublecord))^1.25),
                      simdata_all$relapse_or_death_time)
#Set all rel_death_status to 1 as no censored times at this point
simdata_all$rel_death_status=1

###Now generate censoring times by sampling from uniform dist
#Since all transitions from transplant occurred within the year, only 1-2
#transitions will be censored
simdata_all$cens_time=runif(500000, min = 365, max = 1826)

#update relapse/death time and status for these cases
simdata_all$rel_death_time_uncens=simdata_all$rel_death_time
simdata_all$rel_death_time=ifelse(simdata_all$rel_death_time_uncens>simd
ata_all$cens_time, simdata_all$cens_time,
simdata_all$rel_death_time_uncens)
simdata_all$rel_death_status_uncens=simdata_all$rel_death_status
simdata_all$rel_death_status=ifelse(simdata_all$rel_death_time_uncens>sim
data_all$cens_time, 0,simdata_all$rel_death_status_uncens)
```

3. Functions to apply the missing data mechanisms used in the studies

3.1. MCAR

```
MCAR = function(dset,time,percent)
```

```
# dset = dataset, time = event time, percent = percent missing data as a
decimal
# e.g.
#sample1$time.miss=MCAR(dset=sample1,time=sample1$agvhd_time,pe
#rcent=0.1)
{
 time.miss=c()
 for (j in 1: nrow(dset))
  {time.miss[j]=ifelse(rbinom(1,1,percent)==1,NA,time[j])}
 return(time.miss)
}
```

3.2. MAR with missingness dependent on event type

```
MAR = function(dset, time, percent, event)
# dset = dataset, time = event time, percent = percent missing data as a
decimal
#event = indicator variable: 1 if event type of interest and 0 otherwise
# e.g.
#
sample1$time.miss=MAR(dset=sample1,time=sample1$agvhd_time,perce
nt=0.1,
#event=sample1$agvhd)
{
 time.miss=c()
 for (j in 1: nrow(dset))
  {time.miss[j]=ifelse(rbinom(1,1,percent)==1 & event[j]==1,NA,time[j])}
 return(time.miss)
}
```

3.3. MNAR with smallest event times missing

```
MNAR = function(dset, time, percent, event)
# dset = dataset, time = event time, percent = percent missing data as a
#decimal
#event = indicator variable: 1 if event type of interest and 0 otherwise
# e.g.
#sample1$time.miss=MNAR(dset=sample1,time=sample1$agvhd_time,pe
#rcent=0.1,
#event=sample1$agvhd)
{
 dset.tmp=data.frame(cbind(event.tmp=event,
time.tmp=time,id=c(1:nrow(dset))))
 dset.tmp=dset.tmp[order(-dset.tmp$event.tmp, dset.tmp$time.tmp),]
 dset.tmp$time.miss=dset.tmp$time
dset.tmp$time.miss[1:round(percent*nrow(dset.tmp[dset.tmp$event.tmp=
=1,]))]=NA
 dset.tmp=dset.tmp[order(dset.tmp$id),]
 return(dset.tmp$time.miss)
}
```

4. Functions to apply the multiple imputation methods used in simulation studies

In each function, time = event time with missing times generated as per one of the missing data mechanisms, type = event type indicator, var = covariate and nimp = number of imputations

e.g. imp=NORM(time=sample1$time.miss, type=sample1$agvhd_status, var=sample1$doublecord, nimp=5)

4.1. Linear imputation model with no restrictions on the imputed values

```
NORM = function(time, type, var, nimp)
    {
    midata=data.frame(time=time, type=type , var)
     imp=mice(midata,
    m=nimp,defaultMethod=c('norm','logreg','polyreg'),print=FALSE)
    return(imp)
    }
    NORMNOAUX= function(time, type, nimp) #without covariate
    {
    midata=data.frame(time=time, type=type)
     imp=mice(midata,
    m=nimp,defaultMethod=c('norm','logreg','polyreg'),print=FALSE)
    return(imp)
    }
```

4.2. Type 1 PMM imputation model with no restrictions on the imputed values

```
PMM = function(time, type, var, nimp)
    {
    midata=data.frame(time=time, type=type , var)
     imp=mice(midata,m=nimp,print=FALSE)
    return(imp)
    }
    PMMNOAUX = function(time, type, nimp) #without covariate
    {
    midata=data.frame(time=time, type=type)
     imp=mice(midata,m=nimp,print=FALSE)
    return(imp)
    }
```

4.3. Log-linear imputation model with post-imputation back-transformation

```
LOGNORM = function(time, type, var, nimp)
    {
    midata=data.frame(time=log(time), type=type , var)
     imp=mice(midata,
    m=nimp,defaultMethod=c('norm','logreg','polyreg'),print=FALSE)
    imp[["imp"]]$time=exp(imp[["imp"]]$time)
    return(imp)
    }
```

4.4. Linear regression with restrictions on the imputed values (RESNORM).

```
# Written after looking at the following page on StackExchange:
# https://stats.stackexchange.com/questions/78632/multiple-#imputation-for-missing-values

# Adapt normal imputation method
mice.impute.norm3 = function (y, ry, x,...)
 {
   valid_vals <- rep(NA, length.out = sum(!ry))
   # Counter to avoid endless loop
   cntr = 0
   repeat{
    vals = mice.impute.norm(y, ry, x, ...)
    #Compare with boundaries
    correct = vals > 0 & vals <= 100
     if (all(!is.na(valid_vals) | correct)){
               valid_vals[correct] = vals[correct]
               break
         #stop if all values within boundaries
    }
         else if (any(is.na(valid_vals) & correct)){
             valid_vals[correct] = vals[correct]
    }
     cntr = cntr + 1
    if (cntr > 200){
          if (all(is.na(valid_vals))){
                    valid_vals = vals
     }
    else{
       valid_vals[is.na(valid_vals)] = vals[is.na(valid_vals)]
     }
     break
    }
   }
   return(valid_vals)
 }
RESNORM = function(time, type, var, nimp)
{
midata=data.frame(time=time, type=type , var)
imp=mice(midata, m=nimp, method=c(time = "norm3"),print=FALSE)
 return(imp)
}
```

4.5. MI method proposed by Delord and Genin

```
#Adaptation of the MI.ci function and associated functions from the
#'MIICD' R package
  preproc.crreg_2 <- function(  data = data  , m = m  , trans = trans , status =
  status , cens.code = cens.code ){
     rownames(data)<-seq_len(nrow(data))
   I <- data[,'left'] != data[,'right' ] & data[,'right'] != Inf & data[,status] !=
  cens.code
   data2<-data[I,]
   dim(data2)
   dataE<-data[!rownames(data)%in%rownames(data2),]
   or<-
  order(c(as.numeric(rownames(data2)),as.numeric(rownames(dataE))))
   data1<-t(apply( dataE , 1 , function(x) as.numeric(rep(x['left'] , m ))))
   dim(data1)[1]+dim(data2)[1]
   return( list( data2 = data2 , data1 = data1 , or = or, I = I ) )
  }
  get_z2<-function(data){
   s1<-sapply( 1:ncol(data) , function(y) sapply(1:nrow(data), function(x)
  length(unlist(data[x,y])))  )
   data[s1==0]<-0
   s2<-sapply( 1:ncol(data) , function(y) sapply(1:nrow(data), function(x)
  unlist(data[x,y])))
   return(s2)}
  MI.ci1<-function( m , status ,  trans , data , conf.int = TRUE , cens.code ,
  alpha = 0.05 , ntimes = NULL ){
     if( !is.numeric(m)  ) stop('m must be an integer')
   if( !is.data.frame(data)  ) stop('data must be a data.frame')
   if( !is.logical(conf.int)  ) stop('conf.int must be logical')
   if( alpha <= 0 | alpha >= 1 ) stop('alpha must be in ] 0 , 1 [')
    cl<-match.call()
   #Use interval censored data and generate k sets of imputed data
   sets<-sapply( 1:m , get.set2 , data )
   #Get and sort single times at wich the cumulative incidence will be
  estimated
   times<-as.vector(sets)
   length(times)
   r2<-factor(r2,levels=unique(c(cens.code,unique(r2))))
    r1 <- as.character(  data[ , status ] )
   r1<-factor(r1,levels=unique(c(cens.code,unique(r1))))
    fitCI <- survfit( Surv( time = times , event = r2 , type = "mstate"  ) ~ 1  ,
  weights = rep( 1 / m , length(times) ) ,
                conf.type = 'none' )
    w <- which( fitCI$states == trans )
   pr <- fitCI$pstate[ , w ]
   sd <- fitCI$std.err[ , w ]
```

```
  t0 <- fitCI$time
  #get estimated of cumulative incidence and confidence intervals
    if(! is.null(ntimes) ){
    t1 <- seq( from = range(t0)[1] , to=range(t0)[2] , length = ntimes )
    sd_at_times<-sapply( 1 , get_values_at_times2 , values =  list(sd) , times
= list(t0) , at = t1 , list = T )
    sd_at_times <- get_z2( sd_at_times )
    ci_at_times<-sapply( 1 , get_values_at_times2 , values =  list(pr) , times
= list(t0) , at = t1 , list = T )
    ci_at_times <- get_z2( ci_at_times )
      }else{
    t1 <- t0
    sd_at_times <- sd
    ci_at_times <- pr
  }
  if(conf.int){
    sap<-lapply( 1:m , get_est_mi2 , trans = trans , imp_sets = sets , data =
data , cens.code = cens.code ,  r2 = r1 )
    #obtain data frame of standard errors and point estimates
    cis <- sapply(sap,function(x) x[['est']])
    t3  <- sapply(sap,function(x) x[['time']])
    #get standard errors and point estimates at single times
    cis_at_times<-sapply( 1:m , get_values_at_times2 , values =  cis , times =
t3 , at = t1 , list = F )
    cis_at_times <- get_z2( cis_at_times )
    CI <- post_point_est_CI( beta = cis_at_times , sd = sd_at_times , times =
t1 , conf.int =  conf.int , alpha = alpha )
    CI <- unique(replace(CI , is.na(CI) , 0 ))
  }else{
    CI<-rbind(c(time = 0 ,  est = 0 ) , data.frame( time = t0 , est = pr ) )
    CI <- unique(replace(CI , is.na(CI) , 0 ) )
  }
  ret<-list( est = CI , call = cl , data = data , cens.code = cens.code , status =
status , conf.int = conf.int )
  class(ret) <- 'MI_ci'
  return(ret)
}
print.MI_ci_1 <- function (x , ... ) {
  cat('\nCumulative incidence estimation for interval censored data using
multiple imputation\n')
  cat( "\nCall:\n", paste( deparse(x$call) , sep = "\n" , collapse = "\n" ) ,
"\n\n" , sep = "")
  cat('Interval-censored response for cumulative incidence estimate :\n\n')
  n<-nrow(x$data)
  data<-x$data
  cens <- x$cens.code
```

```
  status<-x$status
  cat('No.Observation:', n , '\n')
  cat('Patern:\n')
  stat<-ifelse(data[,status]==cens,'unknown (right-
censored)',as.character(data[,status]))
  type<-ifelse(data$right==data$left , 'exact' , NA )
  type<-ifelse(data$right!=data$left & data$right!=Inf , 'interval-censored' ,
type )
  type<-ifelse(data[,status]==cens  , 'right-censored' , type )
  print(table('Cause'=stat, type))
  cat('\n')
  cat('$est\n')
  dimest<-paste(dim(x$est)[1] , 'x' , dim(x$est)[2])
  cat(paste('A',dimest,'data frame of required estimates\n'))
  print(head(x$est))
}
# plot method for MI_ci objects
# @param x A MI_ci object
# @inheritParams plot.MI_surv
plot.MI_ci_1 <- function (x , xlab = 'Time' , ylab = 'Cumulative incidence' ,
... )
{
  data <- x$est
  conf.int <- x$conf.int
  plot( data$time , data$est , xlab = xlab , ylab = ylab , type = 's' , ylim =
c(0,1) , bty ='l')
  if(conf.int){
    lines( data$time , data$uci , lty = 2 , type = 's' )
    lines( data$time , data$lci , lty = 2 , type = 's' )
  }
}
get.set2 <- function( k , data){
  data <- data[,c('left','right')]
  df1<-data.frame( data , r = runif(nrow(data)))
  w <- with( df1 , left +   (right - left)  * r )
  if1<-ifelse(w == Inf , df1$left , w)
  return(if1)
}
get_est_mi2<-function( x , status , trans , imp_sets , data = data , cens.code
= cens.code , model = c('Cox','FG') , r2 ){
    t1 <- imp_sets[ , x ]
  fitCI <- survfit( Surv( time = t1 , event = r2 , type = "mstate"  ) ~ 1 )
  w <- which( fitCI$states == trans )
  pr <- fitCI$pstate[ , w ]
  sd <- fitCI$std.err[ , w ]
  t1 <- fitCI$time
```

```
  CI <- list(time = t1 , est = pr , sd = sd )
  return( CI )
}
get_values_at_times2 <- function( j , values , times , at , list = F , unique =
T ){
  if(list==F){
      sapply( at , function( x ) {
     values[ , j ][ tail( which( sort( unique( times[ , j ] ) )  <= x ) , 1 ) ]
    }
    )
  }else{
    if(unique){
     sapply( at , function( x ) {
       values[[j]][ tail( which( sort( ( times[[j]] ) )  <= x ) , 1 ) ]
     }
     )
    }else{
     sapply( at , function( x ) {
       values[[j]][ tail( which( sort( unique( times[[j]] ) )  <= x ) , 1 ) ]
     }
     )
    }
  }
}
MI.ci_2<-function( k , m , data , status , trans , cens.code , conf.int = F ,
alpha = 0.05 ){
  if( !is.numeric(k)  ) stop('k must be an integer')
  if( !is.numeric(m)  ) stop('m must be an integer')
  if( !is.data.frame(data)  ) stop('data must be a data.frame')
  if( !is.logical(conf.int)  ) stop('conf.int must be logical')
  if( alpha <= 0 | alpha >= 1 ) stop('alpha must be in ] 0 , 1 [')
  #if(k <= 1) stop('You may consider the MI.ci function')
    cl<-match.call()
  prep     <- preproc.crreg_2( data = data , m = m , trans = trans , status =
status , cens.code = cens.code )
  data_int <- prep$data2
  data_fix <- prep$data1
  or       <- prep$or
  I        <- prep$I
  r2 <- as.character( rep( data[ , status ] , m  ) )
  r2 <- replace( r2 , r2 == cens.code ,  0 )
  r1 <- as.character(  data[ , status ]  )
  r1 <- replace( r1 , r1 == cens.code ,  0 )
   #Multiple Imputation
  CI <- MI.ci1( m = m , status = status , trans = trans , cens.code =
cens.code,
```

```
            data = data , conf.int = F , alpha = alpha , ntimes = NULL )$est
  CI$diff <- c(0 , diff( CI$est ) )
  for(i in 1:k){
    ss1<-apply(data_int , 1 , function(x ) subset( CI , time >=
as.numeric(x['left']) & time <= as.numeric(x['right']) ) )
    tk2<-lapply(seq_len(nrow(data_int)) ,function(X)  ss1[[X]]$time)
    samples<-t( sapply( seq_len(nrow(data_int)) , function(X) {
      pk2 <- ss1[[ X ]]$diff
      sapply( 1:m , function(x){
        if( sum( pk2 ) & length( pk2 ) > 1 ) sample(  tk2[[ X ]] , size = 1 , prob =
pk2 )
        else   mean( tk2[[ X ]] ) } )  } ) )
        samples2<-rbind(samples,data_fix)[or,]
    times<-as.vector(samples2)
        ci<-Surv( time = times , event = r2 , type = 'mstate')
    fitCI<-survfit( ci ~ 1 , weights = rep( 1 , length( times ) ) / m , conf.type
= 'none')
    w <- which( fitCI$states == trans )
    sd <- fitCI$std.err[ , w ]
    pr <- fitCI$pstate[ , w ]
    t0 <- fitCI$time
    CI<-unique(rbind(c(time = 0 ,  est = 0 ) , data.frame( time = t0 , est = pr )
))
    CI$diff <- c(0 , diff( CI$est ) )
  }
    sap<-lapply( 1:m , get_est_mi2 , trans = trans , imp_sets = samples2  ,
data = data , r2 = r1 )
  #obtain data frame of standard errors and point estimates
  cis <- sapply(sap,function(x) x[['est']])
  t3  <- sapply(sap,function(x) x[['time']])
    #get standard errors and point estimates at single times
    cis_at_times<-sapply( 1:m , get_values_at_times2 , values =  cis , times =
t3 , at = t0 , list = is.list(cis) )
  cis_at_times <- get_z2( cis_at_times )
    #Amended E Curnow
  #Next statement is for checking data that are input into
post_point_est_CI function - #not required
  #ret<-list( est = cis_at_times, sd=sd, times=t0)
   #E Curnow: don't return results of Rubin's rules
  #CI <- post_point_est_CI( beta = cis_at_times , sd = sd , times = t0 ,
conf.int =  #conf.int , alpha = alpha )
    #if(conf.int){
   # colnames(CI)<-c('time','prev','sd','uci','lci')
    #CI <- unique(replace(CI , is.na(CI) , 0 ))
  #}else{
   # colnames(CI)<-c('time','prev')
```

```
  #CI <- unique(replace(CI , is.na(CI) , 0 ))
 #}
 #ret<-list( est = CI , call = cl , data = data , cens.code = cens.code , status =
status , #conf.int = conf.int )
 #class(ret) <- 'MI_ci'
  #Added E Curnow
 return(sap)
}

#Function to apply adapted version of MI.ci
MICI = function(time, type, nimp)
{
#definitions needed for left and right boundary
left=time
right=as.numeric(ifelse(type==0,"inf",time)) #for right-censored data
#apply boundaries for missing times
left=ifelse(is.na(time),0, left)
right=ifelse(is.na(time),100, right)
midata=data.frame(left=left, right=right, status=type)
imp=MI.ci_2(k=5,m=nimp, data=midata, status="status", trans=1,
cens.code=0,
          conf.int = F, alpha = 0.05)
return(imp)}
```

5. Calculating estimates for each imputation in turn and applying Rubin's rules – non-parametric competing risks analysis

5.1. General function for all imputation methods except Delord and Genin's

```
Est.calc = function(imp, nimp)
# imp = output from call of 'mice' using one of the functions above, nimp =
#number of imputations e.g. ests=Est.calc(imp=imp,nimp=5)
   {
    agvhdcuminc_est=c()
    agvhdcuminc_SE=c()
    q2_est=c()
    q2_SE=c()
    for (j in 1:nimp)
            {
                fit=Cuminc("time","type",data=complete(imp,j))
      agvhdcuminc_est[j]=head(fit[fit$time >=100,],1)$CI.1
      agvhdcuminc_SE[j]=head(fit[fit$time >=100,],1)$seCI.1
                q2=fit[fit$CI.1>=0.5,]
                q2_l=fit[fit$CI.1<=0.49,]
                q2_u=fit[fit$CI.1>=0.51,]
                f_q2=(head(q2_u,1)$CI.1 - tail(q2_l,1)$CI.1)/
                    (head(q2_u,1)$time - tail(q2_l,1)$time)
      q2_est[j]=head(q2,1)$time
```

```
            q2_SE[j]=head(q2,1)$seCI.1/f_q2
          }
  q2_results=mean(q2_est)
  q2_SE_results=sqrt(mean(q2_SE^2)+((1+1/nimp)*1/(nimp-
1)*sum((q2_est-mean(q2_est))^2)))
  cuminc_est_results=mean(agvhdcuminc_est)
cuminc_SE_results=sqrt(mean(agvhdcuminc_SE^2)+((1+1/nimp)*1/(nim
p-1)*sum((agvhdcuminc_est-mean(agvhdcuminc_est))^2)))
 ests=cbind(cuminc_est_results, cuminc_SE_results, q2_results,
q2_SE_results)
return(ests)
}
```

5.2. Function to be used for Delord and Genin's method

```
Est.calc.MICI = function(imp, nimp)
# imp = output from call of MICI function, nimp = number of imputations
# e.g. ests=Est.calc.MICI(imp=imp,nimp=5)
  {
   agvhdcuminc_est=c()
   agvhdcuminc_SE=c()
   q2_est=c()
   q2_SE=c()
  for (j in 1:nimp)
          {
            agvhdcuminc_est[j]=head(imp[[j]]$est[imp[[j]]$time>=100],1)
            agvhdcuminc_SE[j]=head(imp[[j]]$sd[imp[[j]]$time>=100],1)
            q2=imp[[j]]$time[imp[[j]]$est>=0.5]
            q2_l=imp[[j]]$time[imp[[j]]$est<=0.49]
            q2_u=imp[[j]]$time[imp[[j]]$est>=0.51]
            q2_l_ci=imp[[j]]$est[imp[[j]]$est<=0.49]
            q2_u_ci=imp[[j]]$est[imp[[j]]$est>=0.51]
            f_q2=(head(q2_u_ci,1) - tail(q2_l_ci,1))/
              (head(q2_u,1) - tail(q2_l,1))
            q2_est[j]=head(q2,1)
            q2_SE[j]=head(imp[[j]]$sd[imp[[j]]$est>=0.5],1)/f_q2  }
  q2_results=mean(q2_est)
  q2_SE_results=sqrt(mean(q2_SE^2)+((1+1/nimp)*1/(nimp-
  1)*sum((q2_est-mean(q2_est))^2)))
  cuminc_est_results=mean(agvhdcuminc_est)
  cuminc_SE_results=sqrt(mean(agvhdcuminc_SE^2)+((1+1/nimp)*1/(nim
  p-1)*sum((agvhdcuminc_est-mean(agvhdcuminc_est))^2)))
  ests=cbind(cuminc_est_results, cuminc_SE_results, q2_results,
  q2_SE_results)
  return(ests)}
```

6. Complete case analysis

6.1. Non-parametric competing risks model

```
 CCA = function(time, type)
{
dset=data.frame(time=time, type=type)
fit=Cuminc("time","type",data=dset)
cuminc_est_results=head(fit[fit$time >=100,],1)$CI.1
  cuminc_SE_results=head(fit[fit$time >=100,],1)$seCI.1
  q2=fit[fit$CI.1>=0.5,]
  q2_l=fit[fit$CI.1<=0.49,]
  q2_u=fit[fit$CI.1>=0.51,]
  f_q2=(head(q2_u,1)$CI.1 - tail(q2_l,1)$CI.1)/
  (head(q2_u,1)$time - tail(q2_l,1)$time)
  q2_results=head(q2,1)$time
  q2_SE_results=head(q2,1)$seCI.1/f_q2
    ests=cbind(cuminc_est_results, cuminc_SE_results, q2_results,
    q2_SE_results)
    return(ests)
}
```

6.2. MSM

```
#1. Cox model for each transition intensity
#Function for RELOS SE
RELOS_SE_boot <- function(data,nboot,agvhdtime,reldeathtime) {
 library(survival)
 library(mstate)
 #Adapting Bakoyannis bssmle_se R code
 tmp <- data.frame()
 for(k in 1:nboot){
  samp=data[sample(dim(data)[1], replace = TRUE),]
  samp$boot=k
  tmp <- rbind(tmp,samp)
 }
 m <- NULL
 tmat <- transMat(x = list(c(2, 3), c(3), c()), names = c("Tx", "aGvHD",
"Rel/Death"))
 res.bt=data.frame(ELOS1=0,ELOS2=0,ELOS3=0)
 for(m in 1:nboot){

  #prep dset in 'long' format
  tmplong <- mstate::msprep(data = tmp[tmp$boot==m,], trans = tmat,
                    time = c(NA, agvhdtime, reldeathtime),
                    status = c(NA, "agvhd_status", "rel_death_status"),
                    keep = c("doublecord","relapse"))
  covs <- c("doublecord", "relapse")
  tmplong_cov <- mstate::expand.covs(tmplong, covs, longnames = FALSE)
```

```
   #allow for errors
   skip_to_next <- FALSE
    tryCatch(survival::coxph(Surv(Tstart, Tstop, status) ~ relapse.1 + relapse.2
   + relapse.3 + doublecord.3 + strata(trans), data = tmplong_cov, method =
   "breslow"),
          error = function(e) { skip_to_next <<- TRUE},
          warning = function(w) {skip_to_next <<- TRUE})
   if(skip_to_next) { next }
   tmpfit=NULL
   tmpfit=survival::coxph(Surv(Tstart, Tstop, status) ~ relapse.1 + relapse.2 +
   relapse.3 + doublecord.3 + strata(trans), data = tmplong_cov, method =
   "breslow")
   baseline <-
   data.frame(trans=1:3,relapse.1=c(0,0,0),relapse.2=c(0,0,0),relapse.3=c(0,0,0),
                    doublecord.3=c(0,0,0),strata=1:3)
   tmpmsf = mstate::msfit(tmpfit, baseline, trans = tmat)
   #calc probs
   tmppt=mstate::probtrans(tmpmsf, predt = 0, variance=FALSE)
   tmpELOS=data.frame(ELOS1=0,ELOS2=0,ELOS3=0)
   for (j in 1:(length(tmppt[[1]]$time[tmppt[[1]]$time<731])-1))
   {

tmpELOS$ELOS1=tmpELOS$ELOS1+(tmppt[[1]]$pstate1[j]*(tmppt[[1]]$time[
j+1]-tmppt[[1]]$time[j]))

tmpELOS$ELOS2=tmpELOS$ELOS2+(tmppt[[1]]$pstate2[j]*(tmppt[[1]]$time[
j+1]-tmppt[[1]]$time[j]))

tmpELOS$ELOS3=tmpELOS$ELOS3+(tmppt[[1]]$pstate3[j]*(tmppt[[1]]$time[
j+1]-tmppt[[1]]$time[j]))
   }

   res.bt[m,]=tmpELOS

  }
  result <- sqrt(diag(var(res.bt[,1:3],na.rm=T)))

  return(result)
}
```

```
####Complete data analysis####
results = data.frame(relapse.1=0,relapse.2=0,relapse.3=0,doublecord.3=0,
relapse.1SE=0,relapse.2SE=0,relapse.3SE=0,doublecord.3SE=0,
agvhd_time_coverage=0,
ELOS1=0,ELOS2=0,ELOS3=0,ELOS_SE1=0,ELOS_SE2=0,ELOS_SE3=0)
for (i in 1:1000)
{
 sample1=simdata_all[simdata_all$sampno==i,]
 tmat <- transMat(x = list(c(2, 3), c(3), c()), names = c("Tx", "aGvHD",
 "Rel/Death"))
 sample1_long <- msprep(data = sample1, trans = tmat, time = c(NA,
 "agvhd_time", "rel_death_time"),
 status = c(NA, "agvhd_status", "rel_death_status"),
 keep = c("doublecord","relapse","sampno","agvhd_time"), id="id")
 covs <- c("doublecord", "relapse")
 sample1_long_cov <- expand.covs(sample1_long, covs, longnames = FALSE)
   fit <- coxph(Surv(Tstart, Tstop, status) ~ relapse.1 + relapse.2 + relapse.3 +
  doublecord.3 + strata(trans), data = sample1_long_cov, method =
  "breslow")
 results[i,1:4]=summary(fit)$coefficients[1:4]
 results[i,5:8]=sqrt(diag(vcov(fit)))
 #gamma12
  fit12 <- coxph(Surv(Tstart, Tstop, status) ~ relapse + doublecord +
  agvhd_time,
 data = sample1_long, method = "breslow", subset=(trans=="3"))
 results[i,9]=ifelse(summary(fit12)$coefficients[3,5]>=0.05,1,0)
 #RELOS
  baseline <-
  data.frame(trans=1:3,relapse.1=c(0,0,0),relapse.2=c(0,0,0),relapse.3=c(0,0,0),
          doublecord.3=c(0,0,0),strata=1:3)
 msf0 <- msfit(fit, baseline, trans = tmat)
 #calc trans probs
 pt0=probtrans(msf0, predt = 0, variance=FALSE)
 ELOS=data.frame(ELOS1=0,ELOS2=0,ELOS3=0)
 for (j in 1:(length(pt0[[1]]$time[pt0[[1]]$time<731])-1))
 {
  ELOS$ELOS1=ELOS$ELOS1+(pt0[[1]]$pstate1[j]*(pt0[[1]]$time[j+1]-
  pt0[[1]]$time[j]))
 ELOS$ELOS2=ELOS$ELOS2+(pt0[[1]]$pstate2[j]*(pt0[[1]]$time[j+1]-
 pt0[[1]]$time[j]))
 ELOS$ELOS3=ELOS$ELOS3+(pt0[[1]]$pstate3[j]*(pt0[[1]]$time[j+1]-
 pt0[[1]]$time[j]))
  }
 results[i,10:12]=ELOS
 results[i,13:15]=RELOS_SE_boot(data=sample1,nboot=50,"agvhd_time",
"rel_death_time")}
```

#2. Weibull model for each transition intensity

```
#define function for ELOS SE
ELOS_delta=function(obj){
msf0 <- msfit.flexsurvreg(obj, t=seq(0,730,by=0.1), trans=tmat,
newdata=baseline,variance=FALSE)
 #calc trans probs
 pt0=probtrans(msf0, predt = 0, variance=FALSE)

 ELOS=data.frame(ELOS1=0,ELOS2=0,ELOS3=0)
 for (j in 1:(length(pt0[[1]]$time[pt0[[1]]$time<731])-1))
 {
  ELOS$ELOS1=ELOS$ELOS1+(pt0[[1]]$pstate1[j]*(pt0[[1]]$time[j+1]-
pt0[[1]]$time[j]))
  ELOS$ELOS2=ELOS$ELOS2+(pt0[[1]]$pstate2[j]*(pt0[[1]]$time[j+1]-
pt0[[1]]$time[j]))
  ELOS$ELOS3=ELOS$ELOS3+(pt0[[1]]$pstate3[j]*(pt0[[1]]$time[j+1]-
pt0[[1]]$time[j]))
 }
 return(ELOS)
}

####Complete data analysis####
results = data.frame(relapse.1=0,relapse.2=0,relapse.3=0,doublecord.3=0,
relapse.1SE=0,relapse.2SE=0,relapse.3SE=0,doublecord.3SE=0,
agvhd_time_coverage=0,
ELOS1=0,ELOS2=0,ELOS3=0,ELOS_SE1=0,ELOS_SE2=0,ELOS_SE3=0)

for (i in 1:1000)
{
 sample1=simdata_all [simdata_all$sampno==i,]

  tmat <- transMat(x = list(c(2, 3), c(3), c()), names = c("Tx", "aGvHD",
  "Rel/Death"))
  sample1_long <- msprep(data = sample1, trans = tmat, time = c(NA,
  "agvhd_time", "rel_death_time"),
  status = c(NA, "agvhd_status", "rel_death_status"),
  keep = c("doublecord","relapse","sampno","agvhd_time"), id="id")
  fit.list <- vector(3, mode="list")
  fit.list[[1]]=flexsurvreg(Surv(Tstart, Tstop, status) ~ relapse, subset = (trans
  == 1), data = sample1_long,dist ="weibullPH",
  inits=c(shape=1.5,scale=36^(-1.5)))
 fit.list[[2]]=flexsurvreg(Surv(Tstart, Tstop, status) ~ relapse, subset = (trans
 == 2), data = sample1_long,dist = "weibullPH",
 inits=c(shape=0.9,scale=120^(-0.9)))
 fit.list[[3]]=flexsurvreg(Surv(Tstart, Tstop, status) ~ relapse + doublecord,
 subset = (trans == 3),
```

```
 data = sample1_long,dist = "weibullPH", inits=c(shape=0.8,scale=160^(-0.8)))
 results[i,1]=fit.list[[1]]$res[3,1]
 results[i,2]=fit.list[[2]]$res[3,1]
 results[i,3:4]=fit.list[[3]]$res[3:4,1]
 #SE
 results[i,5]=fit.list[[1]]$res[3,4]
 results[i,6]=fit.list[[2]]$res[3,4]
 results[i,7:8]=fit.list[[3]]$res[3:4,4]
 #gamma12
  fit12=flexsurvreg(Surv(Tstart, Tstop, status) ~ relapse + doublecord +
  agvhd_time, subset = (trans == 3), data = sample1_long, dist = "weibullPH",
  inits=c(shape=0.8,scale=160^(-0.8)))
  results[i,9]=ifelse(fit12$res[5,2]<0 & fit12$res[5,3]>0,1,0)
 #now apply msfit to obtain estimates of ELOS
 #Define pt with baseline values of covariates
 baseline <- data.frame(trans=1:3,relapse=c(0,0,0),doublecord=c(0,0,0))
  msf0 <- msfit.flexsurvreg(fit.list, t=seq(0,730,by=0.1), trans=tmat,
  newdata=baseline,variance=FALSE)

#calc trans probs
 pt0=probtrans(msf0, predt = 0, variance=FALSE)

 ELOS=data.frame(ELOS1=0,ELOS2=0,ELOS3=0)
 for (j in 1:(length(pt0[[1]]$time[pt0[[1]]$time<731])-1))
 {
  ELOS$ELOS1=ELOS$ELOS1+(pt0[[1]]$pstate1[j]*(pt0[[1]]$time[j+1]-
pt0[[1]]$time[j]))
  ELOS$ELOS2=ELOS$ELOS2+(pt0[[1]]$pstate2[j]*(pt0[[1]]$time[j+1]-
pt0[[1]]$time[j]))
  ELOS$ELOS3=ELOS$ELOS3+(pt0[[1]]$pstate3[j]*(pt0[[1]]$time[j+1]-
pt0[[1]]$time[j]))
 }
 results[i,10:12]=ELOS[1:3]

 #SE using the delta method
 #Obtaining sigma
 #note using res not coef as want to keep params on original scale
 coef=c(as.numeric(fit.list[[1]]$res[1:2,1]),as.numeric(fit.list[[2]]$res[1:2,1]),
      as.numeric(fit.list[[3]]$res[1:2,1]))
  #need to transform each weibull param est from cov as only shown in log-
  #transformed mode
 #using delta method
 cov=matrix(rep(0,36),ncol=6)#assuming independence between strata
 cov[1,1]=(coef[1])^2*(as.numeric(fit.list[[1]]$cov[1,1]))
 cov[1,2]=coef[1]*coef[2]*(as.numeric(fit.list[[1]]$cov[1,2]))
 cov[2,1]=cov[1,2]
```

```
 cov[2,2]=(coef[2])^2*(as.numeric(fit.list[[1]]$cov[2,2]))
 cov[3,3]=(coef[3])^2*(as.numeric(fit.list[[2]]$cov[1,1]))
 cov[3,4]=coef[3]*coef[4]*(as.numeric(fit.list[[2]]$cov[1,2]))
 cov[4,3]=cov[3,4]
 cov[4,4]=(coef[4])^2*(as.numeric(fit.list[[2]]$cov[2,2]))
 cov[5,5]=(coef[5])^2*(as.numeric(fit.list[[3]]$cov[1,1]))
 cov[5,6]=coef[5]*coef[6]*(as.numeric(fit.list[[3]]$cov[1,2]))
 cov[6,5]=cov[5,6]
 cov[6,6]=(coef[6])^2*(as.numeric(fit.list[[3]]$cov[2,2]))

 #Calculate partial derivatives using finite differences
 eps=0.00001
 #Initialise dataset
 partial=matrix(rep(0,18),ncol=3)

 fit.listu1=fit.list
 #fit.listu1[[1]]$res.t[1,1]=fit.list[[1]]$res.t[1,1]+eps
 fit.listu1[[1]]$res.t[1,1]=log(fit.list[[1]]$res[1,1]+eps)
  #change param by eps on original scale
 fit.listl1=fit.list
 #fit.listl1[[1]]$res.t[1,1]=fit.list[[1]]$res.t[1,1]-eps
 fit.listl1[[1]]$res.t[1,1]=log(fit.list[[1]]$res[1,1]-eps)
 partial[1,1:3]=c(as.numeric((ELOS_delta(fit.listu1)-
ELOS_delta(fit.listl1))/(2*eps)))[1:3]

 fit.listu2=fit.list
 #fit.listu2[[1]]$res.t[2,1]=fit.list[[1]]$res.t[2,1]+eps
 fit.listu2[[1]]$res.t[2,1]=log(fit.list[[1]]$res[2,1]+eps)
 fit.listl2=fit.list
 #fit.listl2[[1]]$res.t[2,1]=fit.list[[1]]$res.t[2,1]-eps
 fit.listl2[[1]]$res.t[2,1]=log(fit.list[[1]]$res[2,1]-eps)
 partial[2,1:3]=c(as.numeric((ELOS_delta(fit.listu2)-
ELOS_delta(fit.listl2))/(2*eps)))[1:3]

 fit.listu3=fit.list
 #fit.listu3[[2]]$res.t[1,1]=fit.list[[2]]$res.t[1,1]+eps
 fit.listu3[[2]]$res.t[1,1]=log(fit.list[[2]]$res[1,1]+eps)
 fit.listl3=fit.list
 #fit.listl3[[2]]$res.t[1,1]=fit.list[[2]]$res.t[1,1]-eps
 fit.listl3[[2]]$res.t[1,1]=log(fit.list[[2]]$res[1,1]-eps)
 partial[3,1:3]=c(as.numeric((ELOS_delta(fit.listu3)-
ELOS_delta(fit.listl3))/(2*eps)))[1:3]

 fit.listu4=fit.list
 #fit.listu4[[2]]$res.t[2,1]=fit.list[[2]]$res.t[2,1]+eps
 fit.listu4[[2]]$res.t[2,1]=log(fit.list[[2]]$res[2,1]+eps)
```

```
  fit.listl4=fit.list
  #fit.listl4[[2]]$res.t[2,1]=fit.list[[2]]$res.t[2,1]-eps
  fit.listl4[[2]]$res.t[2,1]=log(fit.list[[2]]$res[2,1]-eps)
  partial[4,1:3]=c(as.numeric((ELOS_delta(fit.listu4)-
ELOS_delta(fit.listl4))/(2*eps)))[1:3]

  fit.listu5=fit.list
  #fit.listu5[[3]]$res.t[1,1]=fit.list[[3]]$res.t[1,1]+eps
  fit.listu5[[3]]$res.t[1,1]=log(fit.list[[3]]$res[1,1]+eps)
  fit.listl5=fit.list
  #fit.listl5[[3]]$res.t[1,1]=fit.list[[3]]$res.t[1,1]-eps
  fit.listl5[[3]]$res.t[1,1]=log(fit.list[[3]]$res[1,1]-eps)
  partial[5,1:3]=c(as.numeric((ELOS_delta(fit.listu5)-
ELOS_delta(fit.listl5))/(2*eps)))[1:3]

  fit.listu6=fit.list
  #fit.listu6[[3]]$res.t[2,1]=fit.list[[3]]$res.t[2,1]+eps
  fit.listu6[[3]]$res.t[2,1]=log(fit.list[[3]]$res[2,1]+eps)
  fit.listl6=fit.list
  #fit.listl6[[3]]$res.t[2,1]=fit.list[[3]]$res.t[2,1]-eps
  fit.listl6[[3]]$res.t[2,1]=log(fit.list[[3]]$res[2,1]-eps)
  partial[6,1:3]=c(as.numeric((ELOS_delta(fit.listu6)-
ELOS_delta(fit.listl6))/(2*eps)))[1:3]

  results[i,13]=sqrt(t(partial[,1]) %*% cov %*% partial[,1])
  results[i,14]=sqrt(t(partial[,2]) %*% cov %*% partial[,2])
  results[i,15]=sqrt(t(partial[,3]) %*% cov %*% partial[,3])
  #results[i,5:10]=ELOS
 }
```

7. B-spline sieve semiparametric maximum likelihood approach of Bakoyannis *et al.*

```
#Adaptation of the bssmle_se function from the 'intccr' R package
 bssmle_se2 <- function(data,nboot) {
  #From bssmle_se R code
  tmp <- list()
  for(k in 1:nboot){
   tmp[[k]] <- data[sample(dim(data)[1], replace = TRUE), ]
  }
   m <- NULL
  no.cores <- parallel::detectCores() - 1
  clst <- parallel::makeCluster(no.cores)
  doParallel::registerDoParallel(clst)
   res.bt <- foreach(m = 1:nboot,
                     .combine = "rbind",
```

```
                .packages = c("intccr", "splines", "stats", "alabama", "utils"))
%dopar% {
                  pb <- utils::txtProgressBar(title = "Progress bar for the
bootstrapping",
                                              min = 0, max = nboot, style = 3)
                  utils::setTxtProgressBar(pb, m)
                  #Amended by E Curnow

tmpfit=ciregic(formula=Surv2(v,u,event=c)~1,tmp[[m]],alpha=c(0,0),nboot=0,
do.par=FALSE)
                  tmppfit <- predict(object = tmpfit, covp=1, times = c(1:100))

pars=c(head(tmppfit$t[tmppfit$cif1>=0.5],1),tmppfit$cif1[tmppfit$t==100])
                  #End of added code

                  return(pars)
                  close(pb)
                }
  parallel::stopCluster(clst)
  rownames(res.bt) <- c()

  result <- list(numboot = if(is.vector(res.bt)) 1 else nrow(na.omit(res.bt)),
                 Sigma = if(is.vector(res.bt)) res.bt else var(na.omit(res.bt)))
  #End of Bakoyannis code
  return(result$Sigma)
}

#Function to calculate estimates
    INTCCR = function(time, type)
    {
    #definitions needed for left and right boundary
      v=as.numeric(time-1)
      u= as.numeric(ifelse(type==0,"Inf", time))
     #apply boundaries for missing times
    v= as.numeric(ifelse(is.na(time),0, v))
    u= as.numeric(ifelse(is.na(time),100, u))
    #Only 1 competing event allowed so recode event type 3
    c= as.numeric(ifelse(type==3,2,type))
    dset=data.frame(v=v, u=u, c=c)
    #Compute the MLE using the Fine and Gray model
    fit=ciregic(formula=Surv2(v,u,event=c)~1,data=dset,alpha=c(0,0),nboot=0,
    do.par=FALSE)
     #Calculate estimates
      #Predict CI at 100 days
      pfit <- predict(object = fit, covp=1, times = c(1:100))
      q2_results=head(pfit$t[pfit$cif1>=0.5],1)
```

```
 cuminc_est_results=pfit$cif1[pfit$t==100]
 #Compute variance using non-parametric bootstrap
 Sigma=bssmle_se2(data=dset, nboot=50)
  q2_SE_results = sqrt(Sigma[1,1])
 cuminc_SE_results = sqrt(Sigma[2,2])
ests=cbind(cuminc_est_results, cuminc_SE_results, q2_results,
q2_SE_results)
return(ests)}
```

8. Performance measures

```
#Calculating theoretical values of cumulative incidence at 100 days and median
#time
cuminc100_T=pnorm(log(100), mean=3.26,sd=0.71)*0.65
median_T=ceiling(exp(qnorm(p=0.5*500/325, mean=3.26,sd=0.71)))
#Run desired missing data mechanism and missing data method for each
simulated dataset
#Initialise results dataset
results = matrix(NA,1000,4)
#colnames(results)=c("cuminc_est_results", "cuminc_SE_results", "q2_results",
#"q2_SE_results")
#Run simulation
for (i in 1:1000)
{
sample1=sample500[sample500$sampno==i,]
#Specify missing data mechanism, for example:
sample1$time.miss=MCAR(dset=sample1,time=sample1$agvhd_time,percent=0.
1)
#Specify missing data method, for example:
results[i,]=CCA(time=sample1$time.miss, type=sample1$agvhd_status)
#Example using an imputation method
#imp=MICI(time=sample1$time.miss, type=sample1$agvhd_status, nimp=5)
# results[i,]=Est.calc.MICI(imp=imp,nimp=5)}
# Calculate standardised bias and average model-based SE for estimands of
interest
bias_cuminc=mean(results[,1],na.rm=T)-cuminc100_T
empSE_cuminc=sqrt(var(results[,1],na.rm=T))
stand_bias_cuminc=bias_cuminc/empSE_cuminc
stand_bias_cuminc
ModSE_cuminc=sqrt(mean(results[,2]^2))
ModSE_cuminc
bias_q2=mean(results[,3],na.rm=T)-median_T
empSE_q2=sqrt(var(results[,3],na.rm=T))
stand_bias_q2=bias_q2/empSE_q2
stand_bias_q2
ModSE_q2=sqrt(mean(results[,4]^2))
ModSE_q2
```